ABSTRACT

Title of Thesis: CLASSIFYING BIAS IN LARGE MULTILINGUAL CORPORA VIA CROWDSOURCING AND TOPIC MODELING

Team BIASES:
Brianna Caljean, Katherine Calvert, Ashley Chang, Elliot Frank, Rosana Garay Jáuregui, Geoffrey Palo, Ryan Rinker, Gareth Weakly, Nicolette Wolfrey, William Zhang

Thesis Directed By: Dr. David Zajic, Ph.D.

Our project extends previous algorithmic approaches to finding bias in large text corpora. We used multilingual topic modeling to examine language-specific bias in the English, Spanish, and Russian versions of Wikipedia. In particular, we placed Spanish articles discussing the Cold War on a Russian-English viewpoint spectrum based on similarity in topic distribution. We then crowdsourced human annotations of Spanish Wikipedia articles for comparison to the topic model. Our hypothesis was that human annotators and topic modeling algorithms would provide correlated results for bias. However, that was not the case. Our annotators indicated that humans were more perceptive of sentiment in article text than topic distribution, which suggests that our classifier provides a different perspective on a text's bias.

CLASSIFYING BIAS IN LARGE MULTILINGUAL CORPORA VIA

CROWDSOURCING AND TOPIC MODELING

by

Team BIASES:

Brianna Caljean, Katherine Calvert, Ashley Chang, Elliot Frank, Rosana Garay Jáuregui,

Geoffrey Palo, Ryan Rinker, Gareth Weakly, Nicolette Wolfrey, William Zhang

Thesis submitted in partial fulfillment of the requirements of the
Gemstone Honors Program, University of Maryland,
2018

Advisory Committee:

Dr. David Zajic, Chair
Dr. Brian Butler
Dr. Marine Carpuat
Dr. Melanie Kill
Dr. Philip Resnik
Mr. Ed Summers

Acknowledgements

We would like to express our sincerest gratitude to our mentor, Dr. David Zajic, for his invaluable expertise and unwavering enthusiasm throughout all three years of our project. Without his continuous motivation and patient guidance, this research and thesis would not have been possible. Next, we want to thank the Gemstone staff and our librarian Eric Lindquist for their steady encouragement and support throughout our research process. We would especially like to acknowledge Assistant Research Scientist Paul Rodrigues, Associate Professor David Sartorius, Professor Madeline C. Zilfi, technical writer Naomi Chang Zajic, and researcher Elena Zotkina for their consultations on our project, as well as our discussants: Dr. Brian Butler, Dr. Marine Carpuat, Dr. Melanie Kill, Dr. Philip Resnik, and Mr. Ed Summers. We would like to thank the FedCentric staff for providing us with computing power, which allowed us to run our tests. In addition, we would be remiss not to acknowledge the considerable contributions of former team member Cassidy Laidlaw to our project design and implementation. Lastly, we would like to express our appreciation to our families for being extremely supportive of our research.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Wikipedia is one of the most used sources of online information, with daily pageviews in the millions (“Wikipedia: Awareness statistics”, 2017). As such, bias in Wikipedia articles may affect millions of readers around the world. Additionally, since textual bias encompasses a broad spectrum of imbalances in information, including bias in topic distribution or in sentiment, it can be difficult for readers to determine if the contents of encyclopedic texts such as Wikipedia contain bias. Conscious of the broad influence of the online encyclopedia, Wikipedia employs a Neutral Point of View (NPOV) policy which states that articles must represent “fairly, proportionately, and, as far as possible, without editorial bias, all of the significant views that have been published by reliable sources on a topic” (“Wikipedia: Neutral point of view”, 2015).

However, there are some factors that could create bias in Wikipedia articles discussing historical events. Wikipedia editors are not trained to understand the full contexts of events, which can lead to modern editors anachronistically applying modern views to historical subjects. Additionally, while general dates and basic facts are correct, editors are influenced by a variety of factors including nationality and culture. Also, computers could possibly identify hidden trends in Wikipedia based on how often topics are discussed (Madeline Zilfi, personal correspondence, September 30, 2016). This insight constitutes the seed of our research.

Analyzing instances of bias is a human task that requires a considerable amount of time to examine a large quantity of sources. Additional complications can arise when sources are in multiple languages or in a language unfamiliar to the reader. Through establishing a computational framework for quantifying bias across multilingual corpora, this cultural analysis can be expedited and applied to larger corpora. Culture is not a homogeneous entity, as it encounters variations due to historical circumstances and geographic distribution. To simplify, this project delineated culture using language, assuming that there is a correspondence between language and culture. The project focused on English and Russian Wikipedia articles on the Cold War and assumed that these articles were reflective of US and USSR stances on Cold War issues. Then, we placed articles reflecting Latin American viewpoints, represented by Spanish articles, on a spectrum between the US and USSR viewpoints. Thus, we create a three-viewpoint model where Spanish articles are placed on a spectrum between the US and USSR viewpoints.

Bias, for the purposes of this project, encompasses imbalances in information (i.e. topic distribution, subject matter, etc.) or the aspects of how information is presented (i.e. tone, sentiment, etc.) that would suggest a preference or prejudice when regarding history. Because the policies of Wikipedia discourage the most easily observable form of bias, emotionally loaded language, we choose to examine more indirect forms such as information imbalance. However, the policy of not allowing direct translations of articles and encouraging local native-language writing can encourage a diversity of information that can introduce new bias. We believe that this type of bias would present itself in an informational imbalance that could suggest a preference for one side of an argument. A preference or prejudice is generally identified by humans through qualitative analysis of the content and is present when the content supports one side of an issue more strongly than another. Qualitative aspects that can indicate bias include the connotations associated with certain words, meanings implied by the organization of the content, and what specific information is presented.

Our research explores the differences in how a topic-modeling approach and a crowdsourced human approach differ in the information they can provide about a Wikipedia article's bias. To explore these questions, we consider the case of Wikipedia articles about the Cold War. We use articles from the English and Russian Wikipedias, which are assumed to have American and Soviet bias respectively, to train an algorithm that models bias along a spectrum between American and Soviet viewpoints. We then use our algorithm to place Spanish Wikipedia articles about the Cold War on this spectrum. This process is referred to as our three-viewpoint model. The algorithm is then evaluated by comparing its results to human judgments of bias in the same articles.

One drawback of using a corpus compiled from Wikipedia is that the various Wikipedias, being differentiated by language, are not homogeneous with respect to culture, just as languages are not homogeneous with respect to culture. Therefore, a Wikipedia edition in a certain language may not represent a single country or even a single region of the world, as writers of articles in one language may not share a culture or country. For instance, because English is spoken in the United States, the United Kingdom, Australia, Canada, and many other countries, the English Wikipedia receives contributions from around the world (Ghosh, Glott, & Schmidt, 2010). One justification for our approach at comparing multilingual Wikipedias is that while languages may not correspond to specific regions around the world, they can still represent cultural viewpoints. Lieberman & Lin (2009), for example, successfully use edit histories to approximate the location of the editor (e.g. if a user edited the pages for the New York Stock Exchange, Central Park, and Fifth Avenue, then they are likely to be located in New York City). Thus, it is reasonable to assume that, in general, Wikipedia edits on historical articles are likely to come from the regions discussed in the article and thus represent culture rather than simply language.

In our research, the US viewpoint on the Cold War is represented by English and the USSR viewpoint is represented by Russian. Documents written in Spanish, reflecting the viewpoints of Latin American authors, were placed along a spectrum between USSR and US viewpoints.

In the literature review, we present previous work from the fields of Wikipedia research, Latin American studies, and computer science (specifically topic modeling). In the methodology section, we present our data collection techniques, which included crowdsourcing human annotations of a subset of a custom corpus of Wikipedia articles about Cold War topics using Amazon's Mechanical Turk and analyzing topic models of that corpus. In the results section, we present our findings, which include a non-significant correlation between the human and computational bias detection methods. In the discussion, we explore how our parallel data collection techniques may have detected different types of bias and thus would not necessarily be correlated. In addition, we discuss opportunities for future work, including the possibility of using topic modeling as a complement to human bias detection.

Chapter 2: Literature Review

Our study bases its methodology on findings from multiple fields. First, we reviewed literature regarding Latin America during and after the Cold War period to understand the prevalent views during that era. Second, we investigated how bias manifests in Wikipedia and its interaction with the site's content policies. We then narrowed our review towards specific studies on cultural bias in Wikipedia, and prior computational methods used to conduct studies on the Wikipedia corpus.

2.1. The Cold War

During the Cold War, the ideological conflict between Western capitalist and Eastern communist countries in the mid-20th century, Latin America was an area where both sides attempted to exert their influence. The United States of America (USA) and the Union of Soviet Socialist Republics (USSR) both had foreign policies that were invested in the aid and preservation of, respectively, anti-communist and leftist governments throughout Central and South America. US-backed actions such as the Bay of Pigs Invasion in Cuba, the installation of Augusto Pinochet as dictator in Chile, and the funding of right-wing Contras as they fought leftist Sandinistas in Nicaragua are some of the more high profile events of the Cold War in Latin America. The USSR also exerted influence in the region, especially through their client state Cuba. Notably, the installation of Soviet missiles in Cuba led to the Cuban Missile Crisis in 1962. The USSR also served as an economic force in the region, selling arms to Peru, importing food from Argentina, and generally increasing trade with Latin American countries from $124 million to $4.9 billion from 1970 to 1981 (Sanchez, 2010).

Historians are divided on how these events affected the attitudes of Latin Americans towards the superpowers. In the case of attitudes toward the United States, some historians believe that a history of political and economic wrongs has created a general sense of resentment, while others assert that despite these wrongs, many see the United States and its cultural exports as representative of greater opportunity (Baker & Cupery, 2013). The goal of analyzing Wikipedia articles is to gain an additional way of understanding these attitudes beyond what traditional analysis of scholarly sources (i.e. the close reading of primary and peer-reviewed secondary sources) can discover. Historical research has been conducted in the same way for a long time: there exists a certain mythos around the "lone historian," spending hours poring over volumes by himself or herself. Many view this as one of the only effective ways to study history (D. Sartorius, personal communication, September 21, 2015). Historians observe that reviewing individual documents gives a wealth of information and provides for easier critical analysis, but is also slow and time-consuming, whereas a computer program can analyze a huge number of documents, but cannot perform in-depth critical analyses (D. Sartorius, personal communication, September 21, 2015).

In the field of US-Latin American relations, historians have found that works analyzing foreign policy focus heavily on the US perspective. Even though the combined populations of just Mexico and Brazil are almost as large as that of the US, in 88.9% of published works in the field of foreign policy history, the focus is on US foreign policy rather than Latin American foreign policy and its effects (Bertucci, 2013). As Wikipedia is edited by users from all over the world, it can provide perspectives on this subject that one might not ordinarily encounter.

2.2. Bias and Wikipedia

Wikipedia was launched on January 15, 2001 as a companion to Nupedia, a free online encyclopedia generated by users (Rosenzweig, 2006). Wikipedia differed from both Nupedia and traditional encyclopedias because it could be edited freely by nearly anyone, not only experts. Wikipedia quickly overtook Nupedia in size and popularity: by the end of its first year, the English-language edition contained around 17,000 articles ("Wikipedia", 2015). On September 9, 2007, the English language Wikipedia surpassed 2 million articles, making it the largest encyclopedia of all time ("Wikipedia", 2015).

Spanish-language Wikipedia was created in June of 2001, but by the end of the year it only included 217 articles ("Wikipedia: Multilingual Statistics (2001)", 2006). Today, it includes over 1.2 million articles, compared to the English Wikipedia's 5 million ("Wikipedia: Estadísticas", 2015; "Wikipedia: Statistics", 2015). The Russian-language Wikipedia was created in May of 2001 and today contains over 1.2 million articles as well. In 2015, it became the sixth largest Wikipedia by number of articles ("Wikipedia: Russian Wikipedia", 2015).

All language versions of Wikipedia are primarily composed by independent authors. Although articles are occasionally human-translated versions of pages in other languages, Wikipedia's official policy strongly discourages machine translation, instead preferring that no article exist until a person can write or translate it ("Wikipedia: Translation", 2015). This policy of preferring user-generated content, as well as the difficulties inherent in translation, may lead to content variations across language editions and thus creates the possibility of relative bias between language editions.

At the same time, Wikipedia's content policies include three central guidelines to prevent such bias: no original research, neutral point of view, and verifiability ("Wikipedia: List of policies and guidelines", 2015). These three policies are designed to maintain the validity of the encyclopedia and provide standards for how the numerous editors should add information. The verifiability and no original research guidelines both state that all content on Wikipedia must originate from a reliable, published work. Verifiability also mandates that in certain cases, such as when using direct quotes or presenting controversial claims, the editor must cite specific sources that verify these claims ("Wikipedia: Verifiability", 2015; "Wikipedia: No original research", 2015). The purpose of the neutral point of view policy is to reduce bias in Wikipedia articles; the policy states that articles should attempt to be as bias-free as possible in both the diction and sources that are used. Likewise, in articles about controversial or disputed topics, an effort is made to provide details on every meritorious viewpoint ("Wikipedia: Neutral point of view", 2015).

Recasens, Danescu-Niculescu-Mizil, & Jurafsky (2013) analyzed the types of bias present in Wikipedia articles. By studying 100 examples of Wikipedia edits designed to preserve a neutral point of view, the authors were able to identify two prevailing types of bias: framing bias and epistemological bias. Framing bias refers to the use of subjective words or phrases linked to a particular viewpoint, while epistemological bias refers to linguistic features that focus on the believability of an assertion (Entman 1993, Boydstun

2013). The authors note that in their sample, changing the word “McMansion”, a negatively connotated term for a large house, to “home” was an example of an edit that removed framing bias. Similarly, changing the word “claimed” to “stated” when referencing the assertion of an author was an example of an edit that removed epistemological bias. Identifying and removing these two types of bias is a task that human editors are well suited for.

However, these content curation policies are not without issue. The nature of Wikipedia’s user-curated content can allow for ‘edit wars’, in which individual or groups of users who disagree on an article’s content frequently try to change the article to fit their narrative (Sumi et al., 2012). The more persistent or larger group can oftentimes drown out the opposing viewpoint. This can violate NPOV and is a demonstration of how careful content policies can be exploited in a user-curated system such as Wikipedia. Likewise, content policies are designated on the granularity of an individual article, and do not control for what content is present across language editions. Hecht & Gergle (2010) conducted an analysis looking at what topics and concepts are present across language editions and found that over 74% of topics are only described in one language edition. This is an apparent violation of Wikipedia’s claim to neutrality and thus informs our decision to use observations of Wikipedia across multiple language editions as the basis for our study.

2.3. Cultural Bias in Wikipedia

Wikipedia has been the focus of several cultural studies because of unique features such as its number of individual contributors, how its content is edited, and its many independent language editions. A good deal of this work has been examinations of differences between language versions of Wikipedia. Computational data analysis has been a large part of this effort, since it allows a larger volume of analysis than human-centric methods.

Pfeil, Zaphiris, & Ang (2006) analyzed culture on Wikipedia using Hofstede's four dimensions (power distance index, individualism vs. collectivism, masculinity vs. femininity, uncertainty avoidance index), a well-known model of cultural norms. The authors hypothesized that there would be significant differences in the number and type of edit actions taken on different language versions of Wikipedia and that these would correlate with the four dimension scores of the countries that correspond to those languages. For this study, it was assumed that French language Wikipedia reflected the culture of France, German language Wikipedia reflected the culture of Germany, Japanese language Wikipedia reflected the culture of Japan, and Dutch language Wikipedia reflected the culture of the Netherlands. The authors had specific hypotheses about how editing actions would be affected by countries' scores on the four dimensions. Many of these were borne out by the data, supporting the idea that there are quantifiable cultural differences across language versions of Wikipedia, and that these language versions can be roughly correlated to cultures. We make a similar assumption in our methodology, correlating English, Spanish, and Russian Wikipedia articles with cultural biases of the USA, Latin America, and USSR.

Hecht & Gergle (2009) investigated whether language versions of Wikipedia exhibit quantifiable bias. Specifically, they looked for "self-focus bias," defined as bias that occurs when Wikipedia contributors in one language encode information that they feel is important and correct, but which may not be considered important or correct to contributors in other languages. To find this bias, they examined Wikipedia as a graph with articles as nodes and links between articles as edges. An article was defined as having greater focus if many other articles contained links to it, while a topic had greater focus if there were more articles about it (here the word topic is used in a general sense, not related to topic modeling). The researchers found that different language versions of Wikipedia have a high degree of self-focus bias, showing that there is bias on Wikipedia, despite its efforts at neutrality.

Another study by Hecht & Gergle (2010) focused on diversity of information presented across language versions of Wikipedia. The data structures of Wikipedia assume that encyclopedic world knowledge is largely consistent across languages and cultures, and the researchers tested this assumption by examining differences in which concepts are presented and what information about those concepts are presented in different language versions. They aligned concepts across languages to analyze how these languages differed in conceptual (article level) and sub-conceptual (content level) information coverage. Concept coverage was not uniform across languages, and sub-concept diversity had a mean overlap coefficient of only 41%, again showing quantifiable cultural differences on Wikipedia's language versions.

Callahan & Herring (2011) performed a case study in cultural differences between Wikipedias of different languages by contrasting English and Polish articles on famous people. Their hypotheses were that systematic biases would be present in English/Polish versions of articles about famous persons and that "local heroes" (people from countries represented by English/Polish languages) would have more content and more favorable coverage on their respective language version of Wikipedia. The researchers examined articles about 15 famous Americans and 15 famous Poles (30 articles in each language), looking at structural characteristics such as length, outlines, lists, references, pictures, links, as well as thematic content such as favorableness of coverage, mentions of personal information, education, nationality, ideology, controversy. Callahan & Herring concluded that English articles about Americans do, in fact, reflect American cultural values and history, and Polish articles about Poles reflect Polish history and culture, further confirming that Wikipedia has a cultural bias across languages.

An exploration of bias in one language version of Wikipedia by Greenstein & Zhu (2012) further confirms that articles do not usually have a completely neutral point of view. Examining political articles in English Wikipedia, the authors used a technique described by Gentzkow & Shapiro (2010), which used a list of phrases from the 2005 Congressional Record to estimate the political (Democrat/Republican) slant of newspaper articles relative to each other, based on their text's usage of the Congressional Record phrases. This algorithm performed well on those articles, identifying coded political language that human annotators might have missed. Articles in most political categories surveyed did have a mean slant, or directional measure of bias towards either the conservative or liberal baseline. Through exploring their edit histories, Greenstein & Zhu found that the editing process Wikipedia relies on to ensure neutral point of view did not perform as expected; most articles had a slant at their inception, and this did not change with time and edits.

Two 2015 studies again looked at how famous or notable people are represented in different language Wikipedias. Eom et al. (2015) looked at the birth date, birth country, and gender of top historical figures on different Wikipedias to see how temporally, spatially, and gender skewed those Wikipedias are. They found that in addition to the top historical figures skewing Western, male, and post-17th century, most countries' local figures were more prominent in that country's associated language Wikipedia. Gloor et al. (2015) examined articles on prominent people in English, Chinese, German, and Japanese Wikipedias by creating networks of contemporary individuals using links between articles. Those with most links for a particular time period were labelled the most prominent. Across languages, there were differences in what kind of person (politician, artist, scientist, religious leader) was usually more prominent and whether they were local to the language being examined.

2.4. Computational Approaches to Bias Detection

Prior studies offer multiple approaches to the bias identification problem posed by Wikipedia. Michel et al. (2011) analyzed a corpus consisting of 4% of all digitized English language literature using n-gram frequency to see culture and lexicon shifts over time. The authors of the study were able to identify groupings of n-gram usages to this purpose, especially with regard to ones that had significant usage over time trends. Two important takeaways from the study are that first, the relative fame of celebrities grows and declines predictably and at a more rapid pace over time; and second, censorship can be uncovered by comparing the popularity trends of terms from geographically close but politically distinct geographic regions. This trend analysis would also be useful for identifying topics relevant to a cultural identity within a selection of text. We could be reasonably confident that if we chose a section of Wikipedia that covered events that were not current, we could obtain a selection that contained moderate cultural bias as opposed to biases reflecting volatile current events. If we chose past events that did not contain modern controversy, then we could be more confident in the language localization. Our selection of the Cold War as our case study was motivated by how there was a clear defined end to the conflict and an abundance of material covering it.

Additionally, these findings suggested that we could identify bias through comparisons of content density, which informed the primary approach to our study. In order to achieve these comparisons, we looked into methods that would allow us to explore the distribution of content among languages at a more granular level than articles.

We were especially interested in the parallel nature of Wikipedia's multilingual corpus, which could be exploited using multilingual topic modeling (Mimno et al., 2009). This study was primarily focused on investigating if multilingual (or polylingual) topic modeling could help in machine translation, but it also discovered that topics were distributed unevenly over different languages due to cultural factors. To find topic models, we took interest in Latent Dirichlet Allocation (LDA) (Blei, 2013), which has been used for topic inferencing by researchers in single and multilingual contexts (Hoffman, 2010; Ni, 2010; Yurochkin, 2016). A related method, Latent Semantic Indexing, has been used for cross-language retrieval of documents based on a set of parallel multilingual documents (Dumais et al., 1997). LDA works under the assumption that each document in a corpus contains multiple topics in different proportions, these topics are distributed over multiple documents, and that we know the number of topics in the corpus ahead of time. The topic distribution is generated by iterating through every word in the vocabulary and putting it in a topic that benefits most from having its probability distribution represented by that word. Through multiple iterations, we should ideally arrive at a collection of coherent topics as well as a few incoherent ones. The benefit of LDA is that it does not need to understand what words in a vocabulary mean; in fact, it can be applied to any object as long as the object is distributed similarly to how words and topics are assumed to be in a multi-document corpus. A topic is a distribution over the vocabulary of words; oftentimes the most prevalent words in a topic have some identifiable thematic connection, like "music", "bands", "song" (Blei, 2013). The output of LDA is a list of topics and a vector of percentages for each document showing the proportions of topics that constitute the document. Topic modeling is a good basis for multilingual bias detection for several reasons. It is a "bag of words" approach, meaning it does not require any knowledge of the semantics of the text (Blei, 2013). Also, topics often reflect major themes within a corpus, so topic modeling is a good first step to becoming familiar with a corpus on a qualitative level as well (Chang, et al., 2009). As we are dealing with a multilingual corpus, this allows us to bypass complications due to translation and usage nuance. These features make multilingual topic modeling a reasonable candidate for bias detection in large text corpora.

Chapter 3: Methodology

In this section we outline our methodology for analyzing bias in Wikipedia articles. We start by explaining our three-viewpoint model and how we assembled our corpus. Then, we discuss how we collected human annotations and topic modeling data. Finally, we explain how we evaluated the results from the two approaches.

In order to examine the bias of one Wikipedia language edition with reference to two others, we utilized a three-viewpoint model. Two of the three viewpoints, the spectrum viewpoints, are used to establish endpoints of a scale. For instance, in our case, these endpoints are the Russian (USSR) and English (US) viewpoints. Past research has used speeches by Democratic and Republican congresspeople in a similar way (Gentzkow & Shapiro, 2010, Greenstein & Zhu, 2012). We can then classify a third viewpoint, the target viewpoint, in relation to the two spectrum viewpoints. In our case, the target viewpoints will come from the Spanish language Wikipedia. In previous research, the target viewpoint (with congressional speeches as endpoints) has been represented by US newspapers (Gentzkow & Shapiro, 2010) or English Wikipedia articles (Greenstein & Zhu, 2012).

3.1 Creating a Corpus

To obtain training data and text for analysis and human annotation, we began by generating a corpus of Spanish Wikipedia articles pertaining to the Cold War. Wikipedia has a hierarchical structure of categories, or main topic classifications, as shown in Figure 1b, that we leveraged to produce a corpus. We manually selected 89 categories under the Guerra Fría (Cold War) category to include in the corpus, and selected every article in those categories, resulting in 1133 unique articles. We then expanded the corpus to include the corresponding articles in English and Russian by writing a script that used Wikipedia's interlanguage links feature. Interlanguage links are "links from a page in one Wikipedia language to an equivalent page in another language" ("Help: Interlanguage Links", 2017), as shown in Figure 1a. For instance, we combined the Guerra Fría, Cold War, and Холодная война articles. Not all Spanish articles had parallel versions in the other two languages, so after removing these, our corpus was reduced to 1021 articles in each language.

Figure 1a, 1b: Left: An example of Wikipedia's interlanguage links that appear in the left sidebar as a list of languages, Right: An example of the category structure of Wikipedia articles

Because Wikipedia articles can be very long, we split each article in the corpus into 8078 “chunks” in order to make them easier for humans to annotate. These chunks were created using an algorithm that attempts to keep text at a similar length (about 180 words) without breaking up paragraphs and sections across multiple chunks. We chose 180 words as the cutoff length because it represents the length of an average readable paragraph.

Once we had assembled our corpus, we proceeded with identifying and quantifying bias within it. This phase of the project contained two parts which were completed simultaneously. Annotators, sourced from Amazon’s Mechanical Turk, manually inspected parts of the corpus, looking for multiple types of bias (imbalances in information, topic selection, and sentiment) and ultimately assigning bias scores for articles along the three-viewpoint model. In parallel, we built a computer algorithm using multilingual topic modeling and regression methods to quantify bias in the corpus along this spectrum. Once both parts were completed, we compared the results from each to test our hypothesis that computers can detect and quantify bias using multilingual topic modeling.

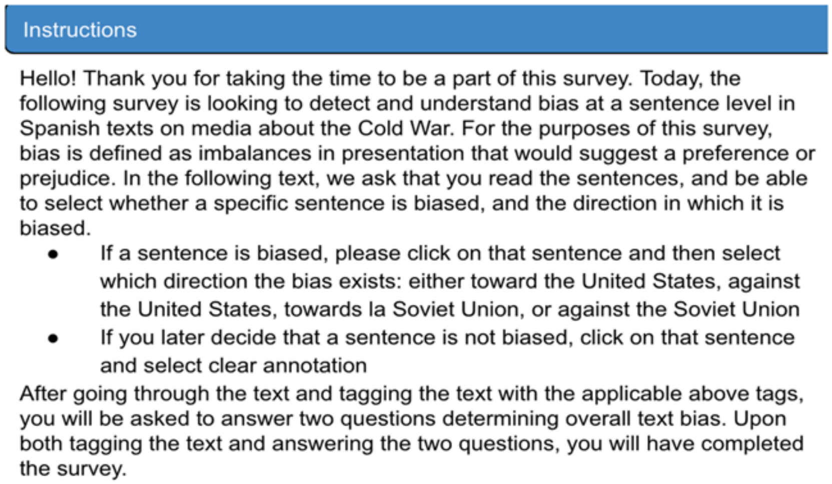
3.2 Human Annotation of Bias in Wikipedia Articles

Yano et al. (2010) demonstrated that Amazon Mechanical Turk workers are capable of detecting bias in political blogs. We took a similar approach, enlisting human annotators to study how humans perceive bias in articles about history. We used Amazon’s online Mechanical Turk marketplace. This service allows users, called Requesters, to list tasks on the website along with a monetary reward for completing the task, and lets other users, called Workers, perform the task. Mechanical Turk specializes in Human Intelligence Tasks (HITs), which are tasks that generally require little skill but would be difficult for an automated system to perform (Mechanical Turk, 2015). We utilized Mechanical Turk to find annotators to complete a HIT requiring the annotation of a one paragraph chunk of text (from our chunking algorithm) in Spanish. Our Workers spent an average of 3 minutes and 6 seconds to complete the task and were paid 35 cents per annotation completed. This comes to roughly $6.77 per hour, which is 93% of the federal minimum wage of $7.25 per hour. It is important to treat crowdsource workers ethically because crowdsourcing has become common in scholarly work, and many of the workers treat this work as their primary source of income (Williamson, 2016). Our payment of the Workers was ethical through how our hourly rate is much closer to the federal minimum wage than the typical rate of $2 per hour (Ross et al., 2010). The payment was increased to 50 cents for each annotation with the average completion time being 3 minutes and 32 seconds.

To ensure that our use of Mechanical Turk met the best practice standards for research, we completed an Institutional Review Board (IRB) Human Subject Determination Research Form through University of Maryland's Research Compliance Office ("IRB process", 2015). Our research was ultimately exempt from needing further IRB approval as we did not ask for the personal information of any Worker.

It was necessary to control for the quality of annotations when using Mechanical Turk. We created a qualification test to ensure that Workers could understand written Spanish and identify bias related to the Cold War. The test consisted of two short texts and two Likert scale questions. We wrote the texts to contain explicitly biased phrases. Workers who scored at least 9/15 on the directional bias (denoting which viewpoint the phrase showed bias toward/against) section and were within two numbers on the Likert scale test were allowed to continue on to complete the HITs.

We used a manual annotation of a sample of articles from the corpus in order to establish a human reference for the quantification of bias. A random sample was taken from the larger corpus and used as the annotation corpus. In order to produce results comparable to the computational methods, the annotations identified the bias through studying directional bias and overall perceptions of bias in the texts.


Instructions

Hello! Thank you for taking the time to be a part of this survey. Today, the following survey is looking to detect and understand bias at a sentence level in Spanish texts on media about the Cold War. For the purposes of this survey, bias is defined as imbalances in presentation that would suggest a preference or prejudice. In the following text, we ask that you read the sentences, and be able to select whether a specific sentence is biased, and the direction in which it is biased.

- If a sentence is biased, please click on that sentence and then select which direction the bias exists: either toward the United States, against the United States, towards la Soviet Union, or against the Soviet Union
- If you later decide that a sentence is not biased, click on that sentence and select clear annotation

After going through the text and tagging the text with the applicable above tags, you will be asked to answer two questions determining overall text bias. Upon both tagging the text and answering the two questions, you will have completed the survey.

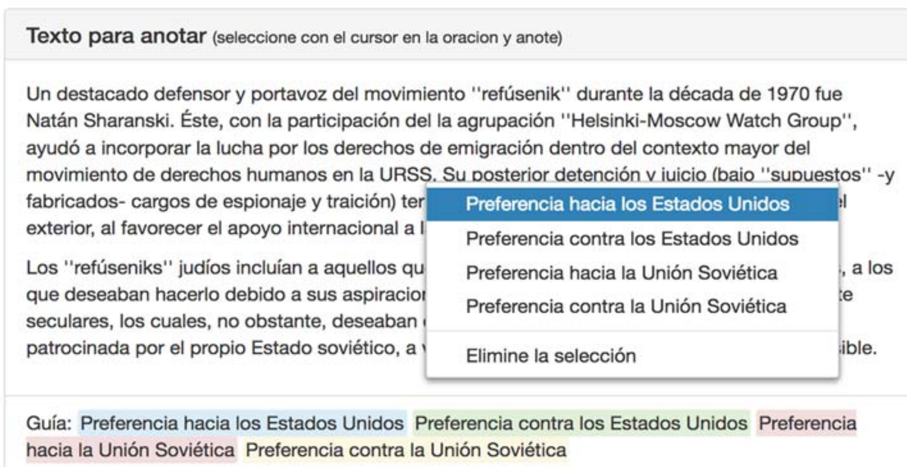


Figure 2a, 2b: Top, Mechanical Turk survey instructions (translated to English). Bottom, the interface for annotating chunks of text from Spanish Wikipedia. Workers could choose four tags from a drop down menu for each sentence (bias towards/against the US/USSR), or choose not to tag it if it didn't contain bias.

As shown in Figure 2a and b, annotators were asked to read a chunk of text in Spanish, consisting of around 180 words. Annotators submitted two types of annotations. First they annotated each sentence with one of four bias tags: towards the US, against the US, toward the USSR, or against the USSR. The annotators could also choose not to tag the sentence if they perceived it as neutral. Second, they were asked to provide two overall bias scores for the entire chunk, each of which was marked on a seven point Likert scale as shown in Figure 3. One bias score considered the chunk's overall bias towards or against the United States, and the other considered the chunk's overall bias toward or against the USSR.

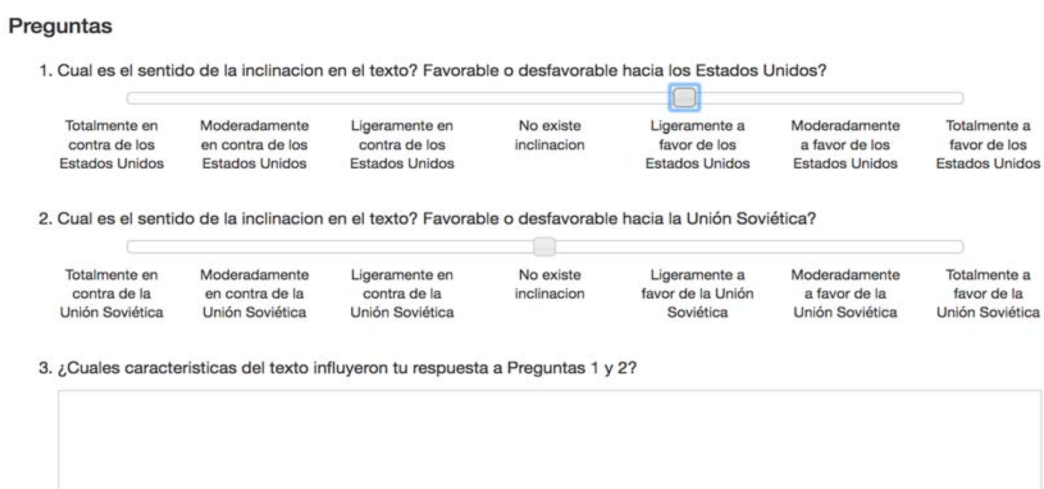


Figure 3: The Likert scale questions following the chunk sentence tagging part of the survey

These Likert scale questions ask "How biased was the text towards or against the United States?" offering a range of 7 options: very biased against, moderately biased against, slightly biased against, not biased, slightly biased towards, moderately biased towards, and very biased towards. The next question asked, “How biased was the text towards or against the Soviet Union?” offering the same Likert scale of 7 answer options.

As our research progressed, we still did not understand clearly why Workers were indicating phrases as biased, as our survey questions focused more on where Workers were finding bias, but not why they classified something as biased. Because an understanding of why humans perceive text to be biased is fundamental to our research, we decided to run a second round of annotations with free form response questions to better understand what made annotators denote something as biased. These questions allowed Workers to identify specific words or phrases that informed their opinions on the biases of texts. The questions were: “What characteristics of the text influenced your response to Questions 1 and 2?”, “Which phrases, if any, influenced your response to Questions 1 and 2?”, “Which words, if any, influenced your response to Questions 1 and 2?”, and “Was there an imbalance of information which influenced your response to Questions 1 and 2?”.

Each chunk was annotated by no fewer than three annotators to control for the biases of individual annotators, as three non-expert annotations have demonstrated very similar accuracy levels to expert annotations (Snow et al., 2008). After each annotator completed their analysis, the results were compiled using various statistical methods to minimize human error. Once the results were compiled, we analyzed them using statistical methods for comparison against computer-generated results.

3.3 Automatic Detection of Bias in Wikipedia Articles

In this phase, we built a system to automatically quantify bias along the three-viewpoint model's spectrum, basing its computational model of bias on a multilingual topic model. Using multilingual topic modeling allowed us to condense large multilingual corpora into a set of "universal" topics that are independent of language. This approach has been shown to work well on multilingual classification problems (Ni et al., 2011). Thus, we trained a topic model over a parallel corpus that consisted of all Spanish, English, and Russian Wikipedia articles. We then used a multinomial logistic regression model to compare topic distributions within articles of the target language (Spanish) to topic distributions within articles of the two spectrum languages (English and Russian) to determine where they fell on the bias spectrum, shown in Figure 4.

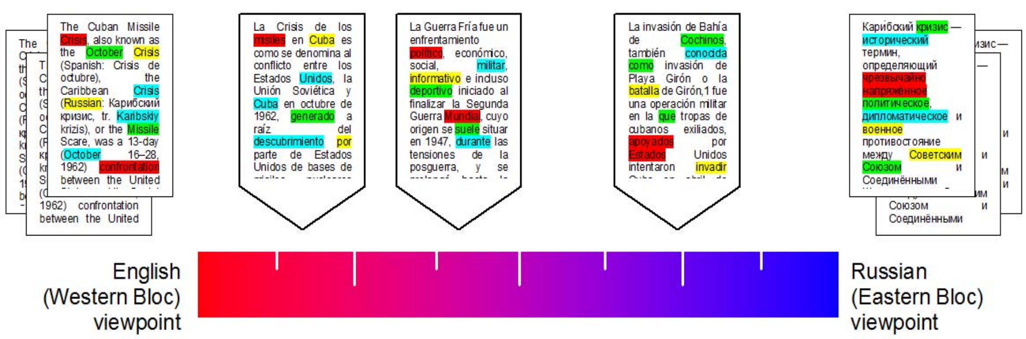


Figure 4: Our research attempts to place Spanish Wikipedia articles related to the Cold War on an American-Soviet spectrum of bias. We assign bias a numerical value from -1 (a completely American viewpoint) to 1 (a completely Soviet viewpoint).

Multilingual Topic Modeling

Topic modeling refers to a group of algorithms that take a corpus of texts and find their underlying topical structure as well as the degree to which each topic appears in each article (Blei, 2013). For our purposes, the inputs are the text in selected articles on all three language versions of Wikipedia, and a few parameters, such as the number of topics, that most topic modeling algorithms require to be adjusted. Probabilistic models such as LDA find the latent, or hidden, structure of topics in the entire corpus, as well as the amount each topic is present in an article (Blei, Ng, & Jordan, 2003).

Topic modeling works best when given a large corpus of text as an input, so we decided to train the topic model over a corpus taken from as much of Wikipedia as possible. We used a multilingual topic model, inspired by the methodology in Jagarlamudi and Daumé (2010). This method required that we create a corpus of Wikipedia articles that appear in all three languages. Ni et al. (2011) named these related groups of Wikipedia articles “concepts”; for example, the “Cuban Missile Crisis” pages in English, Spanish, and Russian would constitute one concept. Using Wikipedia’s interlanguage links as previously described in Section 3.1, we assembled a corpus of every article that appeared in English, Spanish, and Russian, resulting in 1,058,571 total articles (and therefore 352,857 “concepts”). Following Ni et al.’s approach, we then concatenated each article’s English, Spanish, and Russian versions into single documents, giving us 352,857 documents. We trained our LDA topic model over this larger corpus of multilingual documents.

If parallel articles in the three corpora were treated as separate documents, the topic model would generate monolingual topics, because the words of each language would always appear together; i.e. without combining articles from different languages, about a third of the topics would be collections of Spanish words, a third would be English words, and a third would be Russian words. To avoid this, we grouped articles by “concept” and concatenated them into single documents, so that words related to the concept in each of our three languages appeared in proximity to each other. The resulting topic model produced topics that included words in all three languages, as seen below in Figure 5. This “language agnostic” topic model gave us a unified way to describe the content of articles from any one of the three Wikipedia editions.

Possible Interpretation	**Words <language>#<word>**				
Armed forces	en#army	ru#войск	es#ejército	en#forces	ru#cc
Argentine military	en#military	en#argentina	es#argentina	en#argentine	es#militar
Soviet Union	es#soviética	es#unión	en#soviet	en#union	es#república
KGB	es#kgb	en#kgb	ru#кгб	es#seguridad	ru#безопасности
Elections	es#elecciones	en#party	es#partido	ru#партии	es#votos

Figure 5: Examples of top words from multilingual topics and a possible corresponding interpretation.

After creating these trilingual documents, our method followed a traditional procedure for LDA; we processed the multilingual corpus, created bag-of-words representations for each document, and ran the LDA algorithm using 21 passes. To run the LDA algorithm, we used the Gensim package, which implements LDA as described in (Hoffman, 2010). We trained the LDA topic model to produce 100 topics. The LDA model’s output consisted of a list of 100 topic vectors latent to the combined corpus. Each topic vector appeared as a list of words and their associated probabilities of being included in that topic. While each topic in theory includes every possible word that appeared in our corpus, the associated probabilities can be interpreted as weights, indicating which words are most important to that topic. From those most important words, often times we were able to give a description to a topic (for instance, the first topic in Figure 5 could be called "armed forces"). With this LDA model, we could produce 100-element topic vectors from any text in one of our three languages. For instance, given a Wikipedia article, our LDA model would tell us which of the 100 topics it expected to find in the article, and in which amounts.

The multilingual topic model allows us to represent an article from any of the three languages as a normalized vector whose entries correspond to the topic composition of the article. In Figure 6, those topic distributions are represented as pie charts for each language version of the same article.

Figure 6: An example of the different topic distributions, represented as pie charts, for an article in English, Spanish, and Russian.

We assumed that articles in the English Wikipedia would tend to describe a American viewpoint and that articles in the Russian Wikipedia would tend to describe a Soviet viewpoint. Thus, we established a spectrum from 0 (a completely American viewpoint) to 1 (a completely Soviet viewpoint) and assigned all English articles in the Cold War corpus a label of 0 and all Russian articles a label of 1. We then used the topic composition of each of these articles to train a logistic regression model to predict an article's bias.

Logistic Regression

Finally, to calculate a bias score for a target document (a Spanish Wikipedia article), we calculated its topic composition using the LDA topic model and then predicted its bias score using a multinomial logistic regression model. In this multinomial logistic model, we assigned a score of 0 to the topic distributions of English articles and a score of 1 to topic distributions of Russian articles. This response variable fixed the two ends of our spectrum as English and Russian, per the three-viewpoint model. Then, a multinomial logistic regression was trained using the 100-element topic distribution vectors as the features, and the attached scores of 0 or 1 as the desired output depending on if a given topic vector belonged to an English or Russian article. Thus, the bias score represents the probability that a given article's topic distribution a resembles English Wikipedia more versus Russian Wikipedia.

Once the logistic regression was trained, we inputted the topic distributions of Spanish articles into the model to determine which corpus the Spanish article resembled the most, the English or Russian. The resulting prediction was a number on a scale from 0 to 1 that could predict if the topics of Spanish article were more similar to the Russian corpus or the English corpus. An advantage to this approach was that the combination of the topic model and logistic model could be used on any text in Spanish (and to a lesser extent English and Russian), meaning that the bias-recognition abilities of the algorithm could be expanded to texts outside of Wikipedia.

Evaluation of Results

After collecting the data from Mechanical Turk, we analyzed the bias scores given by annotators to determine whether the human annotations of the Spanish articles were correlated with the results from the topic modeling approach. To do this, we generated a single bias score for each article. This bias score ranged from 0 to 1, with 0 indicating complete bias towards the US and 1 indicating complete bias towards the USSR. This score from 0 to 1 mirrored the score generated by the topic modeling system, allowing us to easily compare the results and generate a Pearson's rho correlation coefficient.

First, we generated a normalized average bias score for each chunk. The respondents' answers to the two Likert scale questions, shown in Figure 3, asking about a final determination of bias towards or against the US or the USSR were averaged between the all annotators of a specific chunk. To normalize the averaged 7-point Likert scale and to generate a bias score comparable to the one generated by the automatic system, which ranges from 0 to 1, we combined the two questions asking about bias towards/against the United States and towards/against the USSR using the following formula:

$$B = \frac{b_{USSR} - b_{US} + 6}{12}$$

In this formula, B is the normalized bias score, b_{US} is the average of annotators' Likert scale rankings of bias towards or against the US, and b_{USSR} is the average of annotators' Likert scale rankings of bias towards or against the USSR.

To evaluate the results of the regression models, we used a statistical correlation analysis to compare the results of the topic modeling to those of human annotators. We also did a simple analysis of the inter-annotator agreement to determine if there was a significant difference in how annotators viewed an article. By calculating Fleiss' kappa for each chunk that was annotated, we were able to approximate how closely the annotators' bias scores matched.

The qualitative responses from the second round of annotations were coded based on whether or not the annotator used topical information, semantic information, or other information to explain why they perceived bias in the text. Only the responses which indicated that the text was biased were used for coding, as it was assumed that they would be the only free response questions to contain meaningful responses. Responses which included the phrase "did (not) include" were considered a topic-based decision. Responses which included "seems/feels like," said something was emphasized, listed specific words/phrases, or listed information that contained potentially emotionally-charged words (e.g. "treason") were considered a semantic-based decision. Responses which did not include any of this were considered to be "Other." Blank responses and answers indicating no bias were omitted.

Chapter 4: Results

The final goal of our methodology was to examine the correlation between our computerized results and our human annotation scheme, but our research was designed so that each step would itself produce results that would further our understanding of bias on Wikipedia. We present results pertaining to our custom corpus, human annotations, topic models, and logistic regression here, in the order that they appeared in the methodology.

4.1. Corpus Creation

We selected 89 categories for our corpus. This resulted in 856 articles that appeared in all three languages in our corpus (a list of these can be found in Appendices 2 and 3).

4.2. Analysis of Survey Results [1]

We conducted two rounds of soliciting annotations of one paragraph chunks (about 180 words) of Spanish Wikipedia articles. In the first, we solicited annotations on Mechanical Turk for a total of 45 chunks, and each chunk was annotated by four to five Workers. An overview of the annotation results can be seen in Tables 1 and 2, below. In Table 1, we see that most individual sentences were tagged as neutral, a finding consistent with a later survey (Table 3). The second most prevalent tag was of sentences perceived to be "against the Soviet Union." We hypothesize that this was because most selected articles dealt with subject matters closer to the USSR.

---

[1] Our results can be found at this link: **https://goo.gl/PXR9pA**

Tag	Frequency
Neutral	82.8%
Towards the United States	0.4%
Against the United States	0.4%
Towards the Soviet Union	4.7%
Against the Soviet Union	11.7%

Table 1: First round survey results; distribution of human-annotated tags.

Bias	US	Soviet Union
Completely against	0.0%	4.4%
Moderately Against	0.0%	11.1%
Slightly against	3.0%	25.2%
Neutral	94.1%	45.9%
Slightly towards	3.0%	8.9%
Moderately towards	0.0%	3.0%
Completely towards	0.0%	1.5%

Table 2: First round survey results; distribution of human-annotated bias scores.

To make sure that our task was well-defined and that Workers were obtaining similar results for the same text, we measured inter-annotator agreement. On average, the Fleiss' kappa values for inter-annotator agreement across all text chunks was .171756. Following Landis and Koch's (1977) interpretation of the Fleiss' kappa statistic, this indicates slight agreement. The slight agreement may be due to the low number of annotators used. We also assigned the annotators' Likert scale scores of a chunk's overall bias to a 7-point scale from -3 (completely against the US/USSR) to 3 (completely towards the US/USSR) and then computed the standard deviation for each chunk; the mean of these standard deviations is displayed in Table 2. Even when annotators didn't agree on a bias score, the low standard deviation indicates that they tended to pick similar scores. For instance, one Worker may have chosen "slightly against" and another may have chosen "moderately against."

For our second round of annotations we solicited more in-depth responses about 9 chunks of text, in addition to the previous questions ranking the bias of the chunks. Quantitative analysis of the responses shows that "no bias" again was the most common response for both the US and USSR questions. The results from the coding of the free-response answers are displayed in Figure 7.

Tag	**Frequency**
Neutral	90.0%
Towards the United States	0.8%
Against the United States	0.8%
Towards the Soviet Union	3.9%
Against the Soviet Union	4.6%

Table 3: Second round survey results: distribution of human-annotated tags

Bias	**US**	**Soviet Union**
Completely against	0.0%	0.0%
Moderately Against	0.0%	4.6%
Slightly against	0.0%	18.2%
Neutral	86.4%	68.2%
Slightly towards	9.1%	4.6%
Moderately towards	4.6%	4.6%
Completely towards	0.0%	0%

Table 4: Second round survey results: distribution of human-annotated bias scores

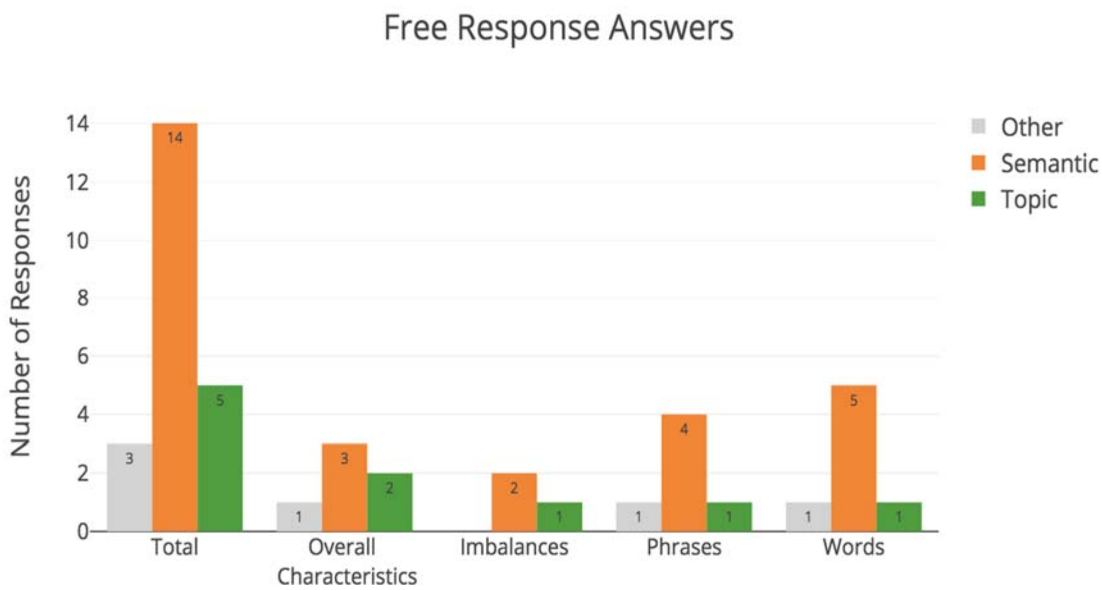


Figure 7: Distribution of free response answers based on type.

4.3. Evaluating Our Topic Model

Each topic in our topic model gives a distribution of how likely it is that a word in our corpus is included in that topic. As it includes every word in our entire multilingual corpus, each topic includes a mix of Spanish, English, and Russian words, each given a particular weight in the topic. We were interested in making sure the topics were truly multilingual, so we summed weights of words in each language. If our topics were balanced, we would expect that each language had a weight of 33.333%.

In order to learn more about the distribution of these words, we used ternary plots to show the weighted distribution of words from each language. In our topic models, each topic consisted of words in English, Spanish, and Russian with certain weights attached corresponding to their importance in the topic model. By summing the weights corresponding to each language, we could determine if any of the three languages was overrepresented in a topic. In Figure 8 below, each data point represents the weighted language composition of a single topic.

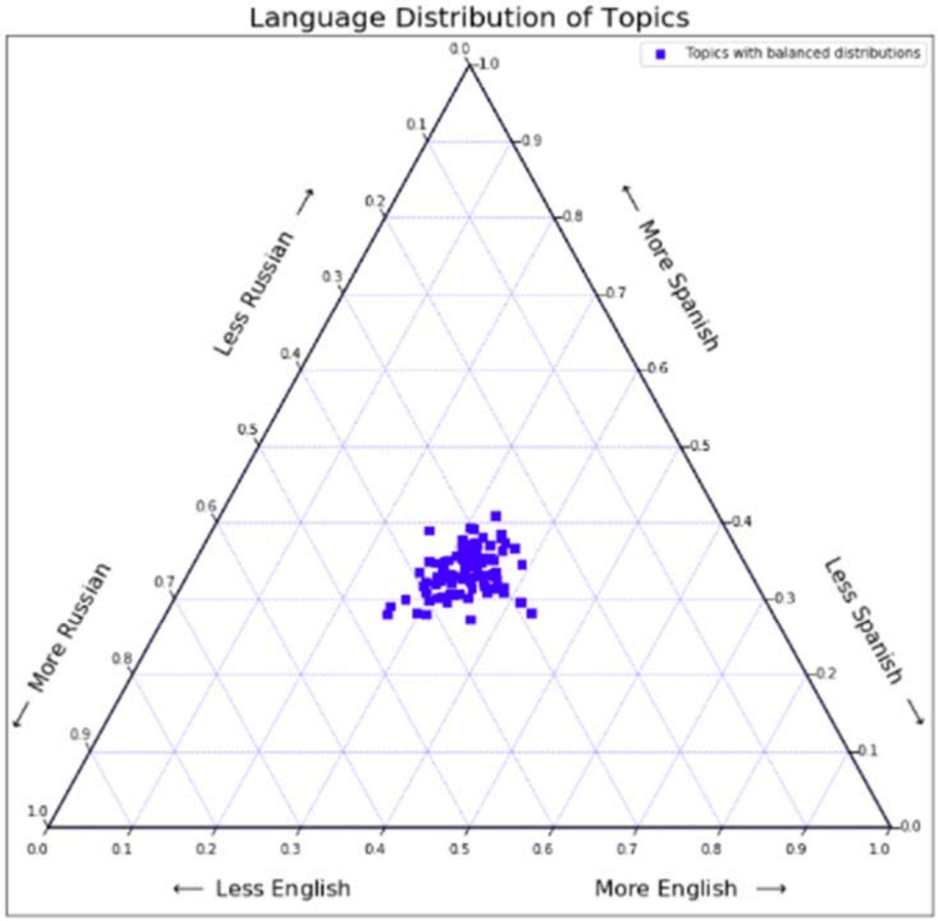


Figure 8: Language distribution of topics

If a topic appears in the middle of the graph, it indicates that words in English, Russian, and Spanish are equally represented in the topic. The graph above demonstrates most of our 100 topics have nearly equal distributions of topics in Spanish, English, and Russian, as evidenced by their proximity to the center of the graph.

4.4. Evaluating Our Logistic Regression

A multinomial logistic regression maps an N-dimensional array onto a range from 0 to 1. Within machine learning, it is used to model membership in two groups: for example, a multinomial logistic regression could provide the likelihood that a person has a disease or not, given other factors about their health. In our algorithm, we assigned English articles a value of 0 and Russian articles a value of 1, and then trained the logistic regression on the topic distributions. The logistic regression used the topic distributions of articles to predict whether an article was more "English" or "Russian." While the structure of this algorithm sounds like it is predicting the language of the article, the input was the topic distributions of various articles and these features were language agnostic.

The multinomial logistic regression model was somewhat effective by this measure. Using k-fold cross validation with k=5, the model was able to predict 72.93% of the articles' origins in our training data. K-fold cross validation is a method of evaluating machine learning techniques, in which 80% of the data (in our case, the Russian and English topic distributions) is used to train an algorithm, and 20% of the data is used to evaluate the algorithm by seeing if the algorithm predicted the results correctly. In our algorithm, a correct prediction means that, given the topic distribution of an article, our algorithm predicted successfully if it was from the English or Russian corpus. This 80%-20% split is repeated five times, and then the final score is averaged. Given the topic distribution of an article in Russian or English, our model could determine whether it came from the Russian or English corpus 72.93% of the time, showing a measurable difference from the null rate of 50%. There were also 4 topics that were not present in any of the English or Russian articles (the topic model was originally trained on the entire Wikipedia corpus, so it was expected there would be some topics which did not apply to the Cold War corpus). These 4 topics were not used as features in the logistic regression, leaving us with 96 features, or independent variables, in our logistic regression. In logistic regression, the coefficient assigned to each independent variable determines how much it contributes to the final classification of “Russian” or “English”. Table 5 displays the five topics which contributed the most to each classification. The full version of this table, with all 95 coefficients ranked, is in Appendix 8. The most important output of this logistic regression were the predicted scores for the Spanish articles that were also annotated by Mechanical Turk workers, as these scores provided an opportunity to compare our computer results to the human reference.

The topics generated were mostly coherent at both the word level and when considering which articles most reflected those topics. For example, the five most “English” topics showed up in articles almost entirely about, respectively, nuclear technology and military strategy, early 20th century Soviet politicians, the geography of the USSR, political popular culture, and inter/intra-governmental organization and diplomacy. The top five most “Russian” topics were reflected in articles concerning actions taken against political dissidents, US and USSR social and economic policies, left-wing politics, inter/intra-governmental cooperation, and Soviet military and police influence on internal affairs.

Topic ID	Topic description	Coefficient
73	*Nuclear technology and military strategy*	-2.892736
74	*Early 20th century Soviet politicians*	-2.812615
37	*Geography of the Soviet Union*	-1.873897
26	*Political popular culture*	-1.688170
90	*Inter/intra-governmental organization and diplomacy*	-1.282192

Table 5: Topics which contributed most to an "English" classification, indicating similarity to the English corpus

Topic ID	Topic Description	Coefficient
40	*Actions taken against political dissidents*	5.532225
83	*US and USSR social and economic policies, and*	4.892671
65	*Left-wing politics*	3.881321
47	*Inter/intra-governmental cooperation*	2.508126
77	*Soviet military and police influence on internal affairs*	2.139590

Table 6: Topics which contributed most to a "Russian" classification, indicating similarity to the Russian corpus

The scores of all 856 articles in the Spanish corpus are shown in the histogram below. The mean is 0.4968 and the median is 0.5085, indicating that our Spanish corpus as a whole is equally biased toward both the English and Russian corpora. The distribution is unimodal, and the standard deviation is 0.1083. While the distribution appears to be approximately normal, it should be noted that it has a skew of -.6170. This statistic corresponds to the longer left tail in Figure 9 below, and means that, while there were roughly equal amounts of articles above and below the score of 0.5, the articles that scored below 0.5 tended to have scores that indicated more bias.

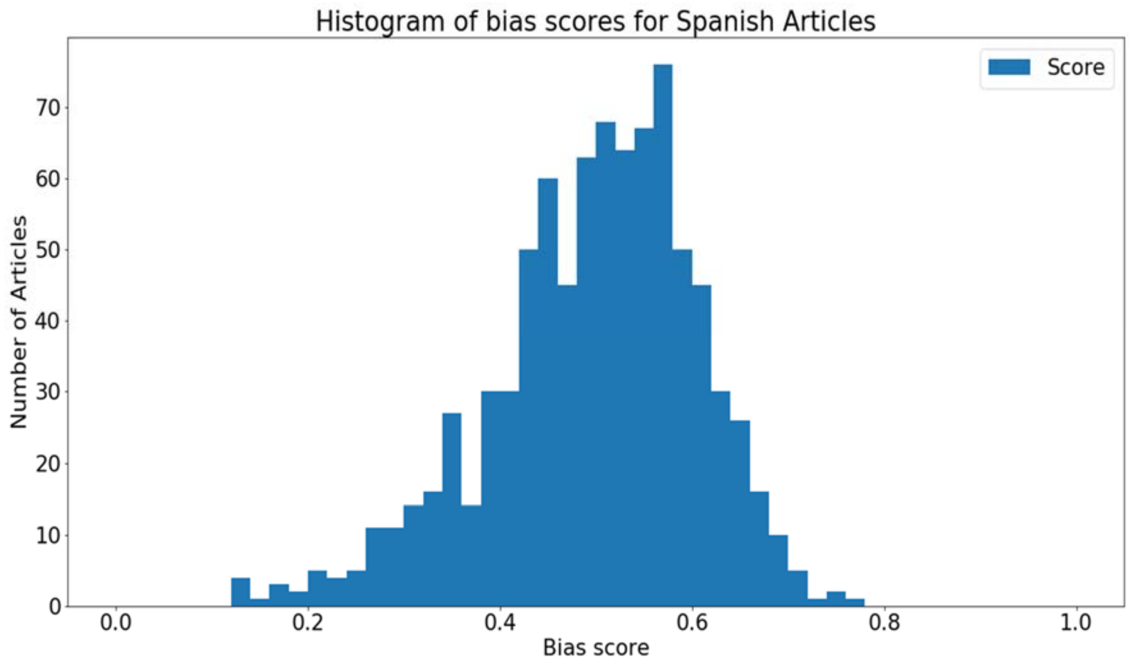


Figure 9: Histogram of Bias Scores for 856 Spanish Articles

4.5. Correlation Between Computer and Human Results

To examine the correlation between the human and computer evaluations, we first normalized the human bias scores (the average of their answers to the Likert scale questions) to a scale of 0 to 1, with 0 indicating bias towards the United States and 1 indicating bias towards the USSR. This allowed for direct comparison between the logistic regression's predicted scores for Spanish articles and the score from the annotators. We assumed that article level topic distribution might be correlated with the human assessment of bias at the chunk level if articles with biased topic distributions also contained bias in sentiment (i.e. a "biased" article might display its bias in more than one way). We obtained a Pearson's coefficient of -0.043 with a p-value of 0.757524, indicating a nonsignificant correlation.

Chapter 5: Discussion

The lack of correlation between human and computer results is indicative of a disconnect between the information that determines how humans perceive bias in a text and the information that our topic modeling approach depends on. An imbalance of topical information in an article does tell us something about the overall bias in an article, but when human annotators analyze smaller portions of those articles, different factors take precedence. The results of our second round of annotations (which included free response questions asking how human respondents determined their bias scores) suggest that individual words and phrases that indicate an author's sentiment are the largest determinant of bias for human readers, at least at the level of paragraph-length chunks of text. Our assumption that the complete articles' topic distribution would match the bias conveyed by these shorter chunks was not borne out, but this type of topic modeling analysis can complement sentiment-based bias identification by providing perspective at a different level than humans can observe (for example, comparing the topic distribution of individual articles to that of an entire Wikipedia, something a human could obviously not do).

In our first round of annotations, we did not request that annotators justify their bias scores. After reading the survey results, we examined the sentences they had scored to find features that could have prompted certain scores. One sentence that four out of five annotators agreed displayed bias against the USSR occurred in an article about the "refusenik" movement of protesters who wished to emigrate from the USSR (the other annotator tagged the sentence as neutral):

“Su posterior detención y juicio (bajo supuestos *-y fabricados- cargos de espionaje y traición) terminó afectando al propio régimen soviético en el exterior, al favorecer el apoyo internacional a la causa* refúsenik.*”*

English translation:

“His subsequent arrest and trial (under assumptions *-and fabricated- charges of espionage and treason) ended up affecting the Soviet regime itself abroad, by favoring international support to the* refusenik *cause.”*

The sentence in question stated that espionage charges against a leader of the Refusenik movement were “fabricado”, or fabricated. Although a very similar sentence occurs in the comparable English article, it does not state that the charges were fabricated and only mentions that the charges existed.

“His arrest on charges of espionage and treason and subsequent trial contributed to international support for the refusenik cause.”

The annotators likely based their tag heavily on the word “fabricated”, which suggests an interpretation of events, otherwise the sentence would have simply stated provable historical facts. This particular case suggests that individual words were more indicative of bias for human annotators, whereas a topic model of the article would most likely not take these words into account unless they occurred relatively frequently.

The second round of human annotations involved the use of free-response questions in which Workers were asked to provide examples of words, sentences, and other factors which may have influenced their decision regarding the bias of the text. Based on the results of this round of experiments, our conclusion that humans tend to rely on semantic information to make their decisions about whether or not a paragraph has bias was supported. These semantic-based decisions were present in answers to all four free response questions (those asking about general characteristics, phrases, words, and information imbalances). One of the responses that identified general characteristics of the text as having been biased was based on this sentence:

> *“Previo a ello, Lenin sufrió un atentado. El mismo fue llevado a cabo por Fanni Kaplán, quien con tres tiros intentó ejecutarle. Lenin había sobrevivido, y Kaplán fue ejecutada.”*
>
> English translation:
>
> *“Prior to that, Lenin suffered an attack. This was carried out by Fanni Kaplan, who with three shots tried to execute him. Lenin had survived, and Kaplan was executed.”*

The response noted: “It seems that there is talk in favor of Lenin surviving.” A response under the information imbalance question actually discusses phrase-level bias in this sentence:

> *“Asimismo Terrie Dodds hace de Barbara Jackson, la mujer que ayudó a encarcelar a quienes antes habían sido sus mejores amigos.”*
>
> English translation:
>
> *“Terrie Dodds also plays Barbara Jackson, the woman who helped imprison those who had previously been her best friends.”*

The annotator singled out one phrase as indicating bias in their response: “Yes, above all the phrase that talks about the treason of the best friends.” The responses to the word-level question identified words with strong positive or negative connotations such as "Stalin, murdered, poisoned" and "treason, best friends, suspicions, paranoia."

These words and the phrases they are often a part of do not appear enough in the text to constitute a topic. Because the topic model-based bias scoring system groups words based on their distribution within the article, these infrequently occurring and non-topic specific words are also unlikely to appear in one of the generated topics. Therefore, the contribution of these individual words would not have much of a noticeable effect on the computer-produced bias score. However, from these responses it is obvious that these single few words contribute heavily to the human bias score. The weight of these words in our computer model of bias is not correlated with the weight of the words in the human model of bias. This implies that for a computer-based system to accurately note the effect of these words, it would most likely require a human-created lexicon of potentially biased language, which might take considerable effort to produce, especially in multiple languages.

Even when human annotators make bias determinations based on information imbalances (essentially what the topic model comparisons are looking for), their base of information is clearly different from that of the topic model. For a human to know that a piece of encyclopedic text is missing an important viewpoint or piece of information, they require prior knowledge of the subject being discussed. This is a potential advantage of the topic modeling approach: since the topics generated usually have recognizable connections between their constituent words, they provide a relatively objective view of the information and subjects covered in a piece of text (and how much coverage they receive). This “bird’s-eye view” of a text’s information content, combined with the ability to directly compare that content with that of a baseline topic distribution (in our case, the spectrum viewpoints of English and Russian Wikipedia editions), gives a perspective about the text that human readers can most likely not provide. If this new perspective could be presented in an easily understandable format for human readers, it could give a more nuanced view of a text’s biases beyond what can be detected based on sentiment alone. Because a human reader would then be able to see the differences in what information is presented across languages, they could detect whether important information is missing in the article they are reading.

Through our review of the free-response questions, we noted that most of the sentences mentioned by the annotators discussed Soviet subjects as opposed to United States or Latin American subjects. Upon further review of the chunks, we found the majority to be centered around Soviet subjects. The focus of the articles on may have caused additional information bias in the results. We hypothesize that this subject bias may have skewed our results, as a greater magnitude of directional bias was reported either toward or against the Soviet Union. The magnitude of directional bias was reported for the United States. Had the subjects been more inclusive of United States and Latin American subjects, the Soviet bias we found may not have held.

One critique we received at the 2017 Chicago Colloquium on Digital Humanities and Computer Science was that our three-viewpoint model may be perpetuating a colonialist narrative by attempting to force Latin American viewpoints, represented by the Spanish language articles, onto a spectrum between the American (English) and

Soviet (Russian) viewpoints. In doing so, the critique suggested, we may have neglected the possibility that viewpoints expressed in the Spanish Wikipedia articles in effect constitute their own viewpoint that is significantly different from both the English and Russian viewpoints. However, our project does not attempt to explore the entire Latin American viewpoint on the Cold War; instead it focuses on exploring the effect, if any, of these two countries on individual subjects (represented by individual articles) within the larger context of the Cold War. We sought to understand if either superpower truly exerted an influence over the Latin American perceptions of events. This region had been a quasi-battleground for the two superpowers of the time, with both attempting to exert some type of power over certain countries in the region. Our project aimed to explore if this attempt at power by these superpowers could have manifested itself in the way the countries viewed historical events. For example, we thought if a country had had a disfavorable interaction with the United States, then they could be likely to give negative attention to the United States, and thus creating a Wikipedia article that would be dissimilar to the English one and show a bias against the United States. We do however, acknowledge the potential negative effects of situating the Latin American perspective in this way and in the future could create a multidimensional scale on which to examine bias. An exploration of the Latin American viewpoint in its own right is certainly a worthwhile endeavor, but it is beyond the scope of our project, which is focused on the similarities between the Latin American articles and the spectrum viewpoints.

Chapter 6: Conclusion and Future Work

Our research has explored a novel application of multilingual topic modeling for the analysis of bias in informational texts. We used multilingual topic modeling to place Spanish Wikipedia articles on a spectrum ranging from an English to a Russian viewpoint, and found that this automatic classification did not correlate significantly with human judgments of bias in the articles. Annotators indicated in free response questions that they looked more closely at words and phrases than topics, which could explain the lack of correlation. However, we acknowledge that it may have been harder for human annotators to analyze topic distributions in the smaller chunk level. Since our classifier looks for bias in topic distribution rather than sentiment, our classifier could serve as a useful complement to human perception of bias in Wikipedia articles.

In addition to our findings, our research also created many opportunities for other researchers to explore questions raised by our exploration. These opportunities include collecting more data on our current corpus to get a more complete idea of human perceptions of bias in Wikipedia, further exploring the usefulness of information provided by both human annotations and topic modeling to the problem of identifying a text's biases, developing a tool that allows Wikipedia users to benefit from our research, applying the system to alternative corpora, refining the classifier, and modeling bias in our current corpus using sentiment and n-gram analysis.

The first opportunity for future work is simply to collect more human annotations in order to have more data to correlate with the automated classifier. Collecting more annotations will provide more data points to compare with the automated classifier and reduce the standard error of our correlation coefficient. If the correlation coefficient is still not significantly different from 0 after collecting more annotations, we can improve our confidence in the conclusion that our topic modeling system cannot accurately replicate human determinations of bias. Additionally, it is possible that the subset of articles in our corpus that we automatically selected for the annotation task were not representative of the entire population, so collecting more data will also improve our confidence that we have annotated a truly representative sample of our corpus. When comparing human-produced and computer-produced bias scores, the second round of human annotations was particularly illuminating because of the free-response questions. If we were to obtain more annotations, especially of articles that Workers find to be strongly biased, the answers to these questions could provide a starting point for developing a more accurate computational method of replicating human bias detection. The information they provide would also generally improve our definition of bias itself, since our attempt to define the concept may not have taken into account how variable the methods of determining bias actually are. This human method of analyzing bias probably differs somewhat from person to person, based on our results, and may differ even more across cultural or linguistic lines. When developing a computational method of analyzing bias, we should take these differences into account, so as to have the most general result possible.

One of our conclusions from analyzing the differences between our topic-based classifier and the human annotations is that our classifier provided a more expansive perspective on an article (where it stands in relation to other articles and other languages).

As noted in the discussion, the classifier could serve as a useful complement to a human Wikipedia reader's perception of bias by indicating differences in topic distribution across language versions, an impossible task for a human reader who cannot read articles in the other languages. We explored various ways to visually present the differences in topic distribution throughout our research, including a visible spectrum where the article is placed somewhere in between two other language versions of the article on a line, a Venn diagram showing shared and unshared topics, and a chart displaying information on the relative frequency of topics in different language versions. Developing a browser plugin or web-based tool accessible to casual Wikipedia readers that presents simple versions of these visualizations could be a useful application of our research for the use of the general public. Additionally, another use case could be a tool designed for Wikipedia researchers that presents more complex visualizations of topic distribution and contains more raw data on relative distribution.

Another avenue for the future is analyzing different document types. Since Wikipedia makes an active attempt to present information without bias through its NPOV policy, it is likely that other sources will provide more biased "endpoints" for the three-viewpoint model than Wikipedia. In our work, there was a case of a Spanish article that was scored by the computer as more biased toward the "American viewpoint" than the English article. This suggests that the low correlation in our results could have been partly due to a lack of relative bias in Russian and English Wikipedias. We have done preliminary research on other non-encyclopedic text sources, including Pravda, formerly the Communist Party's official newspaper in the USSR, and American newspaper archives, though the considerable effort required to text mine these corpora put this project beyond the scope of our research. If our methods were applied to these other sources, having more biased "endpoints" would hopefully clarify differences in topic distribution and sentiment across multilingual, biased corpora and help future researchers to further refine our classifier. Another advantage of non-Wikipedia corpora is that using them requires fewer assumptions on our part. The authorship of Wikipedia articles is mostly unknown, so we were forced to assume that language on Wikipedia correlated with nationality/culture, but for newspapers, the nationality/culture of the authors is more well-defined. In addition, since the text of these newspapers was written contemporaneously with the events of the Cold War, we would not have to conflate current Russian or American views with historical Soviet or American views. However, to fully apply our methods in the context of the three viewpoint model described, we would first need to identify an analogue to Wikipedia's interwiki links. This would provide the required correspondence among equivalent documents across language.

Looking at other sources could also be useful for finding different case studies for the research. The Israel-Palestine conflict is another topic to which the three-viewpoint model can be applied. It has generated enough problems for Wikipedia editors that the site has a "WikiProject" specifically devoted to neutral presentations of the conflict's history, therefore articles on the subject might contain more bias than similar articles on the Cold War ("Wikipedia:WikiProject Israel Palestine Collaboration", 2018). This conflict is ongoing, so the viewpoints and ideologies represented have definite adherents who could be editing Wikipedia. This provides a contrast with the Cold War, as the

USSR no longer exists and conflating its viewpoint with current Russian views is not a completely safe proposition.

Sentiment-based, as opposed to topic-based, computer analysis could be beneficial in replicating human annotation. Human annotators tended to report that they marked something as biased based on emotionally charged words, so it is likely that a better correlation with human annotations could be achieved by a classifier based on sentiment analysis. If a better correlation were achieved by a sentiment analysis system, it would help to confirm our hypothesis that a topic modeling based classifier provides a fundamentally different understanding of a text's bias than a human annotator. While topic modeling provides a unique perspective on the bias of text, sentiment analysis could give a more accurate bias judgment for texts (when comparing to human views). A challenge when assessing bias on a three-viewpoint model (or any relative bias to a specific viewpoint) using sentiment analysis is that this type of analysis usually relies on the connotations of specific words, often whether they are generally positive or negative. Measuring relative bias of a text would require examining the relationships of words to one another to determine how specific topics are discussed. In our case study, for example, a tool that could detect whether positive or negative words are generally used when discussing the US or USSR would provide a better replication of what human annotators found to be biased. The tool would be able to give the reader an idea of what the bias of the text might be without requiring them to read it first. Our research could be refined by integrating this idea: a tool that performs sentiment analysis on an article's topics and looks at how similar the positive and negative topics are to the spectrum viewpoints could improve our approach.

Another way to analyze bias in our corpus is to apply the Gentzkow & Shapiro (2010) method discussed in the literature review. We did some preliminary work on translating phrases from Russian and English into Spanish in order to run the Gentzkow & Shapiro method over the Spanish corpus but did not finish applying the method. Another approach could be to draw biased phrases from left-leaning and right-leaning Spanish-language publications. This method would generate slant scores which could then be compared with our human annotations.

Appendix 1: Topic Distribution Top Words

Topic ID	Top words														
0	en#army	ru#войск	es#ejército	en#forces	ru#cc	en#division	es#soldados	ru#армии	es#tropas	en#ss	en#troops	es#ss	en#soviet	es#división	ru#дивизии
1	en#statistics	ru#отдел	en#struve	es#malaya	es#struve	ru#партизанскую	ru#статистики	ru#войну	ru#великобритании	ru#война	ru#партизан	en#federal	ru#противника	es#estadística	en#malayan
2	en#wife	es#padre	es#prisión	en#prison	en#father	es#hijo	es#muerte	en#assassination	es#meses	en#married	es#tarde	es#madre	en#mother	en#whom	en#arrested
3	en#grand	en#nicholas	es#orden	en#duke	en#alexandra	es#rusia	es#nicolás	en#paul	ru#александра	ru#орден	es#caballeros	es#duque	es#olga	en#russia	es#alejandra
4	es#prisioneros	es#campos	es#gulag	en#prisoners	en#camps	en#gulag	en#camp	es#campo	es#trabajo	ru#right	ru#align	en#soviet	en#labor	ru#заключённых	es#sistema
5	es#finlandia	en#finland	en#finnish	ru#финляндии	es#civil	en#border	es#paz	en#finns	es#finlandeses	ru#война	es#independencia	es#rodesia	es#gobierno	ru#родезии	es#finlandesa
6	en#military	en#argentina	es#argentina	en#argentine	es#militar	ru#видела	ru#аргентине	ru#тысяч	ru#аргентины	en#junta	ru#президента	ru#хунты	ru#хорхе	ru#похищения	en#killed
7	es#atentado	es#mossad	es#estadounidenses	en#stars	es#ataque	ru#ливане	en#glass	en#photograph	es#organización	es#fotografía	es#hecho	en#beirut	en#attack	ru#штаб	es#terroristas
8	en#party	es#partido	en#opposition	es#trotski	en#communist	es#oposición	ru#партии	en#trotsky	en#bukharin	es#comunista	es#trotsky	es#revolución	es#bujarin	es#izquierda	es#política
9	en#road	es#leningrado	ru#тыс	en#falange	en#ice	es#petersburgo	es#camino	es#soviética	en#leningrad	en#spanish	es#club	en#mm	es#acceso	es#carretera	en#lake
10	en#chinese	en#railway	es#rusia	es#ferrocarril	es#china	es#manchuria	es#ruso	en#russian	es#chino	en#soviets	es#soviética	es#rusa	en#eastern	en#china	ru#квжд
11	es#rusia	en#russian	es#óblast	es#ruso	en#oblast	en#russia	es#soviética	en#krai	es#rusa	es#krai	es#unión	en#krasnodar	es#república	es#orden	es#krasnodar
12	en#russian	es#terror	en#terror	es#rojo	en#white	es#blanco	en#red	es#rusa	es#revolución	es#rusia	en#russia	es#rusos	es#civil	en#cheka	es#blancos
13	ru#нквд	en#nkvd	en#stalin	es#nkvd	es#stalin	es#purga	en#soviet	en#executed	es#soviética	en#trial	en#purge	es#ejecutados	es#juicio	en#arrested	es#muerte
14	es#soviética	es#unión	en#soviet	en#union	es#república	es#repúblicas	es#rss	ru#сср	es#letonia	en#republics	es#urss	es#lituania	es#rusia	es#socialista	ru#рсфср
15	es#política	es#represión	es#cosacos	es#soviética	es#ruso	en#cossack	es#unión	en#cossacks	es#genocidio	en#killed	es#muchos	es#grupos	en#don	es#intento	en#soviet
16	en#soviet	en#curtain	es#soviética	ru#занавес	en#iron	es#fría	es#término	es#occidental	ru#железный	es#occidentales	es#telón	es#unión	en#border	en#cold	es#acero
17	en#israeli	es#israel	en#egyptian	en#israel	es#canal	es#israelíes	es#egipto	es#israelí	en#egypt	ru#израиля	es#egipcio	es#suez	en#canal	es#reino	es#sinaí
18	es#soviética	es#quieres	es#jpg	es#propaganda	en#propaganda	es#periódicos	es#mundial	es#conquistar	es#unión	ru#пропаганды	es#prensa	ru#пропаганда	es#urss	es#importantes	es#world
19	es#committee	en#serge	en#novel	es#novela	es#new	es#york	en#russian	en#encyclopedia	es#from	en#german	es#volume	en#don	en#published	en#volume	en#rosenberg
20	en#soviet	es#constitución	ru#советов	es#gobierno	en#russian	en#congress	en#soviets	en#constitution	ru#депутатов	en#assembly	en#party	es#congreso	es#poder	es#asamblea	en#constituent
21	es#vietnam	es#laos	en#laos	en#lao	en#vietnam	es#comunismo	ru#декларация	ru#лао	en#vietnamese	ru#лаоса	es#declaración	en#european	ru#хо	ru#ши	en#air
22	es#kgb	en#kgb	ru#кгб	es#seguridad	ru#безопасности	es#inteligencia	en#security	en#intelligence	en#vlasov	es#directorio	ru#государственной	en#directorate	es#departamento	es#servicio	en#soviet
23	en#nuclear	es#misiles	es#nuclear	es#nucleares	en#missile	en#defense	es#tratado	ru#ракет	es#defensa	es#armas	en#weapons	en#us	en#missiles	ru#про	en#system
24	ru#цк	es#partido	ru#кпсс	es#comité	en#party	en#committee	es#comunista	en#central	es#central	es#soviética	es#unión	es#congreso	es#pcus	es#secretario	en#politburo
25	en#jewish	es#judíos	en#jews	es#judío	ru#евреев	es#casa	es#asesinados	es#judía	es#stern	ru#еврейского	en#building	es#edificio	es#jpg	es#thumb	es#lina
26	ru#мы	ru#бы	ru#перед	ru#дело	ru#ни	ru#тем	ru#даже	ru#день	ru#если	ru#будет	ru#многие	ru#менее	ru#решение	ru#своих	ru#заявил
27	es#ordenadores	en#computers	ru#пк	en#computer	en#soviet	es#ordenador	en#produced	es#soviética	es#personales	en#system	es#ec	en#clone	es#países	en#developed	es#hum
28	es#elecciones	en#party	es#partido	ru#партии	es#votos	ru#партия	en#elections	en#election	ru#выборах	es#partidos	es#políticos	es#resultados	es#parlamentarias	en#votes	es#presidenciales
29	es#soviética	es#unión	es#consejo	es#estatal	es#ruso	es#comité	ru#председателя	es#urss	es#rusia	es#ministro	ru#рсфср	ru#заместитель	en#council	ru#совета	en#ussr
30	ru#операции	es#uss	es#invasión	es#estadounidenses	es#isla	en#navy	en#air	es#ee	es#jemeres	ru#острова	es#rojos	es#fuerzas	en#ship	en#marine	en#force
31	ru#операция	en#pinochet	ru#операции	ru#кондор	en#dina	en#chile	es#prusia	es#mir	ru#пиночета	en#chilean	en#operation	es#chile	es#cóndor	ru#бразилии	en#condor
32	en#language	es#soviética	es#ruso	es#unión	en#soviet	en#tatar	es#rusia	es#lengua	en#languages	en#russian	es#libro	ru#языка	es#lingüistas	es#tártaro	es#tártara
33	en#treaty	en#russia	es#metro	en#brest	es#estonia	es#rusia	en#litovsk	en#peace	en#moscow	es#tratado	en#metro	en#underground	es#ruso	en#central	en#russian
34	en#iran	es#irán	en#iranian	ru#заложников	ru#ирана	ru#операция	es#iraní	en#embassy	en#hostages	ru#иране	es#rehenes	en#shah	en#tehran	es#carter	ru#посольства
35	ru#мы	en#version	ru#песни	es#versión	en#we	en#song	en#orthodox	en#internationale	ru#текст	es#canción	ru#интернационала	es#kombat	ru#песня	ru#сср	ru#русском
36	es#albania	en#albania	es#enver	es#hoxha	en#albanian	es#new	es#york	en#hoxha	es#yugoslavia	en#yugoslavia	es#press	es#publishing	es#ruptura	ru#албании	es#house
37	es#unión	es#soviética	en#border	es#nkvd	es#fuerzas	ru#отдел	es#especiales	en#special	en#soviet	es#miembros	ru#назначения	ru#сэв	en#comecon	es#gru	es#relaciones
38	es#azerbaiyán	en#azerbaijan	en#aliyev	es#grecia	ru#греции	en#greece	en#greek	ru#азербайджана	es#bakú	en#baku	es#corrupción	ru#баку	es#república	es#gobierno	en#junta
39	en#uprising	es#manifestantes	es#alemania	es#riga	es#protestas	es#manifestaciones	en#protests	es#sublevación	es#soviéticas	es#trabajadores	es#berlin	en#protesters	es#trabajo	en#workers	en#berlin
40	en#art	es#arte	en#artists	es#jpg	es#artistas	en#leningrad	en#soviet	es#bw	es#realismo	es#socialista	es#image	ru#художников	en#russian	en#union	en#exhibition
41	es#bush	en#bush	ru#and	es#reagan	en#reagan	ru#буш	ru#in	es#kissinger	ru#рейгана	en#president	es#presidente	en#nixon	ru#блок	ru#восточный	ru#eastern
42	en#latvian	es#letones	es#rumania	en#riflemen	en#romanian	es#sóviet	en#romania	ru#румынии	en#speech	en#latvia	es#soviética	es#doctrina	es#rumano	es#ruso	es#consejo
43	es#gobierno	en#hungarian	ru#правительства	en#communist	en#kun	en#hungary	en#republic	en#minister	ru#aaa	es#habían	en#red	ru#правительство	es#frente	en#soviet	en#army
44	ru#кпрф	en#russian	es#rusia	es#óblast	en#russia	en#oblast	en#medal	en#federation	en#moscow	en#class	es#borís	ru#федерации	es#moscú	es#unión	en#afghanistan
45	ru#сталина	en#stalin	en#khrushchev	es#stalin	en#soviet	es#jrushchov	ru#ленина	en#union	es#política	ru#личности	en#renamed	ru#сталин	es#culto	en#speech	es#personalidad
46	en#cities	en#closed	es#ciudades	ru#область	es#comunismo	en#russia	en#labor	es#rusia	es#estas	en#communism	es#apartamento	es#campesinos	es#familias	en#communal	es#vivir
47	en#soviet	es#economía	en#economic	en#economy	es#producción	en#union	en#production	es#industria	es#soviética	ru#промышленности	es#urss	es#unión	ru#хозяйства	en#plan	en#planning
48	en#gorky	es#gorki	es#rusia	ru#горький	ru#горького	es#lenin	en#novel	es#literatura	es#vladimir	en#maxim	en#writer	es#obras	en#russian	es#escritor	es#escritores
49	en#military	en#forces	es#fuerzas	es#gobierno	es#ejército	es#apoyo	en#support	en#army	en#force	ru#война	ru#силы	en#troops	es#conflicto	ru#правительство	es#tropas
50	es#mir	es#jafar	en#azerbaijan	es#azerbaiyán	en#ssr	en#gold	ru#валют	en#azerbaijani	es#uu	es#ee	es#mundial	ru#багиров	es#azerí	en#mir	es#oro
51	en#japanese	es#corea	es#japón	en#japan	es#mundial	ru#японии	en#korea	en#china	en#korean	es#japonés	en#asia	es#gobierno	ru#мировой	ru#китайской	ru#азии
52	en#site	en#town	es#cerca	en#administrative	en#pechora	en#memorial	es#monumento	es#sitio	en#settlement	es#fosas	es#comunes	es#víctimas	en#monument	es#kilómetros	es#memorial
53	es#china	en#vietnam	es#vietnam	en#vietnamese	es#klan	en#klan	en#china	es#camboya	en#chinese	es#jemeres	en#khmer	es#rojos	ru#вьетнам	en#rouge	es#vietnamita
54	en#victory	es#georgia	ru#победы	es#victoria	en#georgia	ru#грузии	en#anniversary	en#alexei	es#desfile	es#px	en#georgian	es#moscú	en#holiday	en#patriotic	en#medal
55	es#república	es#territorios	es#desaparecidos	es#china	es#soviética	es#unión	ru#width	en#russian	es#oriental	es#popular	es#fundados	en#republic	ru#from	ru#px	ru#gray
56	ru#связи	ru#период	ru#частности	ru#этих	ru#стали	ru#стороны	ru#котором	ru#лишь	ru#кроме	en#moscow	ru#начале	ru#ряд	ru#годов	ru#членов	ru#времени
57	en#rights	ru#комиссии	es#política	en#committee	ru#маккарти	ru#комиссия	en#communist	es#on	ru#red	en#court	en#act	ru#деятельности	es#comité	en#human	en#activities
58	es#ruso	en#khanty	es#rebelión	en#soviet	es#muchos	en#forest	en#rebellion	en#revolt	es#socialismo	ru#советского	es#países	es#revuelta	en#land	es#superficie	es#europa
59	es#checoslovaquia	es#hungría	en#czechoslovakia	es#praga	en#prague	es#república	en#hungary	es#varsovia	es#pacto	en#hungarian	ru#чехословакии	en#warsaw	es#régimen	en#czech	ru#венгрии
60	es#angola	en#angola	ru#анголы	es#cuba	ru#юар	en#africa	ru#против	en#african	en#unita	en#cuban	es#independencia	ru#кубинские	en#mpla	es#cubanos	en#cuba
61	es#academia	es#ciencias	en#institute	en#academy	ru#наук	en#sciences	ru#институт	en#science	ru#академии	es#soviética	en#research	es#unión	en#powers	es#instituto	en#pyotr
62	en#kissinger	es#operación	en#president	en#intelligence	en#operation	en#said	es#derechos	es#inteligencia	es#dictadura	es#asesinato	es#plan	en#kuznetsov	en#according	en#america	en#foreign
63	en#space	es#programa	es#soyuz	es#espacio	en#project	en#soyuz	ru#союз	es#proyecto	en#crew	es#misión	es#bomba	es#pruebas	en#bomb	en#mission	es#espacial
64	ru#никарагуа	es#nicaragua	en#nicaragua	ru#национальной	es#salvador	en#contras	es#sandinista	en#nicaraguan	ru#de	es#presidente	en#somoza	en#contra	es#somoza	es#panamá	ru#гвардии
65	en#jpg	en#file	es#autor	es#derecho	es#derechos	en#present	es#ley	es#soviética	en#copyright	ru#права	en#law	es#unión	es#autores	en#works	es#urss
66	en#soviet	en#german	es#alemania	es#soviética	en#germany	es#unión	en#union	es#invasión	es#pacto	es#tratado	en#pact	ru#германии	es#mundial	es#nazi	es#soviético
67	en#economic	es#económica	es#plan	en#trade	ru#страны	ru#стран	en#plan	ru#px	es#países	en#central	en#european	es#política	en#foreign	en#economy	es#cambio
68	en#party	en#communist	es#partido	es#comunista	es#comunistas	es#partidos	ru#партии	es#políticos	en#communists	ru#партия	en#soviet	en#political	es#unión	es#político	en#election
69	en#stalin	es#stalin	ru#сталин	en#zinoviev	en#socialism	es#política	es#rusia	es#influencia	ru#народа	en#russian	en#proletariat	en#revolution	es#década	es#propio	es#alianza

70	en#gorbachev	es#gorbachov	en#soviet	en#union	es#presidente	es#urss	es#mijaíl	en#president	en#coup	es#golpe	es#unión	es#yeltsin	es#rusia	ru#гкчп	es#borís
71	en#ukraine	en#famine	en#ukrainian	en#million	en#holodomor	en#soviet	es#hambruna	es#campesinos	es#colectivización	es#millones	es#stalin	en#peasants	es#ucrania	ru#млн	ru#усср
72	en#children	es#komsomol	en#komsomol	es#organización	en#women	es#unión	ru#влксм	es#juventud	en#youth	en#young	en#kirov	es#niños	es#kírov	ru#улица	es#comunista
73	en#censorship	es#censura	en#word	es#libros	en#stalin	ru#цензуры	ru#эфиопии	ru#товарищ	en#derg	en#information	en#socialist	ru#цензура	es#etiopía	es#camarada	en#comrade
74	es#misiles	es#cuba	en#missiles	en#missile	en#system	es#sistema	en#kennedy	en#cuban	en#nato	ru#кубе	es#castro	en#radar	en#equipment	es#soviética	es#crisis
75	en#berlin	es#berlín	en#germany	es#alemania	en#german	ru#гдр	ru#германии	ru#фрг	ru#берлина	en#wall	es#rda	es#occidental	es#muro	es#oeste	ru#берлин
76	en#air	en#aircraft	es#zona	es#aérea	es#aviones	en#operation	es#cerca	es#aéreo	es#ataque	ru#операции	es#jpg	en#airlift	en#isbn	es#km	es#territorio
77	es#ucrania	en#ukrainian	en#ukraine	es#república	es#soviética	ru#украины	en#kiev	es#rada	es#ucraniano	es#popular	en#rada	es#http	es#www	en#soviet	en#central
78	es#soviética	es#crimea	es#unión	en#soviet	en#crimean	en#psychiatry	en#psychiatric	en#political	es#tártaros	es#disidentes	es#mundial	en#russian	en#tatars	es#psiquiatría	en#according
79	es#espías	en#spy	en#soviet	en#intelligence	ru#разведки	es#espionaje	en#hanssen	es#inteligencia	en#cia	ru#работал	ru#советской	es#espía	en#fbi	en#espionage	en#blunt
80	ru#против	es#gobierno	ru#правительства	en#army	en#soldiers	es#ejército	en#minister	ru#армии	ru#тот	ru#военного	ru#всё	ru#которое	ru#восстание	ru#эти	ru#переговоров
81	en#military	es#militar	en#army	es#ejército	en#red	en#law	en#academy	es#unión	es#soviética	en#soviet	es#http	es#militares	es#rojo	es#academia	en#commander
82	es#soviética	en#party	en#russian	en#moscow	es#ruso	en#soviet	es#moscú	es#rusia	es#políticos	es#unión	es#partido	es#miembro	ru#член	ru#члены	en#committee
83	en#philby	en#isbn	en#society	ru#isbn	es#york	en#burgess	en#spain	es#burgess	ru#роли	es#española	es#mi	en#london	ru#жизни	en#republican	ru#общество
84	en#even	en#us	en#take	en#came	es#caso	en#never	en#according	es#ante	en#considered	en#possible	en#response	en#sent	en#report	en#do	es#fría
85	es#px	ru#громыко	ru#px	es#premio	en#soviet	en#foreign	es#exteriores	es#rublo	es#embajador	en#ruble	en#rubles	es#urss	es#cargo	es#diplomático	ru#дел
86	en#soviet	es#soviética	es#unión	en#union	es#soviético	es#soviéticos	es#urss	en#ussr	ru#советских	es#soviéticas	en#soviets	ru#советского	en#khrushchev	en#moscow	es#moscú
87	ru#совета	en#soviet	en#stalin	es#stalin	ru#верховного	es#soviética	es#unión	en#molotov	ru#депутаты	es#jrushchov	ru#созыва	es#mólotov	en#brezhnev	es#ministro	es#beria
88	en#church	es#iglesia	ru#церкви	en#unification	ru#объединения	en#religious	en#white	es#ortodoxa	en#catholic	en#fbi	ru#фбр	es#dios	es#moon	en#moon	ru#церковь
89	en#republic	en#president	en#yemen	es#presidente	es#yemen	es#república	en#democratic	ru#министр	en#minister	ru#республики	ru#президента	ru#премьер	en#elected	ru#совета	en#candidate
90	es#radio	en#radio	ru#операция	ru#операции	ru#нато	es#campamento	en#rfe	ru#радио	ru#организаций	ru#италии	ru#стран	es#emisora	es#adiguesia	ru#операцию	en#broadcasting
91	es#países	es#política	es#relaciones	en#foreign	en#policy	en#nations	en#relations	ru#договора	en#security	en#countries	es#gobierno	en#military	en#treaty	en#economic	en#cold
92	es#oriente	es#región	es#ruso	es#extremo	en#far	es#lejano	en#russian	es#rusia	es#pob	es#siberia	en#region	es#federal	en#russia	en#hesse	es#km
93	es#sistema	es#días	en#nuclear	es#central	en#passport	en#system	en#tax	ru#работы	en#power	es#trabajadores	es#reactor	es#ciudadanos	es#registro	es#nuclear	en#plant
94	es#bandera	es#svg	en#red	en#flag	es#rojo	en#star	es#temor	en#file	es#flag	es#cover	es#duck	en#emblem	es#estrella	en#svg	ru#svg
95	es#accidente	en#accident	es#lanzamiento	en#challenger	es#nasa	en#launch	en#crew	es#challenger	en#shuttle	en#nasa	es#transbordador	ru#nasa	ru#экипажа	en#space	es#operación
96	en#polish	es#polonia	en#poland	es#polaco	es#polacos	en#soviet	es#polaca	es#soviéticos	ru#польши	ru#польской	en#army	ru#польских	en#soviets	ru#армии	es#gobierno
97	en#lenin	es#lenin	en#revolution	en#bolsheviks	en#russian	en#bolshevik	es#bolcheviques	es#revolución	en#russia	en#revolutionary	ru#ленин	ru#ленина	en#provisional	en#petrograd	en#socialist
98	es#conferencia	en#conference	ru#конференции	es#roosevelt	es#churchill	en#churchill	es#naciones	es#mundial	en#roosevelt	en#europe	en#truman	es#aliados	es#reino	en#nations	es#unido
99	en#anti	en#movement	es#miembros	es#movimiento	en#organization	ru#организации	es#gobierno	en#communist	es#organización	es#isbn	ru#движения	en#freedom	ru#организация	ru#движение	es#bloque

Appendix 2: Category List

These are the Spanish Wikipedia categories used to compile our corpus.

Agentes del KGB
Alemania Occidental
Anticomunismo
Arte de la Unión Soviética
Bloque del Este
Conferencias de la Segunda Guerra Mundial
Conflictos de la Guerra Fría
Constituciones de la Unión Soviética
Cultura de la Unión Soviética
Derecho de la Unión Soviética
Derechos humanos en la Unión Soviética
Directores del KGB
Disolución de la Unión Soviética
Diáspora soviética
Economía de la Unión Soviética
Ejecutados de la Unión Soviética
Emigrantes de la Unión Soviética
Escuela de las Américas
Espías de la Guerra Fría
Espías de la Unión Soviética
Gran Purga
Guerra Fría
Guerras de la Unión Soviética
Gulag
Historia de Estados Unidos (1945-1989)
Historia de la Unión Soviética
Intervenciones militares de Cuba
KGB
Muro de Berlín
NKVD
Ocupaciones militares de la Unión Soviética
Operaciones de la KGB
Operación Cóndor
Partido Comunista de la Unión Soviética
Política de la Unión Soviética
Políticos de la Unión Soviética
Primavera de Praga
Propaganda anticomunista
Propaganda de la Unión Soviética
Realismo socialista
Relaciones Alemania-Unión Soviética
Relaciones Bulgaria-Unión Soviética
Relaciones Checoslovaquia-Unión Soviética
Relaciones China-Unión Soviética
Relaciones Cuba-Unión Soviética
Relaciones España-Unión Soviética
Relaciones Estados Unidos-Unión Soviética
Relaciones Francia-Unión Soviética
Relaciones Hungría-Unión Soviética
Relaciones India-Unión Soviética
Relaciones Irán-Unión Soviética
Relaciones Mongolia-Unión Soviética
Relaciones México-Unión Soviética
Relaciones Polonia-Unión Soviética
Relaciones Reino Unido-Unión Soviética
Relaciones Rumania-Unión Soviética
Relaciones Suiza-Unión Soviética
Relaciones Turquía-Unión Soviética
Relaciones Unión Soviética-Uruguay
Relaciones Unión Soviética-Vietnam
Relaciones bilaterales de la Unión Soviética
Relaciones internacionales de la Unión Soviética
Represión política en la Unión Soviética
Resoluciones del Consejo de Seguridad de las Naciones Unidas referentes a la Unión Soviética
Revoluciones de 1989
Revolución Sandinista
Sociedad de la Unión Soviética
Soviéticos
Símbolos de la Unión Soviética
Terminología soviética
Terrorismo de Estado en Argentina en las décadas de 1970 y 1980
Tratados de la Unión Soviética
Unión Soviética
Unión de Partidos Comunistas
Zona de ocupación estadounidense

Appendix 3: Articles

The following shows a list of articles in all three language editions that were included in our corpus.

101st kilometre
1924 Soviet Constitution
1936 Soviet Constitution
1948 Czechoslovak coup d'état
1951 Polish–Soviet territorial exchange
1960 U-2 incident
1964 European Nations' Cup Final
1966 Palomares B-52 crash
1968 Thule Air Base B-52 crash
1976 Argentine coup d'état
1977 Soviet Constitution
1983 Beirut barracks bombings
1983 Soviet nuclear false alarm incident
1990 Goodwill Games
1991 Sino-Soviet Border Agreement
1991 Soviet coup d'état attempt
24th Congress of the Communist Party of the Soviet Union
28th Congress of the Communist Party of the Soviet Union
500 Days
99 Luftballons
A-35 anti-ballistic missile system
ANZUS
Able Archer 83
Absamat Masaliyev
Aeroflot Flight 244
Aeroflot Flight 6833
Aftermath of World War II
Agitprop
Ahmad Javad
Akhsarbek Galazov
Akmal Ikramov
Aldrich Ames
Aleksandr Mikhailovich Orlov
Aleksandr Sakharovsky
Alexander Litvinenko
Alexander Pechersky
Alexander Shelepin
Grigori Sokolnikov
Grigory Kulik
Grigory Petrovsky
Grigory Zinoviev
Group of Soviet Forces in Germany
Gulag
Gulf of Sidra incident (1981)
Guy Burgess
Günter Schabowski
Hammer and sickle
Harry Dexter White
Helsinki Accords
Henry Kissinger
Hesse
Heydar Aliyev
Historiography in the Soviet Union
History of Namibia
History of the Soviet Union
History of the United States (1945–64)
Ho Chi Minh trail
Holodomor
Homo Sovieticus
Honghuzi
House Un-American Activities Committee
Hryhoriy Hrynko
Hukbalahap Rebellion
Hungarian Democratic Forum
Hungarian Revolution of 1956
Hungarian Soviet Republic
I Have a Dream
Idania Fernandez
Ignace Reiss
Igor Gouzenko
Igor Panarin
Intermediate-Range Nuclear Forces Treaty
Invasion of Grenada
Ion Mihai Pacepa
Iona Yakir
Project Azorian
Project Mogul
Propaganda Due
Propaganda in the Soviet Union
Propiska in the Soviet Union
Provisional Government of the Republic of China (1937–40)
Pyotr Shirshov
Pyramiden
Pēteris Stučka
Qey Shibir
Quebec Conference, 1943
R504 Kolyma Highway
RYAN
Radio Free Asia
Radio Free Europe/Radio Liberty
Raising a flag over the Reichstag
Ramón Mercader
Reagan Doctrine
Reaganomics
Red Army Military Law Academy
Red Scare
Red Terror
Red Terror (Spain)
Red star
Refusenik
Rehabilitation (Soviet)
Religion in the Soviet Union
Reorganized National Government of the Republic of China
Republic of Mahabad
Revolutionary committee (Soviet Union)
Revolutionary tribunal (Russia)
Revolutions of 1989
Reykjavík Summit
Rhodesian Bush War
Right Opposition

Alexander Shliapnikov
Alexander Tkachov (politician)
Alexander Yegorov (military)
Alexanderplatz demonstration
Alexandra Feodorovna (Alix of Hesse)
Alexandra Kollontai
Alexei Kosygin
Alexei Nikolaevich, Tsarevich of Russia
Alexey Kuznetsov
Alexey Stakhanov
All-Russian Central Executive Committee
Allied intervention in the Russian Civil War
Amethyst Incident
Anandyn Amar
Anatoliy Gekker
Anatoly Dobrynin
Anatoly Lunacharsky
And Quiet Flows the Don
Andrei Gromyko
Andrei Zhdanov
Andrey Vlasov
Andrey Vyshinsky
Anglo-Polish military alliance
Anglo-Soviet Treaty of 1942
Anglo-Soviet invasion of Iran
Angolan Civil War
Anthony Blunt
Anti-Ballistic Missile Treaty
Anti-Bolshevik Bloc of Nations
Anti-Comintern Pact
Anti-Party Group
Anti-Sovietism
Anti-communism
Apollo–Soyuz Test Project
April 9 tragedy
Arcadia Conference
Argentine Anticommunist Alliance
Arkady Rosengolts
Arkady Shevchenko
Armenian Communist Party
Armia Ludowa
Arms race
Army of the Republic of Vietnam
Artek (camp)
Artel
Article 58 (RSFSR Penal Code)
Aslan Dzharimov
Iosif Grigulevich
Ipatiev House
Iran crisis of 1946
Iran hostage crisis
Iran–Contra affair
Irina Baldina
Iron Curtain
Isaak Zelensky
Ivan Belov (commander)
Ivan Silayev
Ivan Skvortsov-Stepanov
Ivan Smirnov (politician)
Ivan Teodorovich
Ivar Smilga
Japanese Red Army
Jewish Anti-Fascist Committee
Jewish Bolshevism
Jiří Dienstbier Jr.
John Birch Society
Joint State Political Directorate
Joseph Stalin
Joseph Stalin Museum, Gori
Jukka Rahja
Jukums Vācietis
Julius and Ethel Rosenberg
Junta of National Reconstruction
Józef Czapski
KGB
Kaliningrad Oblast
Kalmyk Autonomous Oblast
Katyn massacre
Kazym rebellion
Kengir uprising
Khertek Anchimaa-Toka
Khozraschyot
Khrushchev Thaw
Khrushchyovka
Killing of Peter Fechter
Kim Philby
Kitchen Debate
Klaipėda Region
Kola Norwegians
Kolkhoz
Kolyma
Komarovo, Saint Petersburg
Kombrig
Komdiv
Komsomol
Konon Molody
Konstantin Chernenko
Konstantin Rodzaevsky
Korean Air Lines Flight 902
Road of Life
Robert A. Lovett
Robert Eideman
Robert Hanssen
Rock Against Communism
Rudolf Abel
Ruhulla Akhundov
Rundfunk im amerikanischen Sektor
Russian Alsos
Russian Constituent Assembly
Russian Constituent Assembly election, 1917
Russian Constitution of 1918
Russian Far East
Russian Federal State Statistics Service
Russian Provisional Government
Russian Social Democratic Labour Party
Russian naval facility in Tartus
Russian presidential election, 1991
Russification
Russo-Persian Treaty of Friendship (1921)
Ryszard Kukliński
SMERSH
SS Chelyuskin
START I
SWAPO
Saar Protectorate
Salami tactics
Salvadoran Civil War
Samad aga Agamalioglu
Samantha Smith
Samizdat
Sand War
Sandarmokh
Sandinista Popular Army
Scissors Crisis
Second Taiwan Strait Crisis
Secretariat of the Communist Party of the Soviet Union
Securitate
Sergei Kruglov (politician)
Sergey Ilyushin
Sergey Kirov
Sergey Syrtsov (politician)
Sevan–Hrazdan Cascade
Shakhty Trial

Aslan Tkhakushinov
August Kork
Austrian State Treaty
Automotive industry in the Soviet Union
Azerbaijan Communist Party (1993)
Baghdad Pact
Baltic Way
Bandung Conference
Baruch Plan
Basic Treaty, 1972
Basmachi movement
Batallón de Inteligencia 601
Belavezha Accords
Bell P-63 Kingcobra
Ben Linder
Berlin Blockade
Berlin Crisis of 1961
Berlin Wall
Bill Stewart (journalist)
Black January
Black Monday (1987)
Blat (favors)
Blood in the Water match
Bloop
Boris Bazhanov
Boris Feldman
Boris Gromov
Boris Kamkov
Boris Numerov
Boris Ponomarev
Boris Pugo
Bretton Woods system
Brezhnev Doctrine
Bruno Rizzi
Butovo firing range
Béla Kun
COINTELPRO
Cairo Conference
Call of Duty: Black Ops: Declassified
Cambodian Civil War
Cambodian–Vietnamese War
Cambridge Five
Carpathian Ruthenia
Carter Doctrine
Casa Presei Libere
Casablanca Conference
Case of Trotskyist Anti-Soviet Military Organization
Case of the Anti-Soviet "Bloc of
Korenizatsiya
Kotelnicheskaya Embankment Building
Kotlas
Kremlin Wall Necropolis
Kremlin stars
Kremlinology
Ku Klux Klan
Kukryniksy
Kurapaty
Laotian Civil War
Latvian Riflemen
Latvian Soviet Socialist Republic
Lavrentiy Beria
Law of Spikelets
Lazar Kaganovich
Left Opposition
Lend-Lease
Leonid Brezhnev
Leonid Krasin
Leonid Nikolaev
Lev Kamenev
Lev Vasilevsky
Levashovo Memorial Cemetery
Lina Prokofiev
List of heads of state of the Soviet Union
Lithuanian Soviet Socialist Republic
Little Octobrists
Lona Cohen
Louis Adamic
Lubyanka Building
Lukyanivska Prison
Malayan Emergency
Malta Summit
Manfred Stern
Marshall Plan
Mask of Sorrow
Massive retaliation
Matvei Muranov
Maxim Gorky
Mayaguez incident
McCarthyism
Memorial (society)
Metro-2
Mikhail Borodin
Mikhail Chernov (politician)
Mikhail Gorbachev
Mikhail Kalinin
Mikhail Koltsov
Mikhail Suslov
Sharashka
Shoe-banging incident
Shortage economy
Sikorski–Mayski agreement
Sinatra Doctrine
Singing Revolution
Sino-Soviet border conflict
Sino-Soviet conflict (1929)
Sino-Soviet split
Sino-Vietnamese War
Smolensk Archive
Snow Leopard award
Socialism in One Country
Socialism with a human face
Socialist emulation
Socialist realism
Solomon Lozovsky
Solovki prison camp
Sosnogorsk
South African Border War
South Yemen
Soviet (council)
Soviet Border Troops
Soviet Census (1989)
Soviet Information Bureau
Soviet Union
Soviet Union passport
Soviet Union referendum, 1991
Soviet War Memorial (Treptower Park)
Soviet art
Soviet atomic bomb project
Soviet calendar
Soviet dissidents
Soviet invasion of Poland
Soviet occupation of Bessarabia and Northern Bukovina
Soviet people
Soviet ruble
Sovietization
Soviet–Afghan War
Soviet–Albanian split
Soviet–Japanese Neutrality Pact
Soviet–Japanese border conflicts
Sovnarkhoz
Soyuz 28
Soyuz 30
Soyuz 31
Soyuz 33
Soyuz 36
Soyuz 37

Rights and Trotskyites"
Cecilia Bobrovskaya
Censorship in the Soviet Union
Central Committee of the Communist Party of the Soviet Union
Central Council of Ukraine
Charter 77
Checkpoint Charlie
Chernobyl Forum
Children of the Arbat
Chinese Eastern Railway
Christian Rakovsky
Closed city
Cold War
Collectivization in the Soviet Union
Combat (photograph)
Comecon
Cominform
Committee for State Security
Committee of Youth Organisations
Communal apartment
Communarka shooting ground
Communist Party of Belarus
Communist Party of Kazakhstan
Communist Party of Latvia
Communist Party of Lithuania
Communist Party of South Ossetia
Communist Party of Ukraine
Communist Party of the Russian Federation
Communist Party of the Soviet Union
Communist University of the National Minorities of the West
Comrade
Concise Literary Encyclopedia
Congress of People's Deputies of the Soviet Union
Congress of Soviets
Congress of Soviets of the Soviet Union
Constitution of the Moldavian SSR
Constitution of the Moldavian SSR (1941)
Constitution of the Soviet Union
Constitutionalist Liberal Party
Containment
Contras
Mikhail Tomsky
Mikhail Tukhachevsky
Mikhail Vladimirsky
Military Collegium of the Supreme Court of the Soviet Union
Ministry of Education (Soviet Union)
Ministry of Foreign Affairs (Soviet Union)
Mir Jafar Baghirov
Mir mine
Miranda v. Arizona
Molotov–Ribbentrop Pact
Mongolian Revolution of 1990
Montreux Convention Regarding the Regime of the Straits
Moral Code of the Builder of Communism
Morris Cohen (spy)
Moscow Armistice
Moscow Circus on Tsvetnoy Boulevard
Moscow Music Peace Festival
Moscow Peace Treaty
Moscow Victory Parade of 1945
Moscow–Washington hotline
Museum of Soviet Occupation (Tbilisi)
Mutual assured destruction
My God, Help Me to Survive This Deadly Love
Mykola Skrypnyk
NKVD
NKVD Order No. 001223
NKVD Order No. 00447
NKVD prisoner massacres
NKVD troika
Nadezhda Krupskaya
Nahum Eitingon
Nariman bey Narimanbeyov
National Academy of Sciences of Belarus
National Academy of Sciences of Ukraine
National Guard (Nicaragua)
National Intelligence Service (South Korea)
National Liberation Front of Angola
National Liberation Movement (Guatemala)
National Opposition Union (1989)
Soyuz 38
Soyuz 39
Soyuz 40
Soyuz T-11
Soyuz T-6
Space Shuttle Challenger disaster
Spetskhran
Spetsnaz
Sputnik crisis
StB
Stakhanovite movement
Stalin's alleged speech of 19 August 1939
Stalin Note
Stanislav Kosior
Stanisław Pestkowski
State Anthem of the Soviet Union
State Committee on the State of Emergency
State Emblem of the Soviet Union
State Protection Authority
Strategic Arms Limitation Talks
Subbotnik
Suez Crisis
Suppressed research in the Soviet Union
Supreme Soviet
Supreme Soviet of the National Economy
Tamara Press
Tashkent Soviet
Taurida Soviet Socialist Republic
Tear down this wall!
Tehran Conference
Tehri Dam
Ten-Day War
Territories of Poland annexed by the Soviet Union
The Black Book of Communism
The Death Match
The Great Terror
The Gulag Archipelago
The Internationale
The Snow Maiden (1952 film)
The Unbearable Lightness of Being

Copyright law of the Soviet Union
Corfu Channel incident
Council of Europe resolution 1481
Cuban Missile Crisis
Cuban intervention in Angola
Cuba–Soviet Union relations
Culture of the Soviet Union
Cursed soldiers
Curzon Line
Daigo Fukuryū Maru
De-Stalinization
Death flights
Declaration of Independence of Ukraine
Declaration of State Sovereignty of Ukraine
Declaration of the Rights of the Peoples of Russia
Decossackization
Dekulakization
Dem'ianiv Laz
Deportation of the Crimean Tatars
Dictatorship of the proletariat
Die Wende
Dirección de Inteligencia Nacional
Dirty War
Dissolution of the Soviet Union
Division of Korea
Dmitri Polyakov
Dmitri Shepilov
Dmitry Manuilsky
Doctors' plot
Duck and Cover (film)
Duga radar
Dumbarton Oaks Conference
Détente
Eastern Bloc
Eastern Pact
Economy of the Soviet Union
Eduard Shevardnadze
Education in the Soviet Union
Elbe Day
Elena Stasova
Embassy of Cuba in Moscow
Embassy of Russia in Havana
Endel Puusepp
Enemy of the people
Era of Stagnation
Ethiopian Civil War
Eufrosinia Kersnovskaya
Evgeny Pashukanis
National Reorganization Process
Nazino affair
Nestor Lakoba
New Economic Policy
New Forum
Nicaraguan Revolution
Night Wolves
Night of the Murdered Poets
Night of the Pencils
Nikita Khrushchev
Nikolai Bukharin
Nikolai Glebov-Avilov
Nikolai Gorbunov
Nikolai Ivanovich Kuznetsov
Nikolai Kondratenko
Nikolai Krylenko
Nikolai Leonov
Nikolai Patrushev
Nikolai Podgorny
Nikolai Podvoisky
Nikolai Ryzhkov
Nikolai Sukhanov
Nikolai Tikhonov
Nikolai Valentinov
Nikolai Voznesensky
Nikolai Yezhov
Nikolay Burenin
Nikolay Krestinsky
Nikolay Shvernik
Ninth Fort
Nixon Doctrine
Nixon shock
Nomenklatura
Non-Aligned Movement
Norair Sisakian
Norillag
Nuclear Explosions for the National Economy
Occupation of Poland (1939–1945)
Ogaden War
Old Bolshevik
Oleg Gordievsky
Oleg Kalugin
Oleg Lobov
Oleg Penkovsky
Olga Ivinskaya
On the Cult of Personality and Its Consequences
Operation Anadyr
Operation Charly
Operation Chrome Dome
Operation Colombo
Three Mile Island accident
Timurite movement
Tintin in the Land of the Soviets
Treaty of Berlin (1926)
Treaty of Brest-Litovsk
Treaty of Kars
Treaty of Moscow (1920)
Treaty of Rapallo (1922)
Treaty of Tartu (Russian–Finnish)
Treaty of Warsaw (1970)
Treaty on the Creation of the USSR
Treaty on the Final Settlement with Respect to Germany
Trial of the Sixteen
Tripartite Pact
Truman Doctrine
Tudeh Party of Iran
Tuvan People's Republic
UNOVIS
USSR in Construction
Ukrainian National Army
Ukrainian independence referendum, 1991
Ukrainian sovereignty referendum, 1991
Under Fire (film)
Unified Communist Party of Georgia
Union of Communist Parties – Communist Party of the Soviet Union
Union of Fascist Little Ones
Union of Sovereign States
Union of Soviet Composers
Union of Young Fascists – Vanguard (boys)
Union of Young Fascists – Vanguard (girls)
United Nations Conference on International Organization
United Nations Security Council Resolution 2
United Nations Security Council Resolution 3
United Nations Security Council Resolution 5
United States invasion of Panama
United States national missile

Evil empire
FRELIMO
Family members of traitors to the Motherland
Fanny Kaplan
Far North (Russia)
Fayzulla Khodzhayev
Felix Dzerzhinsky
Finno-Soviet Treaty of 1948
First Battle of Târgu Frumos
First Department
First East Turkestan Republic
First Taiwan Strait Crisis
First five-year plan (Soviet Union)
First they came ...
Flags of the Soviet Republics
Foreign relations of the Soviet Union
Forest Brothers
Four Policemen
Francis Gary Powers
Free Territory of Trieste
Free World
Friedrich Werner von der Schulenburg
Fritz Platten
Fyodor Raskolnikov
Fântâna Albă massacre
GIUK gap
GOELRO plan
GOST
GRAU
GSF Explorer
Gabdulkhay Akhatov
Geneva Accords (1988)
Gennady Yanayev
Genrikh Yagoda
Geography of the Soviet Union
George Blake
George H. W. Bush
Georges Agabekov
Georgian independence referendum, 1991
Georgy Aleksandrov
Georgy Chicherin
Georgy Malenkov
Georgy Oppokov
Georgy Pyatakov
German auxiliary cruiser Komet
German reunification
German–Soviet Credit Agreement (1939)
Operation Condor
Operation Cyclone
Operation Gladio
Operation Gold
Operation Long Jump
Operation Opera
Operation Peter Pan
Operation Toucan (KGB)
Operation Trust
Operation Unthinkable
Operation Wigwam
Orgburo
Orlando Letelier
Osip Piatnitsky
Otto Schmidt
Outer Space Treaty
Overman Committee
Pan-European Picnic
Panmunjom
Parasitism (social offense)
Paris Peace Treaties, 1947
Party of Communists of Kyrgyzstan
Party of Communists of the Republic of Moldova
Pavel Bulanov
Pavel Dybenko
Pavel Florensky
Pavel Mif
Pavel Postyshev
Pavel Sudoplatov
Pavlik Morozov
Peace of Riga
Peaceful coexistence
Pechora
People's Commissariat for Nationalities
People's Socialist Republic of Albania
Percentages agreement
Perestroika
Persian Socialist Soviet Republic
Peter Berngardovich Struve
Petrograd Soviet
Piatykhatky, Kharkiv Oblast
Polina Zhemchuzhina
Polish Committee of National Liberation
Polish October
Polish Round Table Agreement
Politburo
Politburo of the Communist Party
defense
Universals (Central Council of Ukraine)
Uprising of 1953 in East Germany
Uskoreniye
Uzbek Soviet Encyclopedia
V-J Day in Times Square
Vadim Bakatin
Valentin Pavlov
Valentina Tereshkova
Valerian Kuybyshev
Varlam Avanesov
Varvara Stepanova
Varvara Yakovleva (politician)
Vasili Arkhipov
Vasili Kuznetsov (politician)
Vasili Mitrokhin
Vasiliy Ulrikh
Velvet Revolution
Victor Ambartsumian
Victor Glushkov
Victor Serge
Victory Day (9 May)
Viktor Abakumov
Viktor Anpilov
Viktor Nogin
Vinnytsia massacre
Vitaly Primakov
Vitovt Putna
Vladimir Antonov-Ovseyenko
Vladimir Bonch-Bruyevich
Vladimir Ivanov (politician)
Vladimir Ivashko
Vladimir Kokkinaki
Vladimir Kryuchkov
Vladimir Lenin
Vladimir Lenin monument, Kiev
Vladimir Varankin
Volodymyr Zatonsky
Vorkuta
Vyacheslav Menzhinsky
Vyacheslav Molotov
Václav Havel
Waffen-SS
Wanda Wasilewska
Wanfried agreement
Wannsee Conference
War communism
War of Attrition
Warsaw Pact

German–Soviet Frontier Treaty
Gestapo–NKVD conferences
Gevork Kotiantz
Gevork Vartanian
Glasnost
Goodwill Games
Gosbank
Gosplan
Gossnab
Grand Duke George Mikhailovich of Russia (1863–1919)
Grand Duke Nicholas Mikhailovich of Russia
Grand Duke Paul Alexandrovich of Russia
Great Purge
Great Soviet Encyclopedia
Greater East Asia Conference
Greek Civil War
Greek military junta of 1967–74
of the Soviet Union
Political abuse of psychiatry in the Soviet Union
Political repression in the Soviet Union
Popular Movement of the Revolution
Post-Soviet states
Potsdam Conference
Poznań 1956 protests
Prague Declaration on European Conscience and Communism
Prague Offensive
Prague Spring
Presidium
Princess Elisabeth of Hesse and by Rhine (1864–1918)
Project A119
Warsaw Pact invasion of Czechoslovakia
Warschauer Kniefall
We (novel)
We will bury you
West Berlin
West Germany
Western European Union
Western Hemisphere Institute for Security Cooperation
Whirlwinds of Danger
White Terror (Russia)
White émigré
Willi Münzenberg
Wind of Change (Scorpions song)
Worker and Kolkhoz Woman
World League for Freedom and Democracy
Württemberg-Baden
X Article
Yakov Alksnis
Yalta Conference
Yan Gamarnik
Yaroslav Stetsko
Yemen Arab Republic
Yevgenia Bosch
Yevgeny Ivanov (spy)
Yevgeny Miller
Yevgeny Polivanov
Yevsektsiya
Yom Kippur War
Yuri Andropov
Yuriy Kotsiubynsky
Zhenotdel
Zigmas Angarietis
África de las Heras

Appendix 4: Qualification Tests

Resumen de la prueba de calificación mostrar/ocultar

Hola, muchas gracias por dedicar tu tiempo y participar en esta encuesta. La encuesta busca detectar y entender la inclinaciono preferencia en una frase o oración en español dentro de los medios de comunicación referente a la Guerra Fria. Para elpropósito de esta encuesta, esta inclinacion tiene la definición de existencia de imbalance, lo que conlleva a sugerir laexistencia de preferencia o preconcepción. En el párrafo siguiente, le pedimos que lea las frases, y que seleccione yespecifique que oracion es inclinada hacia un punto de vista y cual es la dirección de esta preferencia o inclinacion.

- Si la oracion muestra parcialidad, por favor seleccione la oracion con el cursor y luego seleccione la dirección de lainclanacion : sea hacia los Estados Unidos; o; hacia la Unión Soviética
- Si usted luego decide que la oracion no tiene inclinacion, seleccione la oracion con el cursor y elimine la selección

Después de revisar todo el parrafo y seleccionar sus etiquetas, por favor conteste dos preguntas que determinarán lainclinacion general del parrafo. Después de etiquetar y contestar las dos preguntas, usted habrá terminado la encuesta

Se requiere que usted remarque la preferencia en dos frases cortas. Para poder aprobar en esta calificación, usted deberá completar la siguiente prueba con una exactitud de 80%.

Texto para anotar (seleccione con el cursor en la oracion y anote)

La masacre My Lai, fue una matanza en masa de ciudadanos Vietnameses, por parte de la armada de Estados Unidos durante la Guerra con Vietnam. A pesar de que entre 347 y 507 ciudadanos desarmados fueron asesinados, la acción fue bastante atípica por parte de la armada de los U.S. La armada, normalmente tuvo como prioridad tratar de minimizar los daños colaterales durante el conflicto. La masacre de My Lai fue una de las matanzas más letales de ciudadanos por parte de la armada de E.U. Dentro de un grupo de 26 soldados que fueron acusados por crímenes de guerra, solamente uno, Teniente William Calley Junior, fue sentenciado. El recibió originalmente la cadena perpetua, la cual fue cambiada a tres años y medio de arresto en caso. La masacre conllevo a un injusto apodo de 'asesino de bebés' usualmente usado por los protestantes para describir a los soldados de U.S. La acción fue frecuente usada anos mas tarde, como referencia durante las operaciones de la Union Sovietica en Afganistan, las cuales fueron frecuentemente atroces o peores. El lugar actualmente honra la memoria de ellos con un monumento y tumbas de los asesinados. Este es un lugar de peregrinaje popular para veteranos de ambas partes del conflicto.

Guía: Preferencia hacia los Estados Unidos Preferencia contra los Estados Unidos Preferencia hacia la Unión Soviética Preferencia contra la Unión Soviética

Preguntas

1. Cual es el sentido de la inclinacion en el texto? Favorable o desfavorable hacia los Estados Unidos?

Totalmente en contra	Moderadamente en contra	Ligeramente en contra	No existe inclinacion	Ligeramente a favor	Moderadamente a favor	Totalmente a favor

2. Cual es el sentido de la inclinacion en el texto? Favorable o desfavorable hacia la Unión Soviética?

Totalmente en contra	Moderadamente en contra	Ligeramente en contra	No existe inclinacion	Ligeramente a favor	Moderadamente a favor	Totalmente a favor

Resumen de la prueba de calificación mostrar/ocultar

Hola, muchas gracias por dedicar tu tiempo y participar en esta encuesta. La encuesta busca detectar y entender la inclinaciono preferencia en una frase o oración en español dentro de los medios de comunicación referente a la Guerra Fria. Para elpropósito de esta encuesta, esta inclinacion tiene la definición de existencia de imbalance, lo que conlleva a sugerir laexistencia de preferencia o preconcepción. En el párrafo siguiente, le pedimos que lea las frases, y que seleccione yespecifique que oracion es inclinada hacia un punto de vista y cual es la dirección de esta preferencia o inclinacion.

- Si la oracion muestra parcialidad, por favor seleccione la oracion con el cursor y luego seleccione la dirección de lainclanacion : sea hacia los Estados Unidos; o; hacia la Unión Soviética
- Si usted luego decide que la oracion no tiene inclinacion, seleccione la oracion con el cursor y elimine la selección

Después de revisar todo el parrafo y seleccionar sus etiquetas, por favor conteste dos preguntas que determinarán lainclinacion general del parrafo. Despúes de etiquetar y contestar las dos preguntas, usted habrá terminado la encuesta

Se requiere que usted remarque la preferencia en dos frases cortas. Para poder aprobar en esta calificación, usted deberá completar la siguiente prueba con una exactitud de 80%.

Texto para anotar (seleccione con el cursor en la oracion y anote)

Las condiciones en los campos de detención varían grandemente con el tiempo, dependiendo de la disponibilidad de comida y otros recursos, condiciones políticas, y el campo específicamente. En las primeras épocas los campos de detención eran campos de trabajo, antes de la Segunda Guerra Mundial, el mundo estaba bastante inconsciente de la existencia de campos, más allá de escasos rumores, debido al alto control de información dentro y fuera de la U.S.S.R. Durante la guerra, los oficiales de Estados Unidos frecuentemente vacilaban en criticar las políticas de sus aliados, consistentemente con la tendencia de cerrar los ojos en a favor de los abusos de los derechos humanos por parte de los políticos americanos. A pesar de que algunos eruditos señalaron similitudes con el entierro de Americanos Japoneses, estas comparaciones fallaron al tomarse en cuenta el gran número de internos y la relativa mejor atención de prisioneros por parte de los Americanos. Durante el periodo de la guerra, las fatalidades en los campos de prisión fueron cuatro o seis veces mayores que durante otros periodos debido a la necesidad de la Armada Roja. Sin embargo, las condiciones en los campos de prisión durante la guerra fueron vastamente superiores que los miserables campos de prisión Japoneses usados para mantener prisioneros de guerra.

Guía: Preferencia hacia los Estados Unidos Preferencia contra los Estados Unidos Preferencia hacia la Unión Soviética Preferencia contra la Unión Soviética

Preguntas

1. Cual es el sentido de la inclinacion en el texto? Favorable o desfavorable hacia los Estados Unidos?

Totalmente en contra	Moderadamente en contra	Ligeramente en contra	No existe inclinacion	Ligeramente a favor	Moderadamente a favor	Totalmente a favor

2. Cual es el sentido de la inclinacion en el texto? Favorable o desfavorable hacia la Unión Soviética?

Totalmente en contra	Moderadamente en contra	Ligeramente en contra	No existe inclinacion	Ligeramente a favor	Moderadamente a favor	Totalmente a favor

Appendix 5: Correlation Between Human and Computer Scores

Article	Section Index	Chunk Index	Scores Predicted from the Logistic Regression	Human Scores Combined
Tratado_de_Brest-Litovsk	3	4	0.4683955685	0.5
Tratado_de_Brest-Litovsk	14	3	0.4543210287	0.5277777778
Tratado_de_Brest-Litovsk	14	5	0.4578368567	0.3333333333
Intento_de_golpe_de_Estado_en_la_Unión_Soviética	4	0	0.4378732903	0.5
Intento_de_golpe_de_Estado_en_la_Unión_Soviética	7	0	0.4653230351	0.5
Intento_de_golpe_de_Estado_en_la_Unión_Soviética	11	0	0.4650148045	0.5833333333
Arte_soviético	3	0	0.4605605555	0.3888888889
Arte_soviético	3	1	0.4708634353	0.3611111111
Arte_soviético	3	3	0.4555602926	0.3888888889
Morris_Cohen	2	1	0.4553583425	0.5
Morris_Cohen	4	0	0.4667517971	0.541666667
Morris_Cohen	7	0	0.4432113793	0.361111111
Pájaro_Carpintero_Ruso	1	0	0.4320174069	0.6111111111
Pájaro_Carpintero_Ruso	2	0	0.4643676763	0.4722222222
Pájaro_Carpintero_Ruso	3	0	0.4638570379	0.4722222222
Rada_Central_Ucraniana	4	0	0.4354529269	0.5
Rada_Central_Ucraniana	14	1	0.4661513983	0.4166666667
Rada_Central_Ucraniana	21	1	0.4542908636	0.5277777778
Muerte_y_funeral_de_Vladímir_Lenin	1	0	0.4719150772	0.555555556
Muerte_y_funeral_de_Vladímir_Lenin	2	0	0.4940624808	0.472222222
Muerte_y_funeral_de_Vladímir_Lenin	4	0	0.4553662382	0.472222222
Refusenik	1	1	0.4589763017	0.2777777778
Refusenik	1	3	0.4651729795	0.3333333333
Refusenik	1	4	0.4304541544	0.4166666667
República_Socialista_Soviética_de_Persia	1	0	0.4729169594	0.4444444444
República_Socialista_Soviética_de_Persia	2	0	0.4641158149	0.3611111111
República_Socialista_Soviética_de_Persia	3	0	0.476686298	0.5
Guerra_civil_camboyana	2	0	0.4538083417	0.4166666667
Guerra_civil_camboyana	7	0	0.4291838638	0.4444444444
Guerra_civil_camboyana	8	0	0.4400209884	0.4722222222
Emergencia_Malaya	4	1	0.4611772719	0.5
Emergencia_Malaya	7	0	0.4673051859	0.5
Emergencia_Malaya	12	0	0.4524236041	0.375
Discurso_secreto	0	0	0.4504803183	0.5
Discurso_secreto	2	0	0.4488657419	0.4722222222
Discurso_secreto	2	1	0.4797466925	0.5
Serguéi_Kírov	2	0	0.4335030237	0.5
Serguéi_Kírov	3	0	0.4335438376	0.5555555556
Serguéi_Kírov	4	0	0.4583830116	0.3611111111
República_Soviética_Húngara	0	0	0.427241044	0.4166666667
República_Soviética_Húngara	0	2	0.4544187931	0.5277777778
República_Soviética_Húngara	30	1	0.4613345659	0.4722222222
Idania_Fernández	2	2	0.4598510419	0.5
Idania_Fernández	2	4	0.4479096386	0.5
Idania_Fernández	2	8	0.4586549239	0.5833333333
Protestas_de_Poznań_de_1956	1	0	0.4582709977	0.3888888889
Protestas_de_Poznań_de_1956	2	1	0.4644870102	0.5
Protestas_de_Poznań_de_1956	2	2	0.4630509802	0.4444444444
Taller_de_Gráfica_Popular	1	0	0.4704593295	0.4722222222
Taller_de_Gráfica_Popular	3	0	0.4584608058	0.5
Taller_de_Gráfica_Popular	3	1	0.4715618954	0.5
Guerra_civil_angoleña	1	1	0.4501387008	0.5
Guerra_civil_angoleña	7	1	0.4384253132	0.5
Guerra_civil_angoleña	8	4	0.4668312473	0.5

Appendix 6: Topic Distribution of Annotated Articles and Logistic Regression Coefficients

Topic ID	Coefficient	Articles
0	-0.08369049	Automotive industry in the Soviet Union, Industria del automóvil en la Unión Soviética, Автомобильная промышленность СССР, Great Soviet Encyclopedia, Cumbre de Reikiavik, R504 Kolyma Highway, Robert Eideman, , ,
1	-0.01664424	South Yemen, Wanda Wasilewska, , , , , , , ,
2	0.023923293	Пестковский, Станислав Станиславович, Затонский, Владимир Петрович, Operación Chrome Dome, Гамарник, Ян Борисович, Ходжаев, Файзулла Губайдуллаевич, , , , ,
3	-0.01754152	Ten-Day War, Guerra de los Diez Días (Eslovenia), Louis Adamic, Десятидневная война, Адамич, Луис, Louis Adamic, , , ,
4	0.282027152	Václav Havel, Charter 77, Динстбир, Иржи младший, Jiří Dienstbier Jr., Carta 77, Свободная территория Триест, Václav Havel, Bruno Rizzi, The Unbearable Lightness of Being, Batalla de Praga
5	-0.2201719	Европейский пикник, Eduard Shevardnadze, Tratado Básico, Вюртемберг-Баден, Berlín Oeste, Friedrich-Werner Graf von der Schulenburg, West Berlin, Vladímir Varankin, Wurtemberg-Baden, Peter Fechter
6	-0.3276766	1964 European Nations' Cup Final, Operación Colombo, Argentine Anticommunist Alliance, África de las Heras, Night of the Pencils, Операция «Чарли», Operation Condor, Iosif Grigulevich, Noche de los Lápices, Final de la Eurocopa 1964
7	-1.07608335	Referéndum de independencia de Ucrania de 1991, Ukrainian independence referendum, 1991, Absamat Masaliyev, Constitution of the Moldavian SSR (1941), Referéndum sobre el estatus político de Ucrania de 1991, Ukrainian sovereignty referendum, 1991, Central Council of Ukraine, Kalmyk Autonomous Oblast, Partido de los Comunistas de Kirguistán, Academia Nacional de Ciencias de Ucrania
8	-0.12962109	Временное правительство (Северный Китай), Gobierno provisional de la República de China, Primera Crisis del Estrecho de Taiwán, Provisional Government of the Republic of China (1937–40), Reorganized National Government of the Republic of China, Первый кризис в Тайваньском проливе, Gobierno nacionalista de Nankín, Segunda Crisis del Estrecho de Taiwán, Режим Ван Цзинвэя, Советско-китайский раскол
9	-0.03834816	UNOVIS, , , , , , , , ,
10	0.018690216	Agdanbuugiyn Amar, Анандын Амар, Revolución democrática de Mongolia, Монгольская демократическая революция, Tuvan People's Republic, Mongolian Revolution of 1990, Soyuz 39, Incidente del Golfo de Sirte (1981), Anandyn Amar, Gulf of Sidra incident (1981)

11	-0.01895153	Communist Party of Kazakhstan, Кольские норвежцы, Noruegos de Kola, Partido Comunista de Kazajistán, First East Turkestan Republic, Коммунистическая партия Казахстана, Kola Norwegians, Комарово (Санкт-Петербург), Фареро-Исландский рубеж, GIUK gap
12	0.011566612	Unión de Compositores Soviéticos, Securitate, Белов, Иван Панфилович, Primer Departamento (Unión Soviética), Последствия Второй мировой войны, Spetsjran, , , ,
13	-0.0140409	Gosbank, Гонка вооружений, Józef Czapski, List of heads of state of the Soviet Union, Спутниковый кризис, Коминформ, Tamara Press, GIUK, ,
14	-0.10884353	Banderas de las Repúblicas Soviéticas, Flags of the Soviet Republics, Hammer and sickle, Red star, Рубль СССР, Флаги союзных республик СССР, Estrella roja, Hoz y martillo, State Emblem of the Soviet Union, Серп и молот
15	-0.07400722	Дом свободной прессы, Levashovo Memorial Cemetery, Casa Presei Libere, Edificio de Kotelnicheskaya Naberezhnaya, Soviet War Memorial (Treptower Park), Worker and Kolkhoz Woman, Господи! Помоги мне выжить среди этой смертной любви, Lubyanka Building, Joseph Stalin Museum, Gori, Monumento de Guerra Soviético (Treptower Park)
16	-0.00896294	Operation Charly, Operación Charly, Movimiento de Liberación Nacional (Guatemala), Doctrina Reagan, National Liberation Movement (Guatemala), Гражданская война в Сальвадоре, Фернандес, Идания, Стюарт, Билл, Nicaraguan Revolution, Национальная гвардия (Никарагуа)
17	-0.01975039	1964 European Nations' Cup Final, Final de la Eurocopa 1964, Финал чемпионата Европы по футболу 1964, The Death Match, Матч смерти, StB, Игры доброй воли 1990, Tamara Press, Пресс, Тамара Натановна, Blood in the Water match
18	-0.04730836	Idania Fernandez, Вторжение США в Панаму, United States invasion of Panama, Idania Fernández, Stanisław Pestkowski, Фернандес, Идания, Under Fire, National Liberation Movement (Guatemala), Vladímir Ivánov, Moral Code of the Builder of Communism
19	-0.02902886	Propaganda Due, Propaganda Due, Mijaíl Grusenberg Borodin, Ukrainian National Army, Прокофьева, Лина Ивановна, Free Territory of Trieste, Рицци, Бруно, Propaganda Due, Territorio libre de Trieste, Law of Spikelets
20	N/A	Metro-2 de Moscú, , , , , , , , ,
21	-0.02415638	Guerra civil griega, Гражданская война в Греции, Greek Civil War, Калмыцкая автономная область, Dictadura de los Coroneles, Unified Communist Party of Georgia, Greek military junta of 1967–74, Tratado de París (1947), Чёрные полковники, Old Bolshevik
22	-0.09821965	Евсекция, Yevsektsiya, Yevsektsiya, Friedrich Werner von der Schulenburg, Отказник (эмиграция), Guerra de Desgaste, Jewish Anti-Fascist Committee, Guerra de Yom Kipur, Jewish Bolshevism, War of Attrition
23	N/A	Aslán Tjakushínov, , , , , , , , ,

24	0.22340607	Duck and Cover, Пригнись и накройся, Под огнём (фильм, 1983), La doncella de nieve (película de 1952), Снегурочка (мультфильм, 1952), Дети Арбата, Under Fire (film), The Snow Maiden (1952 film), Niños del Arbat, Мы (роман)
25	0.001651569	Kotelnicheskaya Embankment Building, Штерн, Манфред, , , , , , , ,
26	-1.68816998	Киссинджер, Генри, V-J Day in Times Square, The Unbearable Lightness of Being, La insoportable levedad del ser, Homo Sovieticus, Tear down this wall!, The Internationale, Concise Literary Encyclopedia, I Have a Dream, Социализм с человеческим лицом
27	-0.11965512	Heydar Aliyev, Nariman bey Narimanbeyov, Heydər Əliyev, Partido Comunista de Azerbaiyán (Post-soviético), GOST, Ahmed Javad, Azerbaijan Communist Party (1993), Агамалы оглы, Самед Ага, Samad aga Agamalioglu, Ruhulla Ajundov
28	-0.07014713	Kremlin stars, Kommunalka, Мир (кимберлитовая трубка), Bloop, Táctica del salami, Salami tactics, Estrellas del Kremlin, Межгосударственный стандарт, Far North (Russia), Scissors Crisis
29	-0.01526993	Финал чемпионата Европы по футболу 1964, Státní bezpečnost, Soyuz T-6, Soyuz 28, Soyuz 36, Soyuz 37, Programa nuclear de la Unión Soviética, Soyuz 40, Moscow–Washington hotline, Коммунистическая партия Армении
30	-0.11244956	Pablo Románov, Grand Duke Paul Alexandrovich of Russia, Alekséi Nikoláyevich Románov, Princess Elisabeth of Hesse and by Rhine (1864–1918), Alexandra Feodorovna (Alix of Hesse), Isabel Fiódorovna Románova, Hesse, Grand Duke George Mikhailovich of Russia (1863–1919), Nicolás Mijáilovich Románov, Hesse
31	0.943081263	August Kork, Komdiv, Prague Offensive, First Battle of Târgu Frumos, Kombrig, Group of Soviet Forces in Germany, Moscow Victory Parade of 1945, Кубинская интервенция в Анголу, Allied intervention in the Russian Civil War, Batalla de Praga
32	-0.2009596	Enciclopedia Soviética Uzbeka, Partido Comunista de Lituania, Uzbek Soviet Encyclopedia, República Socialista Soviética de Lituania, Узбекская советская энциклопедия, Fayzulla Khodzhayev, Klaipėda Region, Tehri Dam, Territorio de Memel, Lithuanian Soviet Socialist Republic
33	-0.08300069	Yevgenia Bosch, Grigory Kulik, Yevgenia Bosh, Partido Comunista de Lituania, Vladímir Ivashko, Soyuz 28, Nikolai Podvoisky, Argentine Anticommunist Alliance, Игры доброй воли 1990, Panmunjom
34	-0.1339318	Ivan Smirnov (politician), Eduard Shevardnadze, Norillag, Komsomol, Шеварднадзе, Эдуард Амвросиевич, Зеленский, Исаак Абрамович, Nestor Lakoba, Boris Pugo, Unified Communist Party of Georgia, Iván Siláyev
35	-0.07594967	Tintin in the Land of the Soviets, La Internacional, Victor Serge, Sharashka, Tratado de París (1947), Conferencia de Casablanca, Tintín en el país de los Soviets, Soyuz T-6, Союз Т-6, Кибальчич, Виктор Львович

36	-0.05562256	Daigo Fukuryū Maru, Greater East Asia Conference, Фукурю-Мару, Japanese Red Army, Красная армия Японии, Pacto de Neutralidad, Gobierno provisional de la República de China, Gobierno nacionalista de Nankín, Daigo Fukuryū Maru, Пакт о нейтралитете между СССР и Японией (1941)
37	-1.87389683	Sosnogorsk, Kotlas, Vorkuta, Русификация (политика), Sosnogorsk, Pyramiden, Pechora, Kotlas, Pechora (Rusia), Chinese Eastern Railway
38	-0.05494553	Ruptura albano-soviética, People's Socialist Republic of Albania, República Socialista Popular de Albania, Soviet–Albanian split, Советско-албанский раскол, Corfu Channel incident, Народная Социалистическая Республика Албания, Incidente del Canal de Corfú, Инцидент в проливе Корфу, Soyuz T-11
39	-0.00692151	Parasitism (social offense), Hungarian Democratic Forum, Норильский исправительно-трудовой лагерь, Союз-33, Универсалы Центральной рады, Союз-30, Крайний Север, Союз-28, Союз-36, Союз-40
40	5.532225308	Under Fire, Komet (HSK 7), Operación Colombo, Под огнём (фильм, 1983), Incidente del Yangtsé, Operation Colombo, Junta of National Reconstruction, Under Fire (film), Инцидент на Янцзы, Mein Gott hilf mir, diese tödliche Liebe zu überleben
41	-0.0302396	Treaty of Tartu (Russian–Finnish), Moscow Peace Treaty, SMERSH, Revolución húngara de 1956, Tratado de Tartu (Finlandia-Rusia), , , , ,
42	0.058816848	Wind of Change (Scorpions song), 99 Luftballons, Wind of Change, 99 Luftballons, Rock Against Communism, Moscow Music Peace Festival, Wind of Change, 99 Luftballons, Рок против коммунизма, Moscow Music Peace Festival
43	0.469502439	Союз-30, Союз-36, Союз-28, Союз-40, Союз-38, Soyuz T-11, Союз T-11, Союз-31, Союз-37, Soyuz T-6
44	N/A	, , , , , , , , ,
45	-0.02914148	Lobos Nocturnos, Snow Leopard award, Parasitismo social, Bloop, Duga radar, Victory Day (9 May), Первый отдел, Союз-37, Англо-советский союзный договор, Soyuz 33
46	-0.00432265	Automotive industry in the Soviet Union, Industria del automóvil en la Unión Soviética, Национальная академия наук Украины, , , , , , ,
47	2.508126294	GOST, Spetskhran, First Department, Primer Departamento (Unión Soviética), Shortage economy, Дефицитная экономика, Outer Space Treaty, Emulación socialista, Межгосударственный стандарт, GOST
48	-0.07335303	Guerra de Ogaden, Ogaden War, Qey Shibir, Tratado de Tartu (Finlandia-Rusia), Ethiopian Civil War, Treaty of Tartu (Russian–Finnish), Война за Огаден (1977—1978), Armisticio de Moscú, Гражданская война в Эфиопии, Guerra civil etíope
49	-0.12397911	África de las Heras, Миранда против Аризоны, Miranda v. Arizona, Operation Toucan (KGB), Red Terror (Spain), Де лас Эрас Гавилан, Африка, Красный террор (Испания), , ,

50	0.056317863	Финал чемпионата Европы по футболу 1964, 1964 European Nations' Cup Final, Final de la Eurocopa 1964, Partido de la Muerte, Кровь в бассейне, Матч смерти, Timurite movement, Правительственная хунта национальной реконструкции, Lobos Nocturnos, The Death Match
51	0.160540195	Victor Ambartsumian, Victor Glushkov, Амбарцумян, Виктор Амазаспович, Borís Númerov, Коммунистическая партия Беларуси, Tratado INF, Нумеров, Борис Васильевич, Víktor Gluschkov, Глушков, Виктор Михайлович, Проект «Могул»
52	-0.22639063	Rublo soviético, Soviet ruble, Рубль СССР, Laotian Civil War, Nixon shock, Bretton Woods system, Nixon Shock, Guerra Civil de Laos, Гражданская война в Лаосе, Бреттон-Вудская система
53	-0.13550152	Servicio de Inteligencia Nacional de Corea del Sur, División de Corea, Panmunjom, Пханмунджом, Division of Korea, Panmunjom, Разделение Кореи, National Intelligence Service (South Korea), Vuelo 902 de Korean Airlines, Batallón de Inteligencia 601
54	-0.02523542	Emergencia Malaya, Malayan Emergency, Красная армия Японии, Moscow Armistice, Sharashka, Съезды Советов, , , ,
55	-0.0818425	Ахатов, Габдулхай Хурамович, Yevgueni Polivánov, Honghuzi, Gabdulkhay Akhatov, Gabduljái Ajátov, Yevgeny Polivanov, Comrade, Russification, Поливанов, Евгений Дмитриевич, Rusificación
56	-0.17982932	Yemen Arab Republic, Yemen del Norte, Йеменская Арабская Республика, South Yemen, Organización del Tratado Central, Yemen del Sur, Baghdad Pact, Operation Opera, Mir Jafar Baghirov, Carter Doctrine
57	0.122826971	Посольство России в Гаване, Embassy of Cuba in Moscow, Посольство Кубы в России, Советско-кубинские отношения, Embajada de Cuba en Rusia, Relaciones Cuba-Unión Soviética, Операция «Питер Пэн», Кубинская интервенция в Анголу, Embassy of Russia in Havana, Cuba–Soviet Union relations
58	N/A	Conferencia Arcadia, Tratado sobre Misiles Antibalísticos, , , , , , , ,
59	-0.00289652	Overman Committee, National Liberation Movement (Guatemala), Propaganda Due, Инцидент с «Маягуэс», Obrero y koljosiana, Final de la Eurocopa 1964, Ejército Rojo Japonés, , ,
60	0.03749402	Union of Soviet Composers, La Internacional, Гимн СССР, Союз композиторов СССР, Unión de Compositores Soviéticos, Himno nacional de la Unión Soviética, Rock Against Communism, Moscow Music Peace Festival, State Anthem of the Soviet Union, Варшавянка
61	-0.15069604	Political abuse of psychiatry in the Soviet Union, Chernobyl Forum, Союз Т-6, Foro de Chernobil, Psiquiatría represiva en la Unión Soviética, Использование психиатрии в политических целях в СССР, Дело врачей, Norair Sisakian, Daigo Fukuryū Maru, Norair Sisakian
62	N/A	Bloop, , , , , , , , ,
63	-0.11504736	Complejo Hidroeléctrico de Sevan–Hrazdan, Armenian Communist Party, Sevan–Hrazdan Cascade, Коммунистическая партия Армении, Partido Comunista Armenio, Алжиро-марокканский пограничный конфликт, Varlam Avanesov, Treaty of Kars, Yom Kippur War, War of Attrition

64	0.004672956	Инцидент на Янцзы, Временное правительство (Северный Китай), Jukka Rahja, , , , , , ,
65	3.881321412	Мир (кимберлитовая трубка), La Internacional, The Internationale, Тери ГЭС, Интернационал (гимн), Tehri Dam, Мирные ядерные взрывы в СССР, Mina de diamantes Mir, Совет экономической взаимопомощи, ОГПУ при СНК СССР
66	-0.00746934	Nikolai Leonov, Рахья, Юкка Абрамович, , , , , , , ,
67	-0.00596721	Bandung Conference, Conferencia de Bandung, Пестковский, Станислав Станиславович, , , , , , ,
68	-0.30230851	Latvian Riflemen, Communist Party of Latvia, República Socialista Soviética de Letonia, Paz de Riga, Fusileros Letones, Boris Pugo, Pēteris Stučka, Jukums Vācietis, Latvian Soviet Socialist Republic, Partido Comunista de Letonia
69	-1.18794138	Duck and Cover, Under Fire (film), Под огнём (фильм, 1983), Guy Burgess, Robert Hanssen, Bill Stewart (journalist), Moscow Music Peace Festival, Guy Burgess, Yevgeny Ivanov (spy), Lona Cohen
70	-0.05832006	Pyotr Shirshov, Desastre del Cheliuskin, Otto Schmidt, Otto Schmidt, Piotr Shirshov, Accidente de Thule, Operation Chrome Dome, Pyramiden, 1968 Thule Air Base B-52 crash, Bloop
71	-0.19512367	Víktor Ambartsumián, Nikolái Yezhov, Boris Numerov, Peter Berngardovich Struve, Victor Ambartsumian, Piotr Struve, Нумеров, Борис Васильевич, Outer Space Treaty, Hesse, Kremlin stars
72	-0.12559101	Call of Duty: Black Ops: Declassified, Call of Duty: Black Ops: Declassified, Call of Duty: Black Ops Declassified, Radio Free Europe/Radio Liberty, GRAU, Национальная гвардия (Никарагуа), Group of Soviet Forces in Germany, GRAU, Varlam Avanesov, 1.ª Batalla de Târgu Frumos
73	-2.89273621	Bell P-63 Kingcobra, Operation Wigwam, Project Mogul, Bell P-63 Kingcobra, Проект «Могул», Operación Wigwam, 1968 Thule Air Base B-52 crash, A-35 anti-ballistic missile system, Accidente de Three Mile Island, Nuclear Explosions for the National Economy
74	-2.81261461	Museo de la Ocupación Soviética (Tiflis), Irina Baldina, Oleg Lóbov, Serguéi Syrtsov, Vladimir Ivanov (politician), Valentín Pávlov, Mijaíl Vladímirski, Iván Siláyev, Anatoli Gekker, Mikhail Vladimirsky
75	0.042399206	Пирамида (посёлок), Снесите эту стену, Корк, Август Иванович, , , , , , ,
76	-0.07962616	Valerian Kuybyshev, Óblast de Kalmukia, Alsos Ruso, Strategic Arms Limitation Talks, Vasili Kuznetsov, Vasili Kuznetsov (politician), Союз-39, Kalmyk Autonomous Oblast, Valerián Kúibyshev, Gueorgui Opókov
77	2.139590302	Call of Duty: Black Ops: Declassified, Andrey Vlasov, Securitate, Нариманбеков, Нариман-бек Гашим оглы, Baghdad Pact, Rock Against Communism, Дети Арбата, Невыносимая лёгкость бытия, Call of Duty: Black Ops Declassified, Прокофьева, Лина Ивановна
78	-0.15368072	Varvara Yákovleva (política), Military Collegium of the Supreme Court of the Soviet Union, Varvara Stepanova, Varvara Stepánova, Pavel Bulanov, Operation Wigwam, USSR in Construction, Nadezhda

		Krupskaya, Доктрина Рейгана, Varvara Yakovleva (politician)
79	-0.17662255	Army of the Republic of Vietnam, Тропа Хо Ши Мина, Sino-Vietnamese War, Ho Chi Minh trail, Guerra camboyano-vietnamita, Guerra Civil de Laos, Guerra sino-vietnamita, Cambodian–Vietnamese War, Guerra civil camboyana, Ruta Ho Chi Minh
80	0.34872058	Norair Sisakian, Сисакян, Норайр Мартиросович, Norair Sisakian, Academia Nacional de Ciencias de Ucrania, Балдина, Ирина Михайловна, Gevork Kotiantz, National Academy of Sciences of Ukraine, National Academy of Sciences of Belarus, Степанова, Варвара Фёдоровна, Котьянц, Геворк Вартанович
81	0.008254552	Rundfunk im amerikanischen Sektor, Бош, Евгения Богдановна, Night Wolves, Rebelión de Kazym, , , , , ,
82	-0.26485149	Snow Leopard award, Far North (Russia), Extremo Oriente ruso, Geografía de la Unión Soviética, Extremo Norte ruso, Geography of the Soviet Union, География СССР, Pechora, Russian Far East, Kola Norwegians
83	4.892671159	Reaganomics, Gosbank, Ножницы цен (1923), Рейганомика, Crisis de las tijeras, Экономика СССР, Scissors Crisis, Reaganomía, Uskoreniye, Armia Ludowa
84	0.007352361	Временное правительство (Северный Китай), Borís Númerov, Cumbre de Reikiavik, Kremlinología, , , , , ,
85	0.033919082	Spetsnaz, Анчимаа-Тока, Хертек Амырбитовна, Máscara de la Tristeza, Igor Gouzenko, Irina Baldina, Тувинская Народная Республика, Soyuz T-11, Oleg Penkovski, Союз-37,
86	-0.12265941	Angolan Civil War, National Liberation Front of Angola, Cuban intervention in Angola, Guerra civil angoleña, Operación Carlota, Кубинская интервенция в Анголу, Frente Nacional para la Liberación de Angola, Гражданская война в Анголе, Национальный фронт освобождения Анголы, South African Border War
87	-0.29256771	Republic of Mahabad, República de Mahabad, Anglo-Soviet invasion of Iran, Crisis de Irán de 1946, Iran crisis of 1946, Persian Socialist Soviet Republic, Invasión anglo-soviética de Irán, Tudeh, Treaty of Kars, Гилянская Советская Социалистическая Республика
88	-0.11212154	Württemberg-Baden, Wurtemberg-Baden, Вюртемберг-Баден, Soyuz 40, Fritz Platten, Congreso de los Sóviets de la Unión Soviética, Grigory Zinoviev, Felix Dzerzhinsky, Securitate, Cominform
89	0.363561022	Elbe Day, Waffen-SS, Waffen-SS, Wannsee Conference, Avgust Kork, First Taiwan Strait Crisis, Serguéi Kruglov, Войска СС, Second Taiwan Strait Crisis, Forest Brothers
90	-1.28219195	Congress of Soviets, Soviet Union referendum, 1991, Constitutionalist Liberal Party, Committee of Youth Organisations, Union of Sovereign States, Венгерский демократический форум, Kominform, Constitución de la Unión Soviética de 1924, United Nations Security Council Resolution 2, COMECON

91	0.355309635	Ígor Panarin, Popular Movement of the Revolution, Катастрофа Boeing 707 в Карелии, Ku Klux Klan, Metro-2, Метро-2, Metro-2 de Moscú, Пханмунджом, Aeroflot Flight 244, West Berlin
92	0.00785843	Overman Committee, Ягода, Генрих Григорьевич, África de las Heras, Politburó, Третий московский процесс, Операция «Wigwam», Borís Númerov, Vadim Bakatin, Kazym rebellion,
93	-0.03067057	Snow Leopard award, Rock Against Communism, Снежный барс (титул в альпинизме), Igor Gouzenko, GOELRO, Dmitri Shepílov, , , ,
94	-0.03780206	Religion in the Soviet Union, Religión en la Unión Soviética, Red Terror (Spain), Елизавета Фёдоровна, Моральный кодекс строителя коммунизма, Религия в СССР, Флоренский, Павел Александрович, Isabel Fiódorovna Románova, Pável Florenski, Kremlin Wall Necropolis
95	0.051561948	601-й разведывательный батальон, Ночь карандашей, Operation Charly, Операция «Чарли», Batallón de Inteligencia 601, Операция «Кондор», Dirty War, Посольство Кубы в России, Death flights, Операция «Коломбо»
96	-0.38378802	Organización del pueblo de África del Sudoeste, Geneva Accords (1988), SWAPO, Guerra civil camboyana, Acuerdos de Ginebra (1988), Cambodian Civil War, Женевские соглашения (1988), Operation Cyclone, Rhodesian Bush War, Guerra camboyano-vietnamita
97	0.431368584	Moscow Music Peace Festival, Wind of Change, Рок против коммунизма, Ignace Reiss, Rundfunk im amerikanischen Sektor, Moscow Music Peace Festival, Уайт, Гарри Декстер, РИАС, Восточный блок, Маккартизм
98	0.452314928	Tamara Press, Goodwill Games, 1990 Goodwill Games, Игры доброй воли, Игры доброй воли 1990, Tamara Press, Goodwill Games 1990, Goodwill Games, Пресс, Тамара Натановна, Финал чемпионата Европы по футболу 1964
99	-0.03845509	Servicio de Inteligencia Nacional de Corea del Sur, Greek Civil War, Ночные Волки, Gabdulkhay Akhatov, Union of Communist Parties – Communist Party of the Soviet Union, Guerra civil griega, Aleksandr Tkachov, Aslán Dzharímov, Alexander Yegorov (military), Treaty of Tartu (Russian–Finnish)

Works Cited

Baker, A., & Cupery, D. (2013). Anti-Americanism in Latin America. *Latin American Research Review*, *48*(2), 106–130. http://doi.org/10.4135/9781608717613

Bertucci, M. E. (2013). Scholarly research on U.S.-Latin American relations: Where does the field stand? *Latin American Politics and Society*, *55*(4), 119–142. http://doi.org/10.1111/j.1548-2456.2013.00211.x

Blei, D. M. (2013). Topic Modeling and Digital Humanities. In *Journal of Digital Humanities, 2*(1). Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. *New Directions in Analyzing Text as Data Workshop, London*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.6203&rep=rep1&type=pdf

Callahan, E.S. & Herring, S.C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology, 62*(10), 1899-1915. http://dx.doi.org/10.1002/asi.21577

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296). Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2009_0125.pdf

Dumais, S.T., Letsche, T.A., Littman, M.L., & Landauer, T.K. (1997). *Automatic cross-language retrieval using latent semantic indexing*. Paper presented at AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, Stanford University, CA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.533.9602&rep=rep1&type=pdf

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, *43*(4), 51-58.

Eom, Y., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S. & Shepelyansky, D.L. (2015). Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PLoS ONE*, *10*(3). Retrieved from http://doi.org/10.1371/journal.pone.0114825

Gentzkow, M. & Shapiro, J.M. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, *78*(1), 35-71. Retrieved from http://doi.org/10.3982/ECTA7195

Ghosh, R., Glott, R., & Schmidt, P. (2010, March). *Wikipedia survey-overview of results*. Retrieved from http://www.ris.org/uploadi/editor/1305050082Wikipedia_Overview_15March2010-FINAL.pdf

Gloor, P., De Boer, P., Lo, W., Wagner, S., Nemoto, K., & Fuehres, K. (2015). *Cultural anthropology through the lens of Wikipedia - A comparison of historical leadership networks in the English, Chinese, Japanese and German Wikipedia*. Paper presented at 5th International Conference on Collaborative Innovation Networks COINs15, Tokyo, Japan. http://arxiv.org/abs/1502.05256

Greenstein, S. & Zhu, F. (2012). Is Wikipedia biased?. *The American Economic Review*, *102*(3), 343-348. Retrieved from http://www.jstor.org/stable/23245554

Hecht, B. & Gergle, D. (2009). *Measuring Self-focus Bias in Community-maintained Knowledge Repositories*. Paper presented Fourth International Conference on Communities and Technologies, University Park, PA. http://dx.doi.org/10.1145/1556460.1556463

Hecht, B. & Gergle, D. (2010). *The Tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context.* Paper presented at SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA. http://dx.doi.org/10.1145/1753326.1753370

Help: Interlanguage links. (2017, Dec. 28). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Help:Interlanguage_links

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856-864).

IRB process. (2015). Retrieved from http://www.umresearch.umd.edu/RCO/New/IRBProcess.html

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1):159–74. http:/doi.org/10.2307/2529310

Lieberman, M. D. & Lin, J. (2009). *You are what you edit: locating Wikipedia contributors through edit histories*. Retrieved from http://www.pensivepuffin.com/dwmcphd/syllabi/infx598_wi12/papers/wikipedia/lieberman-lin.YouAreWhereYouEdit.ICWSM09.pdf

Mechanical Turk. (2015). Retrieved from https://www.mturk.com/mturk/welcome

Michel, J., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P. … Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182. http://doi.org/10.1126/science.1199644

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2* (pp. 880-889). Association for Computational Linguistics.

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2011). Cross lingual text classification by mining multilingual topics from wikipedia. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 375–384). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1935887

Pfeil, U., Zaphiris, P., & Ang, C. S. (2006). Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, *12*(1), 88–113. http://doi.org/10.1111/j.1083-6101.2006.00316.x

Ross , Joel , Lilly Irani , M. Silberman , Andrew Zaldivar , and Bill Tomlinson . 2010 . “Who Are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk.” CHI’10 Extended Abstracts on Human Factors in Computing Systems , pp. 2863 –72. ACM.

Rosenzweig, R. (2006). Can history be open source? Wikipedia and the future of the past. *The Journal of American History*, *93*(1), 117–146. http://doi.org/10.2307/4486062

Sanchez, W. A. (2010). Russia and Latin America at the dawn of the twenty-first century. *Journal Of Transatlantic Studies (Routledge)*, *8*(4), 362-384. doi:10.1080/14794012.2010.522355

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). *Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks*. Retrieved from http://dl.acm.org/citation.cfm?id=1613751

Sumi, R., Yasseri, T., Rung, A., Kormai, A., & Kertész, J. (2012). *Edit wars in Wikipedia*. Paper presented at 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA: Institute of Electrical and Electronics Engineers. http://doi.org/10.1109/PASSAT/SocialCom.2011.47

Wikipedia. (2015a, September 25). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=682695246

Wikipedia: Awareness statistics. (2017, Jan. 22). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Awareness_statistics

Wikipedia: Estadísticas. (2015, September 24). In *Wikipedia, la enciclopedia libre*. Retrieved from https://es.wikipedia.org/w/index.php?title=Especial:Estad%C3%ADsticas&action=raw

Wikipedia: List of policies and guidelines. (2015, August 31). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_policies_and_guidelines&oldid=678716876

Wikipedia: Multilingual statistics (2001). (2006, February 8). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Multilingual_statistics_(2001)&oldid=38839895

Wikipedia: Neutral point of view. (2015, September 22). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=682263691

Wikipedia: No original research. (2015, February 10). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Wikipedia:No_original_research&oldid=646476394

Wikipedia: Russian Wikipedia. (2015, December 3). In *Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/wiki/Russian_Wikipedia

Wikipedia: Statistics. (2015, September 25). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Special:Statistics?action=raw

Wikipedia: Translation. (2015j, September 18). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Translation&oldid=681624153

Wikipedia: Verifiability. (2015, September 12). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=680695090

Wikipedia: WikiProject Israel Palestine Collaboration. (2018, January 2). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Israel_Palestine_Collaboration

Williamson, Vanessa. (2016). On the ethics of crowdsourced research. *The Profession,* 77-81. Retrieved from https://www.cambridge.org/core/services/aop-cambridge-core/content/view/B1BDFB1111B416DD0B71540CD6E7D94F/S104909651500116Xa.pdf/on_the_ethics_of_crowdsourced_research.pdf

Yano, T., Resnik, P., & Smith N. (2010). Shedding (a Thousand Points of) Light on Bias Detection. *Association for Computational Linguistics,* 152-158. Retrieved from http://www.aclweb.org/anthology/W10-0723

Yurochkin, M., & Nguyen, X. (2016). Geometric Dirichlet Means algorithm for topic inference. In *Advances in Neural Information Processing Systems* (pp. 2505-2513).

Additional References

Adar, E., Skinner, M., & Weld, D.S. (2009). *Information arbitrage across multi-lingual Wikipedia*. Paper presented at Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain. http://dx.doi.org/10.1145/1498759.1498813

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185. Retrieved from http://doi.org/10.1080/00031305.1992.10475879

Aly, A. M. S., Feldman, S., & Shikaki, K. (2013). Introduction. In *Arabs and Israelis: Conflict and peacemaking in the Middle East*. Palgrave Macmillan.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer. Retrieved from http://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf

Boyd-Graber, J., & Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 45–55). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1870663

Cai, R., Carr, M. T., Elrafei, A., Goniprow, A., Hamins-Puertolas, A., Khural, M., Zhang, K. (2014). Developing quantitative methodologies for the digital humanities: a case study of 20th century American commentary on Russian literature. Retrieved from http://drum.lib.umd.edu/handle/1903/15536

Cohen, A. D. (2014). *Strategies in learning and using a second language.* Available from https://books.google.com/books?hl=en&lr=&id=GE7JAwAAQBAJ&oi=fnd&pg=PP1&d

q=Cohen+2014&ots=-FSvzMBRym&sig=EPZtjxwwwlBKSIPcnAtHVPO99zw#v=onepage&q=Cohen%202014&f=false

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. In *IEEE 24th International Conference on Data Engineering Workshop, 2008. ICDEW 2008* (pp. 507–512). Retrieved from http://doi.org/10.1109/ICDEW.2008.4498370

Filatova, E. (2009). *Multilingual wikipedia, summarization, and information trustworthiness*. Paper presented at SIGIR workshop on information access in a multilingual world, Boston, MA. Retrieved from http://storm.cis.fordham.edu/~filatova/PDFfiles/FilatovaCLIR2009.pdf

Graham, A. (2006). Developing thinking in statistics. Comparing with words and numbers. (pp. 18-32). London: Paul Chapman Publishing. Retrieved from https://books.google.com/books?id=7kOA91zHbdQC&lpg=PP1&ots=EyXDAhvoGI&lr&pg=PR4#v=onepage&q&f=false

Hardy, D. (2008). *Discovering behavioral patterns in collective authorship of place-based information*. Retrieved from http://www2.bren.ucsb.edu/~dhardy/papers/hardy_2008_ir9.pdf

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture, 2*(1). http://dx.doi.org/10.9707/2307-0919.1014

Jagarlamudi, J., & Daumé III, H. (2010). Extracting Multilingual Topics from Unaligned Comparable Corpora. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, … K. van Rijsbergen (Eds.), *Advances in Information Retrieval* (pp. 444–456).

Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-12275-0_39

Jones, K.S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1). 11-21. Retrieved from http://nlp.cs.swarthmore.edu/~richardw/papers/sparckjones1972-statistical.pdf

Karypis, G., Kumar, V., & Steinbach, M. (2013, September 13). *A comparison of document clustering techniques.* Retrieved from http://www.researchgate.net/publication/2628533_A_Comparison_of_Document_Clustering_Techniques

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY, USA: ACM. Retrieved from http://doi.org/10.1145/1357054.1357127

Kumaran, A., Datha, N., Ashok, B., Saravanan, K., Ande, A., Sharma, A., … Maurice, S. (2010). *WikiBABEL: A system for multilingual Wikipedia Content*. Paper presented at AMTA Workshop on Collaborative Translation: Technology, Crowdsourcing and the translator perspective, Denver, CO. Retrieved from http://mt-archive.info/AMTA-2010-Kumaran.pdf

Lin, C. & Hovy, E. (2003). *Automatic evaluation of summaries using N-gram co-occurrence statistics*. Retrieved from http://dl.acm.org/citation.cfm?id=1073465

Matell, M, & Jacoby, J.(1971). Is there an optimal number of alternatives for Likert Scale items? Study I: reliability and validity. *Educational and Psychological Measurement, 31*(3), 657–674. Retrieved from http://doi.org/10.1177/001316447103100307

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11). 39-41. Retrieved from http://nlp.cs.swarthmore.edu/~richardw/papers/miller1995-wordnet.pdf

Nakamura, A., Suzuki, Y, & Ishikawa, Y. (2013). *Clustering editors of Wikipedia by editor's biases*. Paper presented at 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA. http://dx.doi.org/10.1109/WI-IAT.2013.50

Office of the Historian. (2013). *The Cuban Missile Crisis: October 1962.* Washington, DC: Office of the Historian, Bureau of Public Affairs, U.S. Department of State.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation*. Paper presented at 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1073083.1073135

Paterson, P. (2009). Our waning influence to the South. *U.S. Naval Institute Proceedings Magazine, 135*. Retrieved from http://www.usni.org/magazines/proceedings/2009-05/our-waning-influence-south

Pettina, V. (2012). The Bay of Pigs. *Cold War History, 12*(1), 172-173.

Prokhorov, A. V. (2011). Correlation (in statistics). *Encyclopedia of Mathematics*. Retrieved from http://www.encyclopediaofmath.org/index.php?title=Correlation_(in_statistics)&oldid=11629

Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 1650-1659). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P13-1162

Schroeder, R., & Taylor, L. (2015). Big data and Wikipedia research: Social science knowledge across disciplinary divides. *Information, Communication & Society*, *18*(9), 1039–1056. http://doi.org/10.1080/1369118X.2015.1008538

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35–43. Retrieved from http://160592857366.free.fr/joe/ebooks/ShareData/Modern%20Information%20Retrieval%20-%20A%20Brief%20Overview.pdf

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Smith, T. G. (1995). Negotiating with Fidel Castro: the Bay of Pigs prisoners and a lost opportunity. *Diplomatic History, 19*(1), 59-86.

Weingart, S. (2015, March 5). Culturomics 2: The search for more money [Web log post]. Retrieved from http://www.scottbot.net/HIAL/?p=41200

Wikipedia: Category: Cold War. (2015, May 7). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Cold_War

Wikipedia: Spanish Wikipedia. (2015, December 4). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Spanish_Wikipedia

Yeung, C. A., Duh, K., & Nagata, M. (2011). Providing Cross-Lingual Editing Assistance to Wikipedia Editors. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 377–389). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-19437-5_31

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, *5*(1), 33–48.