IDEAH • Vol. 2, Iss. 1 (DHSI 2019 & 2020)

Identifying Relevant Topics in Ecuadorian Presidential Speeches from 2007 – 2020

Luis Meneses¹

¹University of Victoria

Published on: Aug 19, 2021

DOI: 10.21428/f1f23564.b3ecc3fe

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Introduction

In this paper, I present an exploratory analysis for identifying the common topics in Ecuadorian Presidential speeches from thirteen years (between 2007 and 2020) by drawing inferences over those thirteen years using topic modelling. I will also contrast the Ecuadorian speeches with speeches given by Fidel Castro, whose ideology has had great influence in Ecuador and Latin America. I will address three research questions in this paper. First, as the two Ecuadorian Presidents in those years, how far apart are the political ideologies of Correa and Moreno? Second, can we identify any overlapping topics in their ideologies? And finally, are the ideologies of Correa and Moreno similar to the political ideals of Castro?

Some of the political discourses in the document corpus that I will be analyzing in this paper have been taken offline. Thus, this paper contributes methods that can be used to explore sensitive situations where a document corpus is not readily available; these situations are increasingly relevant in political scenarios. My exploratory analysis using topic modelling yielded two core results. First, it outlined two document clusters that are clearly delimited with multiple intersecting topics; and second, it showed that increasing the topic modelling iterations increases the prominence of the overlap between the topics. The goals of this paper are to engage in an exploratory analysis of the document corpus, identify the common trends among the three sets of discourses, and emphasize the strengths of our field, while addressing how my research questions fit into the interdisciplinary conversation of the digital humanities.

Background

La República del Ecuador is a small country in South America. It is 283561 km², making it 35 times smaller than Canada. It has a population of approximately 17 million people. The official language is Spanish, and the country uses the US Dollar as their official currency. Ecuador's economy is highly dependent on commodities, namely petroleum and agricultural products. Ecuador has a very unique ecological heritage, hosting many endangered plants and animals, including those found in the Galápagos Islands.

Politically, Ecuador has been on an interesting and peculiar path since its return to democracy in 1979. I would even dare to say that it is currently in a state of political transition. After a series of ousted governments in the early 2000s, Rafael Correa was elected as the President of Ecuador from 15 January 2007 to 24 May 2017. He was

immediately succeeded by Lenin Moreno, a member of his political party and one of his former Vice Presidents.

Correa's political ideology was labeled as the "Socialism of the 21st Century," and it was heavily influenced by Cuba's socialism and by Fidel Castro's style of government as Prime Minister and President from 1959 to 2008. Correa's presidency was characterized by a difficult relationship with the press and was not shy of controversies. In 2020, Correa was found guilty of bribery and sentenced to serve eight years in prison. This sentence has not been executed on Correa because he left Ecuador right after his mandate ended. He is currently living in Belgium.

As it can be expected, Moreno has tried to distance himself from Correa with different policies on journalism, freedom of speech, and tackling corruption. The Ecuadorian Presidency removed Correa's discourses and they are no longer available on their website, which is shown in the screenshots in Figures [1](#) and [2](#). However, the political undercurrent of Moreno's government has been very similar to the one established by Correa.

Figure 1: Screenshot of the website of the Ecuadorian Presidency in 2017. Retrieved from <https://www.presidencia.gob.ec/discursos/>.

Figure 2: Screenshot of the website of the Ecuadorian Presidency in 2018. Retrieved from <https://www.presidencia.gob.ec/discursos/>.

Methodology

I have systematically harvested the presidential speeches given by Correa, Moreno and Castro. These speeches were published online by the Ecuadorian Presidency (Presidencia de La República Del Ecuador) and by the Cuban Government (República de Cuba). Interestingly, Correa’ s speeches were taken offline in 2017. I downloaded the discourses before they were taken offline from the website of the Ecuadorian Presidency using a Python 3 script (van Rossum 1–65). I chose to use Python because of its Beautiful Soup library (Richardson), which I used to identify and extract the HTML links to the discourses. Once the relevant links were identified by the parser, the script issued an HTTP request to download each document using the Python Requests library, “an elegant and simple HTTP library for Python” (Reitz). All the Ecuadorian discourses were originally in Microsoft Word format, and their text was extracted using the Python-based library docxpy (Shah) and saved in plain text files. Plain text files are portable, and this ensures that their contents are accessible and readable in the future.

Since Fidel Castro’ s discourses were available in HTML format, I did not use a custom script to obtain them. I downloaded Castro’ s discourses using Wget which is “a free software package for retrieving files that employs widely used Internet protocols” (Free Software Foundation). Castro’ s discourses were available in different language. I chose Spanish for consistency and to set a common baseline with the other discourses. The text from Castro’ s discourses was extracted using Beautiful Soup and saved into plain text

files as well. As a whole, the document corpus contained 2307 Presidential speeches: 806 from Rafael Correa, 360 from Lenin Moreno, and 1141 from Fidel Castro.

One limitation of these methods is the potential introduction of errors during harvesting and text extraction of the documents. But this is less of a concern with the libraries I employed because they are widely used and tested by the Python community. Wget is also widely used among Unix users and distributed with current Linux distributions. These two factors minimize this limitation and strengthen the soundness of my data.

Preliminary Analysis

I conducted two preliminary experiments as part of the exploratory analysis to get a better sense of the similarities between the speeches. Employing a form of stylometric analysis, a quantitative study of literary style that assumes authors write in consistent and uniquely recognizable ways, the first experiment analyzed Correa's and Moreno's speeches and employed an analysis based on the chi-square statistic. The chi-square statistic is one of the measures that can be used to compare sets of documents (Kilgariff). In this case, I used this statistic to measure the distance between the vocabularies employed in the two sets of discourses to determine their authorship. The more similar the vocabularies are in the two sets of texts, the more probable it is that the same author wrote them both.

To carry out this analysis I generated a random sample from the discourses. I selected 100 documents from each set corresponding to Correa and Moreno, while leaving 50 documents for the test subset and making sure there was no overlap between the training and testing subsets. I then proceeded to tokenize the documents by putting all text into lowercase, remove stop words, and removing punctuation. To do this I used Python's Natural Language Toolkit (NLTK Project). Figure 3 shows the results of running 100 iterations of this analysis. If the values represented in the histogram for Correa and Moreno were closer to 1, that would mean that the analysis detected statistical evidence of similarities in the speeches. In my opinion, the results of this analysis are inconclusive. This may indicate that both sets of discourses were written by different groups of political aides.

Figure 3: Results of analysis using the chi-square statistic.

My second preliminary experiment used cosine similarity, which is often used to measure document similarity in text analysis. Cosine similarity measures the similarity between documents, by calculating the cosine of the angle between the projection of its vectors in a multi-dimensional space. For this analysis I used a random sample of 100 documents from one set (Correa' s or Moreno' s discourses) against one document from the other set. I also tokenized, lowercased, and removed stop words and punctuation using Python' s Natural Language Toolkit. Figure 4 shows the results of running this analysis for 1000 iterations. The results of this analysis can be interpreted as the document sets not sharing many similarities. However, cosine similarity can provide an inaccurate measure of the associations between documents (Meneses et al. 333).

Figure 4: Results of the analysis using cosine similarity.

Topic Modelling

Topic modelling can solve the problem of finding and retrieving related information, which is the reason I used it to answer my research questions. Topic modelling “not only helps the researcher to determine the trending themes or related fields with respect to their field of interest but also helps them to identify new concepts and fields over time” (Lamba and Madhusudhan 477). Topic modelling is useful when dealing with a large number of documents, particularly if they share a common theme. In other words, as David M. Blei notes, “topic modelling algorithms are statistical methods that analyze the words in texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time” . An important feature of topic modelling is that it organizes and summarizes documents at a scale that would be impossible for humans to do manually, or that would require great deal of effort to do so (Blei).

Latent Dirichlet Allocation (commonly known as LDA) is the simplest form of topic modelling (Blei et al.). In our paper, “Is Falstaff Falstaff? Is Prince Hal Henry V?,” Laura Estill and I explain that,

The main idea behind LDA is that the text in documents do not belong to a single topic exclusively, but can belong to multiple topics at the same time [. ...] Formally, a topic can be defined as a probability distribution over a fixed vocabulary. For example, a topic about English literature has words about literary works with high

probability, whereas a topic about computer science has words about engineering with a high probability. (3-4)

LDA can be explained using two steps. First, the algorithm randomly chooses a distribution of words over topics. And second, for each word in the document the algorithm randomly chooses a topic from the distribution over topics in step 1. As part of this second step, the algorithm also randomly chooses a word from the corresponding distribution over the vocabulary. Estill and I go on to explain that

[t]his process is usually carried out in multiple iterations. Therefore, a topic model with greater iterations is usually considered to provide a better representation of the data [. ...] Consequently, the output of an LDA model usually consists of the probability distribution of documents and terms over different topics (4).

One of the affordances of topic modelling is that it can point researchers towards new avenues for inquiry and validating research questions. Keyword and thematic analysis have been a mainstay of humanities criticism, and topic modelling offers a mode of literary analysis that expands the understanding of word use beyond themes and pre-selected keywords. This latter form of inquiry assumes that scholars know what terms or topics are worth searching. Topic modelling, on the other hand, provides scholars with possible terms and topics to pursue further. Topic modelling, of course, is not analysis unto itself, but it can promote further analysis of a document corpus.

In the context of the other studies within the digital humanities, Lisa Rhody has shown how LDA topic models can fail for figurative language. Estill and I discuss elsewhere the concerns expressed by Andrew Goldstone and Ted Underwood: “despite the popularity of topic modeling and LDA models for critical inquiry, the ‘black box’ nature of topic models and algorithms warrants attention” (4). In this sense, I do not intend to present a model as a way to offer a final analysis, but to present a different method for approaching a document corpus. Along these lines, I have applied LDA to TEI-encoded plays, which allowed the analysis of speeches by individual characters (Estill and Meneses), and used it for metadata generation in a bilingual corpus (Meneses, Arbuckle, et al.).

In the exploratory analysis that I am describing in this paper, I chose to use LDA because of its relative simplicity compared to other topic modelling algorithms. My reasoning is to fully take advantage of the affordances that topic modelling provides as a machine learning technique and discover the relationships within a collection of documents, which falls right under the scope of LDA. Since I am dealing with a collection of political

discourses these relations translate to the hidden connections between the ideas expressed in these public addresses.

I used Gensim for this analysis. Gensim is a “Python library for topic modelling, document indexing and similarity retrieval with large corpora” (Řehůřek). Gensim is widely used in the the natural language processing and information retrieval communities, though it is gaining prominence in other communities as well. I also used two additional libraries in my work, lxml and pyLDAvis to help me interpret the results. Lxml is a Python library which allows for easy handling of XML to record the output from the topic modelling (lxml); PyLDAvis (Mabey), is a Python library for interactive topic model visualization.

Results

I applied LDA to the discourses of Correa, Moreno, and Castro to investigate the relationships between the documents, their potential clustering, and if patterns started to emerge. My expectation was to see clusters emerge using this unsupervised modelling method that were related to the socialist concept of equality, government, and progress as a nation. This expectation is connected to my research questions, as I hypothesized that the three sets of speeches shared similar political ideologies. Following this hypothesis, which assumed that the discourses shared a common theme and will have multiple overlapping topics, I applied LDA to the discourses using 20 topics and 50 iterations as parameters as a starting point. In this sense, choosing a larger number of topics would have not aligned with my initial hypothesis. The output of this initial experiment is shown in Figure 5. Two clusters of documents are very prominent in this visualization.

Figure 5: Visualization of the results of the first experiment, using 20 topics and 50 iterations as parameters.

After this initial experiment, I decided to reduce the number of topics while increasing the precision. As I have stated earlier in this paper, a topic model with greater iterations is usually considered to provide a better representation of the data. The result of the modelling using 20 topics as a parameter showed that there was a lot of overlap between the topics. My intention with reducing the number of topics was to gain a concise overview of the documents in the corpus. More specifically, I used a similar analysis to the initial experiment but using 15 topics and 100 iterations as parameters. The output of this second experiment is shown in Figure 6. As expected, more details are starting to surface as a result of the increased number of iterations. In the visualization of the results using this new set of parameters, two clusters are still very identifiable, but topic 5 shows very little in common with the others, becoming almost an outlier.

Figure 6: Visualization of the results of the second experiment, using 15 topics and 100 iterations as parameters.

The results from the second experiment showed that there were many points of intersection between the topics, so I decided to reduce the number of topics even further to gain a better sense of the underlying themes the documents in the corpus. As a third experiment, I reduced the number of topics further while retaining the same level of precision from the second experiment: employing 10 topics and 100 iterations as parameters. A visualization of the results from this third experiment is shown in Figure 7. In the visualization of the results, two clusters are clearly delimited, with topic 5 being completely separated from the others. I interpret this as refinement in the clustering obtained in the second experiment.

Figure 7: Visualization of the results of the third experiment, using 10 topics and 100 iterations as parameters.

Discussion

The results from the experiments using LDA make it clear that the political discourses from Correa, Moreno, and Castro share similar topics. The three experiments described in the results section show the overlap that exist between their discourses: outlining two document clusters that are clearly delimited with multiple intersecting topics. This overlap tells us that the discourses do share common ideologies, and that topic modelling (as a machine learning technique that can help discover the relationships within a collection of documents) was an appropriate methodology for the exploratory analysis that I have described in this paper. This overlap becomes more evident as the precision of the experiments is increased by a larger number of iterations. In the third experiment (10 topics, 100 iterations) there are clear commonalities between topics 1, 2, 3, and 4. The term distribution for these topics include the following translated terms: “revolution,” “people,” “government,” “country,” “production,” and “system.”

It can be argued that topic 5 is an outlier with respect to the other topics. Topic 5 deals mostly with economic programs, the environment, and strategic alliances between countries. Looking further at the document distribution for topic 5, the documents with the higher probability are from Correa in the second experiment (15 topics, 100

iterations). In the third experiment with 10 topics, all three Presidents are equally represented in topic 5.

There are topics in all three experiments that did not show commonalities, and these topics can be considered outliers. Topic 8 in the second experiment (15 topics, 100 iterations) falls into this category, and deals with overall themes of freedom of speech and the press. The translated terms with the highest representation are “media,” “freedom,” “newspapers,” “press,” “communication,” “journalists,” “television,” and “information.” The documents in this topic with the highest probability belong to Correa and Moreno, with Fidel Castro having minimal representation in this topic. Correa’s government was characterized by having a difficult relationship with the press, so I find this topic particularly interesting.

PyLDavis was a crucial asset to the analysis that I have outlined in this paper, allowing the topics extracted in the three experiments to be visualized in a web browser. However, I must address that these visualizations can potentially introduce some bias into the analysis, hindering some details in its interpretation. Analyzing the raw data obtained from the topic modelling (XML files) can provide a closer interaction with the results of the analysis and reveal further details. This is a direction for future iterations of this work, along with expanding the corpus of documents to allow comparisons with other Presidents such as Hugo Chavez and Nicolás Maduro from Venezuela, and Evo Morales from Bolivia.

Conclusions

In this paper, I have taken the first steps towards identifying the common topics in the Presidential speeches of Rafael Correa, Lenin Moreno, and Fidel Castro. As answers to the research questions that I have stated in the introduction of this paper, I can conclude that the ideologies of Correa and Moreno are not very different from each other, showing clearly overlapping topics. My analysis also shows that there are similarities with the late Fidel Castro from Cuba.

Additionally, the analysis that I have outlined in this paper fulfills its objective of exploring sensitive situations where a document corpus is not readily available. Once a document corpus disappears, which is quite frequent for online materials, it is impossible to analyze it and draw inferences from it. Ideally, I would have analyzed the entire set of discourses from all the Presidents of Ecuador, presenting a more complete study and a potentially complex analysis spanning close to two centuries. In this sense, I had to make compromises to overcome the limitation of the availability of data. My work

has explored a data set from different perspectives and drawn inferences, which in turn emphasizes one of the key strengths of the digital humanities, and addresses how its methodologies fit into its interdisciplinary conversation and are increasingly relevant to analyze political scenarios.

Works Cited

Blei, David M. “Probabilistic Topic Models.” *Communications of the ACM*, vol. 55, no. 4, Apr. 2012, pp. 77–84. *ACM Digital Library*, doi.org/10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research*, vol. 3, 2003, pp. 993–1022.

Estill, Laura, and Luis Meneses. “Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare’s Plays.” *Digital Studies/Le Champ Numérique*, vol. 8, no. 1, Jan. 2018, pp. 1–22, doi.org/10.16995/dscn.295.

Free Software Foundation. “GNU Wget.” *GNU Operating System*, 2018, www.gnu.org/software/wget/.

Goldstone, Andrew, and Ted Underwood. “The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us.” *New Literary History*, vol. 45, no. 3, 2014, pp. 359–84. *Project MUSE*, doi.org/10.1353/nlh.2014.0025.

Kilgariff, Adam. “Comparing Corpora.” *International Journal of Corpus Linguistics*, vol. 6, Nov. 2001, pp. 97 – 133. *ResearchGate*, doi.org/10.1075/ijcl.6.1.05kil.

Lamba, Manika, and Margam Madhusudhan. “Mapping of Topics in DESIDOC Journal of Library and Information Technology, India: A Study.” *Scientometrics*, vol. 120, no. 2, Aug. 2019, pp. 477–505. *Springer Link*, doi.org/10.1007/s11192-019-03137-5.

lxml. *lxml - XML and HTML with Python*. 2018, lxml.de/.

Mabey, Ben. *PyLDAvis: Interactive Topic Model Visualization. Port of the R Package*. 2.1.2, 2018, github.com/bmabey/pyLDAvis.

Meneses, Luis, et al. “Aligning Social Media Indicators with the Documents in an Open Access Repository.” *KULA: Knowledge Creation, Dissemination, and Preservation Studies*, vol. 3, no. 1, Feb. 2019, p. 19, doi.org/10.5334/kula.44.

—. “Restoring Semantically Incomplete Document Collections Using Lexical Signatures.” *Research and Advanced Technology for Digital Libraries*, edited by Trond Aalberg et al., Springer, 2013, pp. 321–32. Springer Link, doi.org/10.1007/978-3-642-40501-3_33.

NLTK Project. *NLTK 3.6.2 Documentation*. 2015, www.nltk.org.

Presidencia de La República Del Ecuador. *Presidencia de La República Del Ecuador» Discursos*. www.presidencia.gob.ec/discursos/. Accessed 15 Oct. 2020.

Řehůřek, Radim. *Gensim: Topic Modelling for Humans*. radimrehurek.com/gensim/. Accessed 12 Apr. 2017.

Reitz, Kenneth. *Requests: HTTP for HumansTM*. 2019, requests.readthedocs.io/en/master/.

República de Cuba. *Discursos e Intervenciones Del Comandante En Jefe Fidel Castro Ruz*. www.cuba.cu/gobierno/discursos/. Accessed 1 Oct. 2020.

Rhody, Lisa. “Topic Modeling and Figurative Language.” *Journal of Digital Humanities*, vol. 2, no. 1, Winter 2012, journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Richardson, Leonard. “Beautiful Soup: We Called Him Tortoise Because He Taught Us.” *Crummy*, www.crummy.com/software/BeautifulSoup/.

Shah, Ankush. *Docxpy: A Pure Python-Based Utility to Extract Text, Hyperlinks and Images from Docx Files*. 0.8.5, 2017. PyPI, github.com/badbye/docxpy.

van Rossum, Guido. *Python Tutorial*. 1.2, Stichting Mathematisch Centrum, 1996. ir.cwi.nl/pub/5007/05007D.pdf.