Computational Content Analysis.
Camacho Jonathan E.
Research Project.

Introduction

From 2000 to 2008, Venezuela has experienced a rare increase of migration out of the country. (Mayda, 2010; Subero, 2012) Most Venezuelan scholars cite political and economic factors as motivating reasons for Venezuelans to migrate. (Alvarez, 2012) However, in recent ethnographies such as (Alvarez, 2012; Subero, 2012) , crime and insecurity; and political and economic reasons have taken a prominent place among the interviewees as a reason for Venezuelans to leave the country. For this reason, I want to explore if the perception about crime could be a critical factor among the reasons motivating Venezuelans to consider leaving the country.

This proposal uses quantitative data from Latinobarometro to compare the political, economic, and crime perceptions of Venezuelans and Ecuadorians. I hypothesize that the perception that crime frequency has increased in the country, especially in Venezuela, correlates with an increase of the intentions to emigrate. This study sheds light on the migration phenomenon in the region and in general. In particular, expanding the understanding of migration is important since according to demographers, migration will increase in the 21st century due to demographic growth, in incomes, and in security and human rights. (Fitzgerald, Leblang, & Teets, 2014; Martin, 2015) Furthermore, this study contributes to understanding an important side of migration, the subjective side that has been little explored since most research takes a structural deterministic outlook. (Fernández, 2006) Finally, this study contributes importantly to the field of migration and Cognitive Sociology; the latter a growing subfield in Sociology.

Developing a Corpora.

This first section of the report drawn techniques from two first weeks of class: Intro and Corpus Linguistics. First step for this analysis was to develop a corpus using twenty-four interviews of Venezuelans living in the USA and Canada. The interviews are part of a qualitative project I was part of. The twenty-four surveys were already transcribed, so they only needed to be translated from Spanish to English. This process took between seventeen and twenty hours. After translations were done, they were checked by an independent reader. The independent reader randomly selected seven interviews to check if the English translations matched the Spanish original transcript.

Finally, after the twenty-four interviews were translated, I created a data frame utilizing Microsoft Excel version 15.32 for Mac. Then the data frame containing the data was saved as a .csv file. Then, the data was loaded into the Jupyter Notebook "Migrants Perspectives" using a python script into the data frame *raw_data*. The data frame contains four columns (*gender, gender_code, arrival_year, and text*). Most of the data frame variables names are self-explanatory except the variable *gender_code* which is a binarization of the variable *gender.* The text for these interviews is the variable *text* to facilitate processing such as tokenization. Below there is a snapshot of the first five registers.

Figure 1. *raw_data* data frame loaded in Python.

	gender	gender_code	year	text
1	m	1	1997	Only four interviewees in my research used the...
2	f	0	1998	The story of Luisa, husband, and children in A...
3	f	0	2000	There is an atmosphere of paranoia that is liv...
4	m	1	2001	Rafael Mirabal gave his family a lifestyle in ...
5	m	1	2001	Luis Alvaray, 43, is a social communicator. It...

Processing the Text (Filtering and Normalizing Text)

To start the process of comparison between the different interviews, started by filtering and normalizing the text. This process of normalization and filtering was conducting using the function provided in class two (Corpus Linguistic): *normlizeTokens*. This function lowers case of words, drops the non-word tokens, removes some 'stop words' in English, and stems the remaining words to remove suffixes and prefixes (using the porter and snowball stemmers form nltk), and lemmatize (using SnowballStemmer from nltk because the porter stemmer seems to do a poor job stemming some tokens) tokens by grouping the inflected forms of the same word. Then, the resulting normalized tokens are added to the new column: *normalized_tokens*. Bellow there is a snapshot of the data frame with the two new columns.

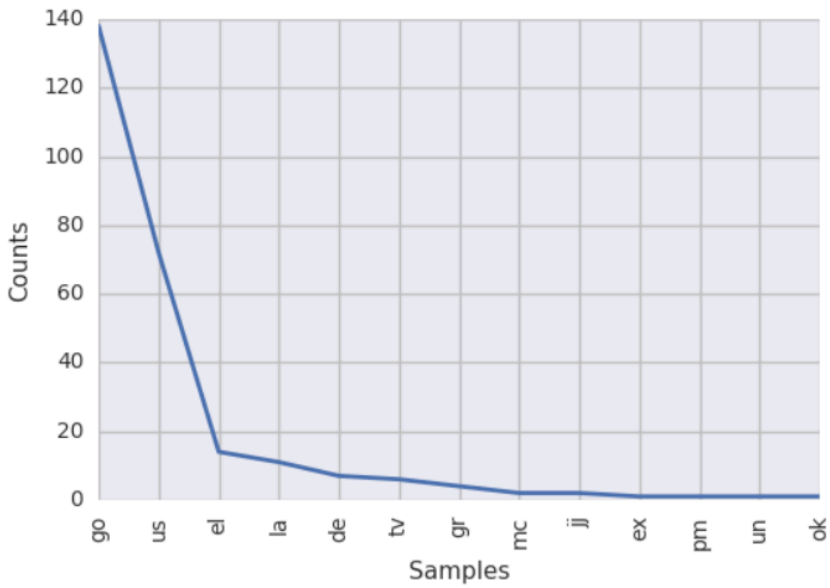
Figure 2. *raw_data* data frame with tokenized columns added.

	gender	gender_code	arrival_year	text	tokenized_text	normalized_tokens	normalized_tokens_count
1	m	1	1997	Only four interviewees in my research used the...	[Only, four, interviewees, in, my, research, u...	[onli, four, interviewe, research, use, phrase...	655
2	f	0	1998	The story of Luisa, husband, and children in A...	[The, story, of, Luisa, ,, husband, ,, and, ch...	[stori, luisa, husband, children, atlanta, geo...	1537
3	f	0	2000	There is an atmosphere of paranoia that is liv...	[There, is, an, atmosphere, of, paranoia, that...	[atmospher, paranoia, live, even, outsid, coun...	673
4	m	1	2001	Rafael Mirabal gave his family a lifestyle in ...	[Rafael, Mirabal, gave, his, family, a, lifest...	[rafael, mirab, gave, famili, lifestyl, caraca...	341
5	m	1	2001	Luis Alvaray, 43, is a social communicator. It...	[Luis, Alvaray, ,, 43, ,, is, a, social, commu...	[lui, alvaray, social, communic, attract, life...	192

Now that the text in the data frame *raw_data* is normalized, it is possible to start analyzing it. First, I start by finding frequency distributions for all the interviews together. For this process, I used the "TheConditionalFreqDist" class which reads tuples of a condition and a focal word. As a condition, I utilized the lengths of the words. The total number of words according to this class is 12254 across all the interviews.

Then, I wanted to know how words of different lengths were distributed across all the interviews. I tried several lengths, and some of them were difficult to visualize or were meaningless when printed, particularly word with a length between the interval three to eleven characters. Words lenngths that were particularly interesting were words of two and twelve characters. I tried with words' length above twelve characters but they are not represented in the interviews corpus.

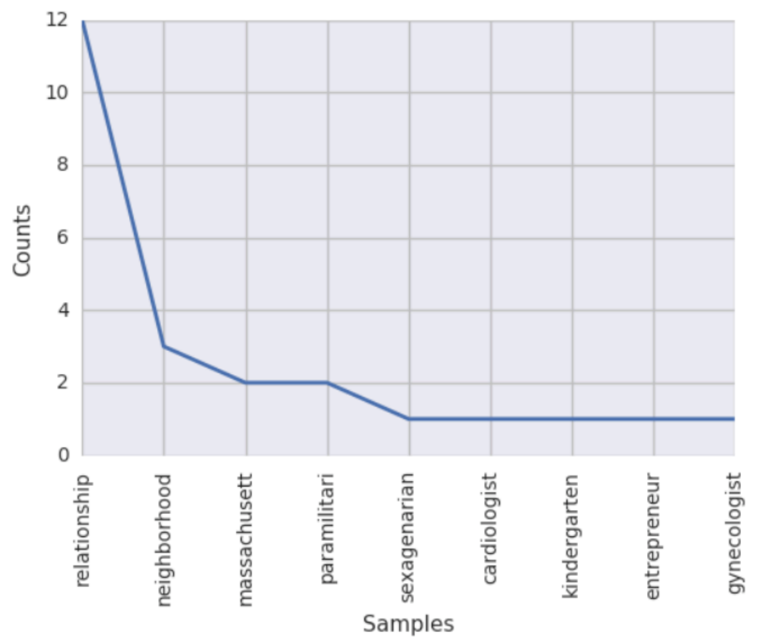
Figure 3. Frequency distribution of words with length two.



In Figure 3., we can see the frequency distribution of words with a length of two. It is interesting that the most common words across the survey seems to be “go” and “us.” In the case of the word “go”, it makes sense that it will have a high representation across all the survey since the process of migration is general act that requires sub actions in many dimensions. For example, actions of coordination of resources and connections, processes that need to happen to facilitate migration such as getting passports, contacting connections, and finding information. Finally, the last act, which is the act of moving or migrating itself. (Subero, 2012) Similarly, the prominence of plural personal pronouns such as “us.” According to The New Economics of Migration, the migratory decision is a decision taken by the family or the community and not a decision take by an individual, as other theories of migration such as

Wage Maximization argues. Thus, the fact that the word "us" has a high frequency distribution in this corpus goes along with the notion of communal decision of The New Economics of Migration.

Figure 4. Frequency distribution of words with length twelve.



The frequency distribution of words with length twelve is also interesting. In Figure 2., we can see that the word with the highest frequency is "relationship" and "neighborhood" This strengthens the idea exposed above about the importance of social relations in the process of migration. I remember from the process of reading and translating interviews' transcript the centrality of family in the process of migration, this notion is corroborated by the frequency of words such as "us," "relationship," and "neighborhood." Other words with prominence according to the distribution and of different lengths are “get” (frequency 72), “one” (frequency 102), and “live” (frequency 108).

Frequency Distributions and Parts of Speech (POS)

Another critical feature that I explored was the role that words play in the sentences in this corpus. Because of the found prominence that verbs such as “go” and nouns such as “us,” “relationship,” and “neighborhood,” I wanted to determine what is the importance of these words in the sentences in this corpus. Specifically, I wanted the frequency of each part of speech by word. I used the *nltk.pos_tag* class to tag and classify part of speech or POS across the corpus. The tags used correspond to the Brown Corpus tagset. This process resulted in a new column with the parts of speech. Then, I constructed a conditional probability distribution with a total of 2424 conditions using the function ConditionalProbDist from the ConditionalFreqDist. The model used for generating the model for the probability distribution was ELEProbDist.

First, I checked the most common adjectives and superlatives but the results were uninteresting. Then, I considered the five most common nouns and verbs. The most common noun is work with a probability of 0.01923, followed by year, because, state, and Venezuela. I am not sure why the command *POStoWord['NN']* identified the conjunction because as a noun. Beside that, the more common nouns seems to make sense but are uninteresting. More interesting are the most common verbs found by *POStoWord*. As mentioned before, because migration is in essence an act and the prominence in the frequency distribution of words such as “go,” I wanted to see other important verb in the corpus. The six most important are go (66), get (41), take (22), happen (14), see (13), and feel (10).

Figure 5. More common verbs.

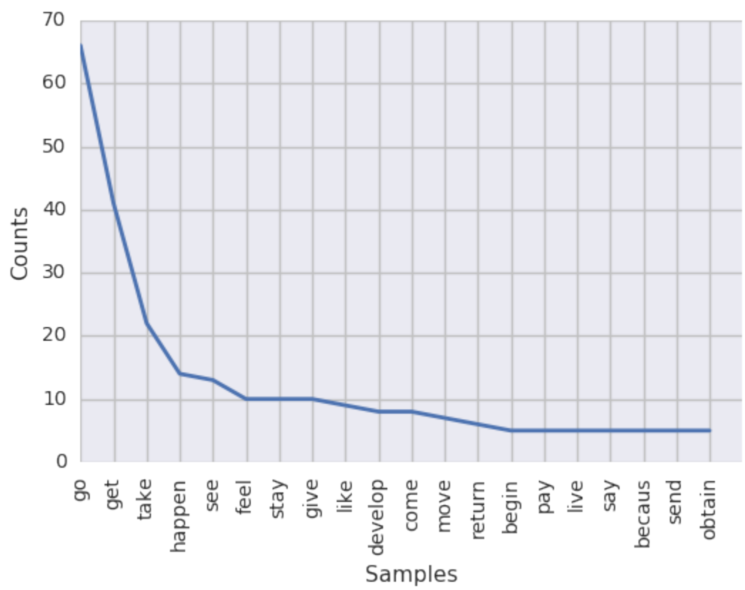
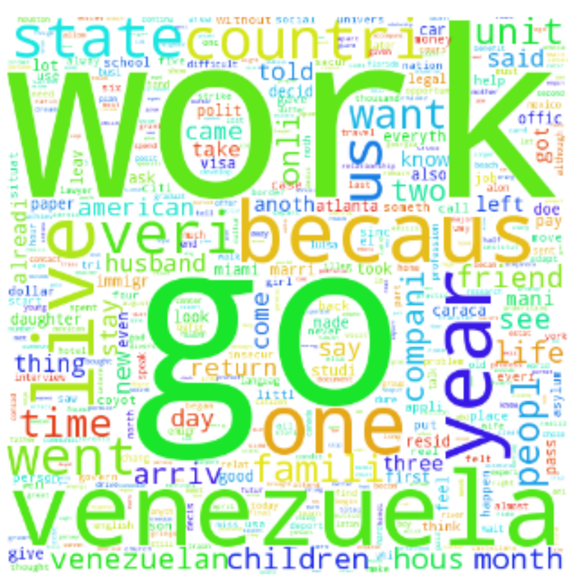


Figure 6. Nouns and verbs word cloud.



To facilitate the visualization and inference about the most important nouns and verbs, I plotted the twenty most common verbs and I created a Word Cloud with the most important nouns and verbs. (See Figure 5 and 6) The verb *go* and *take* make sense because the reasons exposed above. The importance in migration of acting, moving, and taking actions. Furthermore, contextual knowledge can add a possible explanation to the prominence of verbs such as get, take, and happened. In the context of migration obtaining is a key part of the process, in the case of Venezuelan migration. There are many resources that migrants need to obtain. For example, the visas, permits, licenses, resources, etc.

Bigrams, trigrams, and n-grams.

I continued the analysis of this corpus, finding the most common bigrams, trigrams, and n-grams. For this task, I use the following classes from nltk: nltk.collocations.BigramCollocationFinder, and andnltk.collocations.TigramCollocationFinder. Initially, I started looking at the raw counts. The more common bigrams are: united - state, new - York, and year – old and as it can be noticed most of them are uninteresting. However, the presence of the bigrams such as return – Venezuela, would – like, go - back, go – Venezuela, and one – day seems interesting. The reason is that in the literature exploring Venezuelans' subjective experiences the idealization of a return to the country of origin is a recurring topic. In other words, Venezuelans as a migratory group are characterized as having intentions to go back to their country, if the political and economic situation allows. (Alvarez, 2012; Subero, 2012). Similarly, it is interesting that the bigrams permanent - resident, and political – asylum are present among the most common bigrams. This corroborates the centrality of migratory strategies present in the literature, asking for permanence residency or political asylum.

To ensure that the appearance of these bigrams was not due to change, I computed several statistics for bigrams, trigrams, and n-grams: student_t, chi_sq, likelihood_ratio, and pmi. The statistics that produced the results that make the more sense were student_t, and, likelihood_ratio. In Figure 7. we can see the most common bigrams using two different measurements tudent_t, and likelihood_ratio. We can see that return to Venezuela and permanent residence continue being among the most common bigrams.

Figure 7. Nouns and verbs word cloud.

tudent_t,	likelihood_ratio.
united, state - 7.6960294173454	unit, state, - 670.854581426250
new, york - 3.7303597432852	real, estate, - 182.780999570861
year, old - 3.6990840275269	new, york, - 168.589822937107
real, estate - 3.6002550519485	permanent, resident - 90.5119009389434
return, Venezuela - 3.324310669	social, security - 80.5786256108842

Then, I proceeded calculating the trigrams and n-grams. Since the measurement student_t was the one that produced the results that make more sense according to the line of analyses of this exercise, and disregarded the other measurements. There are several trigrams that go alone with the knowledge about subjective experiences of migrants. Social-security-number (student_t score 2.2360566601994223) is one of the most prominent trigrams. This makes sense since one of the first and more important aspects of moving to the USA is the ability to work legally and have access to other benefits, which can only be obtained with a social security number. The trigrams citizen-united-state (student_t score 1.7318138856100618, at first glance, seems to contradict the notions expressed above about Venezuelans wanting to go back to their country. However, right after the appearance of the trigrams go-back-venezuela (student_t score 1.7303667007687362) indicates that, it is possible, as common sense suggests, that migrant will have a period of mental contradiction between the possibility of stabilizing in the receiving country or going back to their country of origin. Finally, an interesting trigram is cross-mexican-border (student_t score 1.7320445096433899. The reason is that among the literature on Venezuela migration, it is uncommon to find stories of illegal emigration. Clearly, the prominence of this trigram in this corpus indicates the peculiarity of this set of interviews or a recent changing trend in Venezuelans' strategies of migration.

Information Extraction

After, finding interesting words, bigrams, and trigrams related to possibly expressing Venezuelans desire to either go back or stay, I thought that using strategies from information extraction will allow me to use of computation and linguistic models to parse precise claims from the interviews.

Figure 8. First fifth-teen tagged POS in the migrants' corpus.

```
23    [[(Only, RB), (four, CD), (interviewees, NNS),...
22    [[(The, DT), (story, NN), (of, IN), (Luisa, NN...
21    [[(There, EX), (is, VBZ), (an, DT), (atmospher...
20    [[(Rafael, NNP), (Mirabal, NNP), (gave, VBD), ...
19    [[(Luis, NNP), (Alvaray, NNP), (,, ,), (43, CD...
18    [[(Adolfo, NN), (D, NN), ('Erizans, NNS), (,, ...
17    [[(Angela, NNP), (Maria, NNP), (Urdaneta, NNP)...
16    [[(JosŽ, NN), (L—pez, NN), (Padrino, NN), (is,...
15    [[(Carlos, NNP), (Lares, NNP), (studied, VBD),...
14    [[(Elia, NNP), (Mata, NNP), (poses, VBZ), (on,...
13    [[(Darcy, NNP), (Perez, NNP), (had, VBD), (bee...
12    [[(Jose, NNP), (Colina, NNP), (Pulido, NNP), (...
11    [[(Mariana, NNP), (Torrealba, NNP), (,, ,), (4...
10    [[(When, WRB), (Eira, NNP), (Ramos, NNP), (agr...
```

For this task, I started by initializing all the necessary tools the taggers (NER tagger and POS Tagger) and the parser. Then, I POS tagged the corpus. Below there is a set of the first fifteen tagged POS in the migrants' corpus. Then, I proceeded to count nouns. The five most common and interesting are country (counts 72), time (57), family (56), life (54), company (47), and husband (46). Then continuing with the focus on action in the exploration, I checked the most common verbs: be (18), have (17), get (14), do (11), and change (9). Interestingly the verb go (5), which was central in the Corpus Linguistic analysis, do not have the same prominence as a tagged verb. When trying to see which words surrounded some of these prominent nouns, the most interesting was the adjective classifying the noun family. It is well established among the literature on Venezuelans migration that most migrants are from the middle section of the socioeconomic statuses. This is corroborated by the prominence of adjectives such as whole, middle-class, and low-middle-class. Similarly, the adjective *whole* highlights the centrality of the family in this emigrants' narratives; which is common among Latin Americans.

Named-Entity Recognition

Figure 9. First ten most common non-objects and organizations.

```
23    [[(Only, O), (four, O), (interviewees, O), (in...
22    [[(The, O), (story, O), (of, O), (Luisa, PERSO...
21    [[(There, O), (is, O), (an, O), (atmosphere, O...
20    [[(Rafael, PERSON), (Mirabal, PERSON), (gave, ...
19    [[(Luis, PERSON), (Alvaray, PERSON), (,, O), (...
18    [[(Adolfo, PERSON), (D, O), ('Erizans, O), (,,...
17    [[(Angela, PERSON), (Maria, PERSON), (Urdaneta...
16    [[(JosŽ, O), (L—pez, O), (Padrino, O), (is, O)...
15    [[(Carlos, PERSON), (Lares, PERSON), (studied,...
14    [[(Elia, PERSON), (Mata, PERSON), (poses, O), ...
13    [[(Darcy, PERSON), (Perez, PERSON), (had, O), ...
```

Then I proceeded with a classification task using Named Entity Recognition (NER) to identify object and entities. I started by running NER in the entire corpus. First, I checked common entities which were uninteresting. Then, I focused on the list of the most common non-objects and organizations.

Interestingly, the most common non-objects seem to cluster the countries of origin, in this case Venezuela (counts 138), and to which they are moving to united states (60), Miami (30), and Atlanta (30). Similarly, among the most common organizations are *university (8), Harvard (5), and Intevep (4).* These results are interesting. It seems that among the migrants in this corpus, institutions of higher education such as Harvard are common. This goes along with the literature of Venezuelan migrants, it seems many come to the USA to study. The organization Itevep is an interesting finding. I checked, and it use to be an organization part of PDVSA which is state owned oil company. In 2002, more than ten thousand workers were fired because political reasons. Many of them emigrated to the USA to cities such as Miami, Atlanta, and Houston.

Parsing

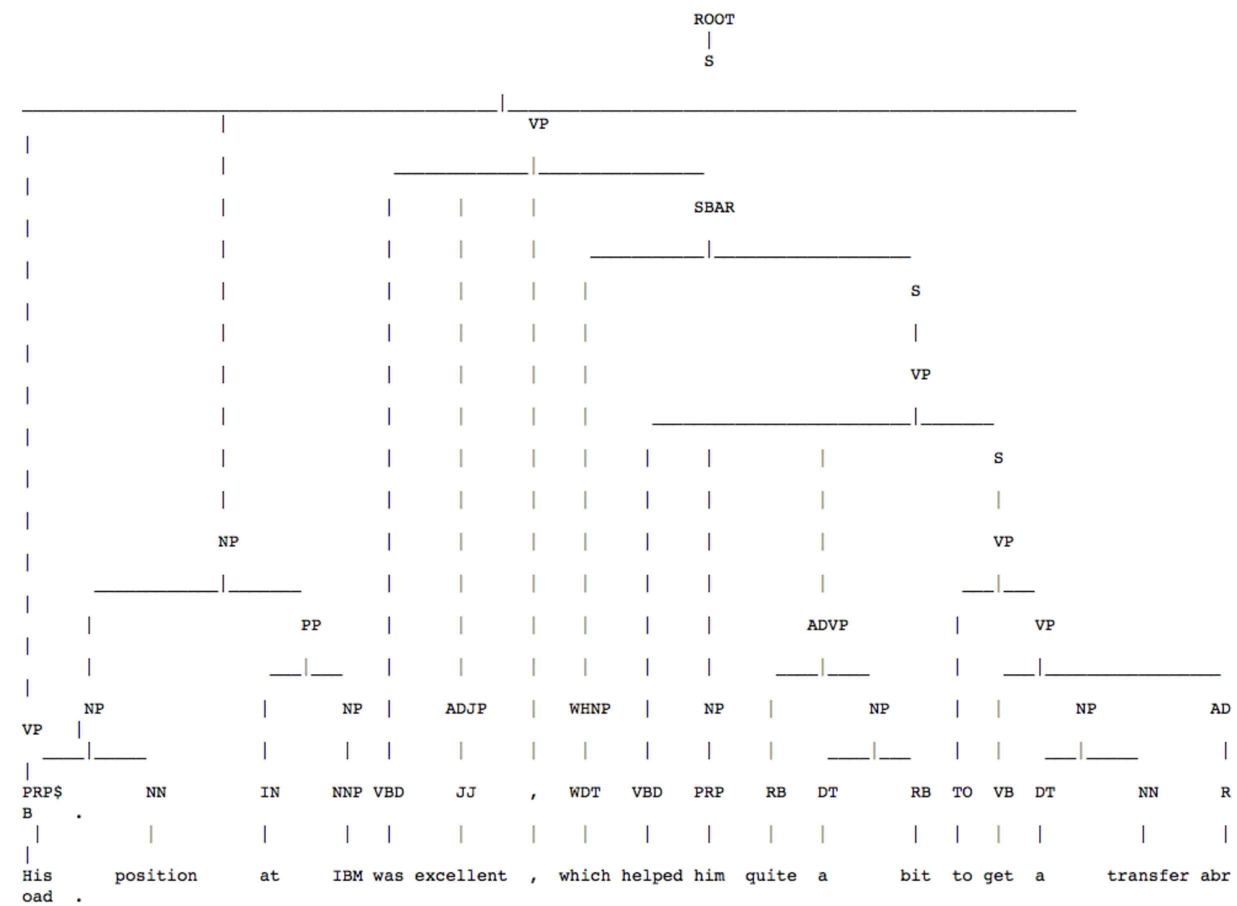


Figure 9. First ten most common non-objects and organizations.

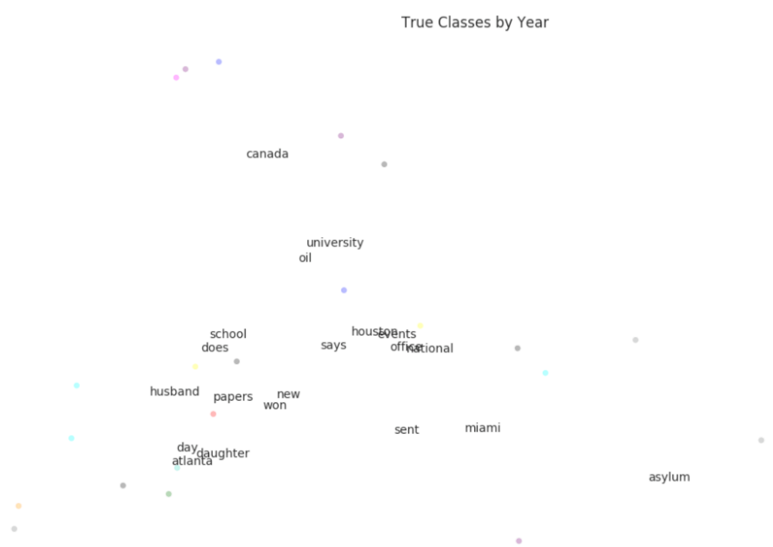
Then, to explore the relationship between different parts of speech in this corpus, I utilized the Stanford Parser with the corpus. I tried several levels of parsing: 3, 5, and 7. All this produced at least one sentence or dependency. These levels revealed deeply nested threes. For example, [Tree('ROOT', [Tree('S', [Tree('NP', [Tree('DT', ['This']). These deep nested structures revealed a high level of complexity of sentences and ideas. This level of complexity is characteristic of biographical accounts.

In the tree presented we can see the expression of one individual utilizing the opportunities in social connections through his employer to emigrate. This is quite interesting since one of the most common ways to emigrate for Venezuelans is through employment; to the extent that this Venezuelan migration has been categorized as labor migration.

Topic modeling

I wanted to see if I could model some topics using techniques from the week three. However, the models did not produce interesting results for two main reasons. First, the sample size is too small for predicting good topics using models such LDA or PCA. Secondly, even when these interviews are from a homogeneous group such as Venezuelans, the fact that their time of migration varies across twenty years could mean there is not a sufficiently unified voice across documents or interviews.

Figure 10. True Classes grouped by the variable year.



I also wanted to see how well this documents clustered. Thus, I conducted hierarchical clustering using Wald's method and Cosine Similarity. In figures 11 and 12, we can see that documents in general groups in five to six distinctive groups.

Figure 11. Hierarquical clustering of interviews Wald's method.

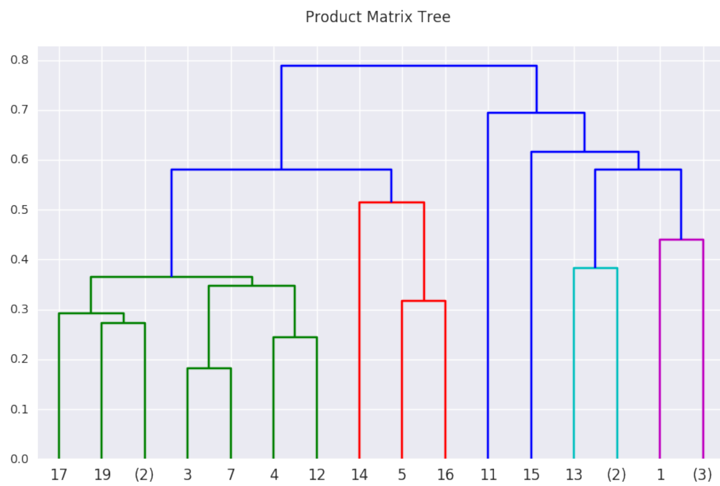
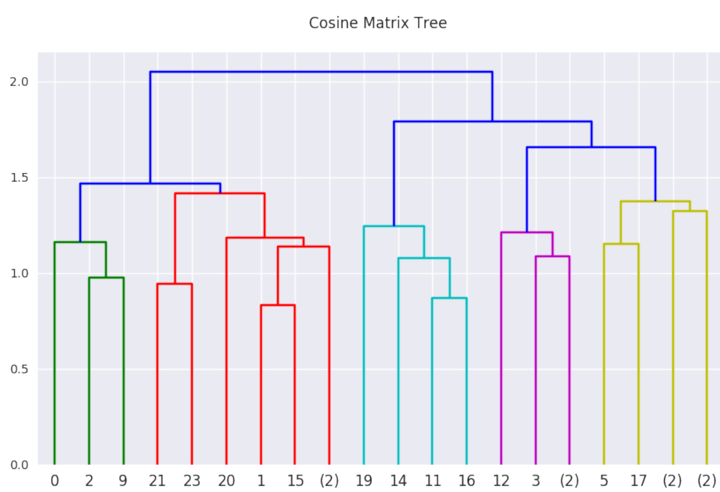


Figure 12. Hierarquical clustering of interviews Cosine Similarity.

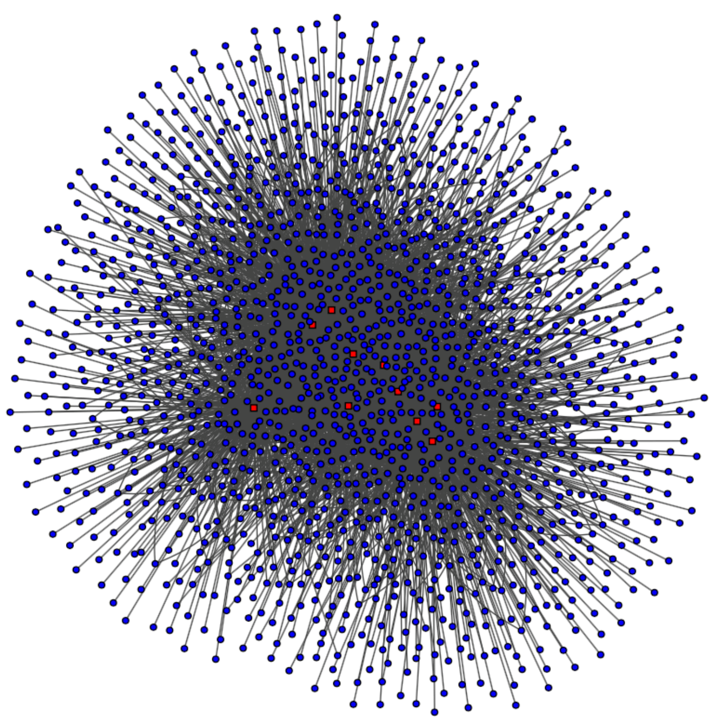


Semantic Networks.

I tried to use word embeddings techniques but I did not find any relevance in the result for understanding this corpus, so I proceeded with sematic networks to see If I can better visualize the relationships between words-to-words and documents. I stated by checking the word concurrence in the corpus, but I got an uninteresting clustering of terms.

So, I proceeded to build a two mode networks two-mode network. In Figure 10, we can notice that most words correspond to a larger cluster of documents (button), while there are two few subset of documents.

Figure 10. A two-mode network document-words.



Finally, I calculated some measures of centrality to see if which terms mediated the connections between words according to two measurements. Betweenness Centrality and Betweenness Centrality. These two measurement were the more interesting between all the measurements of centrality. The other measurements produced tightly clustered graphs. The words with more betweenness centrality are Miami - 1.0768283622852464, Asylum - 0.7103514496467788, and Husband - 0.6038394923002777. Similarly, the words with the highest closeness centrality are Miami – 110, Asylum – 94 and, Husband – 59. This highlights mainly two aspects of Venezuelan migration. First, the centrality of Miami as a migratory corridor; a place where migrants move temporarily as a platform for more stable locations. Also, the prominence of Asylum in this network of works correlates with the increasing number of petitions for political asylum by Venezuelan migrants. Finally, words such as "husband" talk about the importance of relationships among this groups of migrants.

Figure 11. Between-ness centrality of words.

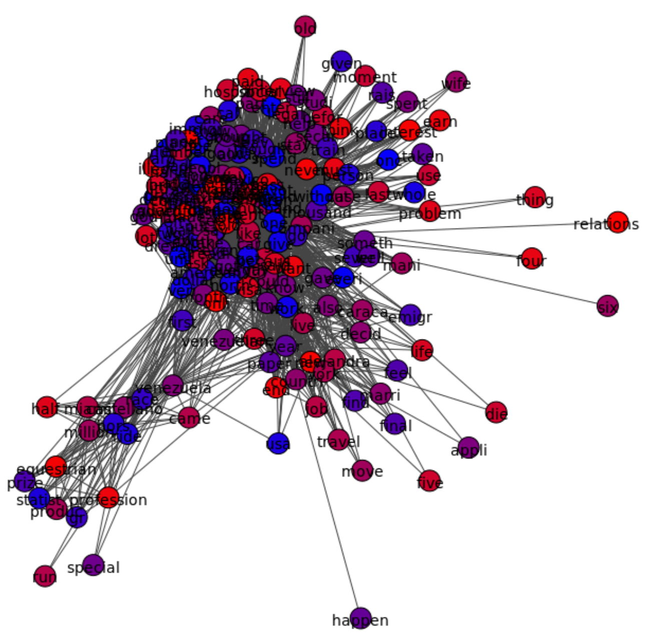
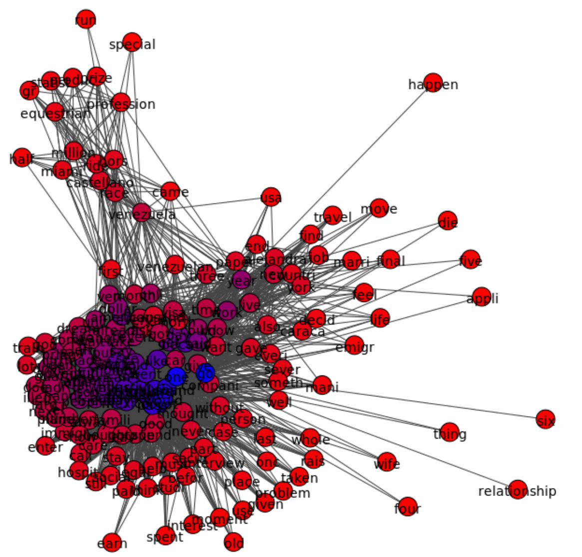


Figure 12. Closes-ness centrality of words.



Conclusions.

This exploratory exercise resulted in two main findings across all the tools used from five week: intro, Corpus Linguistics, Topic Modeling, Information Extraction, and Semantic Networks. The two main findings are the centrality of the cognitive struggle of Venezuelan migrants between staying in the USA or Canada, or returning to their country. This is interesting since these groups of migrants reflect their intentions to go back to their countries, if the political and economic situations allow.

The second, finding is related to the importance of relationships among this subset of Venezuelan migrants. The importance of this aspect was reflected in the corpus linguistic analysis with the prominence of terms such as “relationship,” “us,” “neighborhood,” and “husband.” The centrality of relationships also appeared in the Network Semantics analysis section. The importance of these two aspects of Venezuelan migration correspond to what authors have found: Venezuelans migrants are highly ethnocentric and want to go back to their country of origin if possible, and family places an important role in their process of migration. (Alvarez, 2012; Subero, 2012)

References

Alvarez, E. O. (2012). *La migracion Venezolana durante la rebolucion.* (M. Phelan, Ed.). Caracas.

Fernández, C. M. (2006). Nuevas direcciones para estudios sobre familia y migraciones internacionales. *Aldea Mundo*.

Fitzgerald, J., Leblang, D., & Teets, J. C. (2014). Defying the law of gravity: The political economy of international migration. *World Politics*. http://doi.org/10.1017/S0043887114000112

Martin, P. L. (2015). Managing International Labor Migration in the 21st Century. *South-Eastern Europe Journal of Economics*, *1*(1).

Mayda, A. M. (2010). International migration: a panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, *23*(4), 1249–1274. http://doi.org/10.1007/s00148-009-0251-x

Subero, C. (2012). La alegria triste de emmigrar.