Morena or PRD? An analysis from the left-wing Mexican parties

Text as Data Term Project[*]

Word count: 2998

Github-Repo: https://github.com/edufierro/TADproject

Eduardo Fierro Farah (eff254) and Akash Kadel (ak6201)

May 11, 2017

Abstract

In this work, we analyze the policy implication brought by the left wing separation of the main left-wing political party. Using a corpus of 4,251 bills introduced between September 1, 2015 to April 11, 2017 and a Latent Dirichlet Allocation model, we found that topics usually identified with the left in Mexico were split between these two parties. While MORENA is more associated with Human Rights and Indigenous people, PRD appears to be pushing for an agenda on women, health and People with Disabilities and Youth. We also find weak evidence that the Congressional Lingo used by the two parties is different.

[*] Prof. Arthur Spirling; Teaching Assistant: Mr Patrick Chester

1 Introduction

Mexico is currently undergoing one of the most important changes in it's political party system on it's short democratic history. The *Movimiento de Regeneración Nacional* (Movement for a National Regeneration, or MORENA), founded by Andrés Manuel López Obrador and recognized as a Political Party in 2014 (Diario Oficial de la Federación 2014), came to break an already broken left-wing on the Mexican party system. This breakup came when the two-times presidential candidate, AMLO, broke with his former political party, the *Partido de la Revolución Democrática* (Democratic Revolution Party, or PRD) just after the election, when the PRD decided to recognize the results and their defeat (Excelsior 2012).

Since then, steadily, MORENA has been gaining ground on the Mexican electorate[1]. From January 2016 to February 2017, for example, MORENA went from 4th to the second most prefered candidate towards the 2018 presidential election, bearly behind the*Partido Acción Nacional* (National Action Party, or PAN), Mexican right-wing conservative party (Mitofsky 2017).

It's under this context that one of the most important questions regarding the breakup between PRD and AMLO arises: Was the separation due to differences on policy preferences? Or was it only due to for AMLO to pursue the presidential candidacy toward 2018 for a third time?

To asses this, we have downloaded all the proposed bills in both, the *Cámara de Diputados* and the *Cámara de Senadores* for the current Congress, and performed a topic model analysis to them to test whether there are significant differences in the topics proposed by each party.

2 Literature Review

Lately, a significant amount of analysis has been recorded in the Western countries to predict which party will win the elections based on the topics discussed through the bills. People have started to use data science techniques to scrutinize the reports which can help them perform political analysis through the parliament

[1] Other than MORENA and PRD, there are two other minor left-wing political parties in Mexico: *PT* and *Movimiento Ciudadano*

bills. Analysis has been made through classification models and topic modeling to draw how legislative text is influencing decision taken by public.

Some of the most relevant work comes with the motivation to predict vote and ideological positions of law makers. For example, Gu et.al propose a "topic-factorized ideal point estimation model for a legislative voting network in a unified framework" (Gu et al. 2014). Sean M. Gerrish and David M. Blei use generative probabilistic models to also associate voting patterns and text from bills (Gerrish and Blei 2011). Justin Grimmer, using 24,000 press releases from senators in United States, develops a Bayesian Model with Dirilecht priors to estimate the topics on them, concluding that the model performs robustly by evaluating with stories from local news (Grimmer 2010).

3 Data

3.1 Data Sources

All the data used in this project comes from the *Sistema de Información Legislativa* (Legislative Information System), controlled by the *Secretaría de Gobernación* (Secretariat of the Interior). This system monitors and registers all the activity from the Congress (both, the Senate and the Deputies).[2].

The data was initially scraped on April 11, 2017, and it includes information only about the current *Legislatura* (Legislature), the *LXIII Legislatura.* This means that it covers the period ranging from September 1, 2015 to April 11, 2017. Although data on other Legislatures is available on the same system, because of potential computational restrictions we have decided to work on the first draft of this project with only this current Legislature. For the same reasons, the data is also limited to the *Iniciativas* (Initiatives or proposed Bills), leaving out other congressional documents out of the sample such as Points of order, Points of Agreement, among others.

[2] In Mexico, the congress is composed of two chambers: the *Cámara de Diputados* and the *Cámara de Senadores.* The former is composed of 500 members and the latter, representing the states of Mexico, is composed of 128 members, 4 per state.

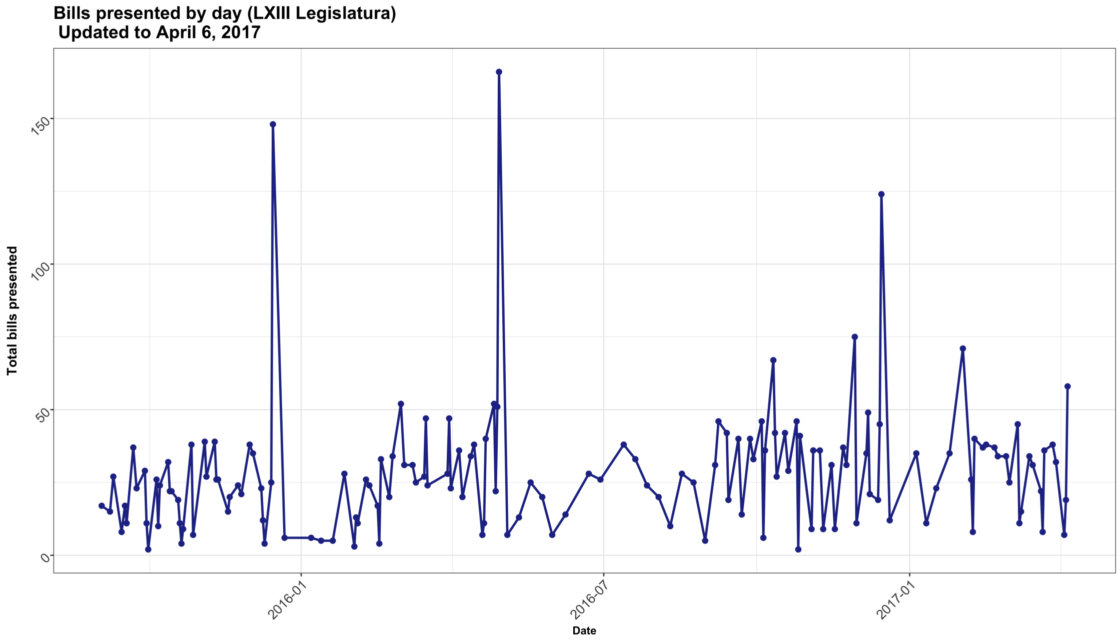

Figure 1: Bills proposed by day

3.2 Basic Data Exploration

From September 1, 2015 to April 11, 2017 there were 4,361 proposed bills in Mexican congress. As seen in Figure 1, there are 3 days where the number of bills proposed reach what seems to be an outlier number: December 12, 2015 (148); April 29, 2016 (166) and December 15, 2016 (124). These are the last days the congress is in session: for winter and summer break, the Mexican congress breaks and permanent congress is erected with members from the Deputies Chamber and from the Senate.

By far, the Political party that has proposed the most bills is the *Partido de la Revolución Institucional* (Party of the Institutional Revolution, or PRI). As seen in Figure 2, together with PAN and PRD, they have proposed more than 62% of all the bills that have been proposed in Congress. It's worth mentioning that there are three political parties as well that have more members in Congress: PRI has 205 Deputies and 52 seats in the Senate, about 42% and 40% respectively.

As a side note, it's important to note that, out of 4,361 bills in the system, we were unable to scrape 32 of them, mostly because there is no PDF file in the system for them. This are evenly spread across all parties, and are mostly from recent days: 19 of them from April 6, 2017 and 5 of them from the day before.

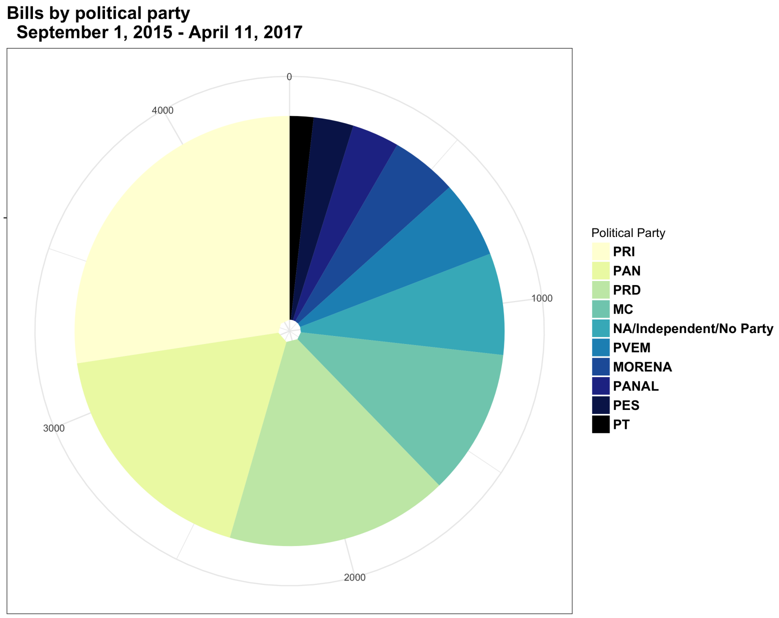

Figure 2: Bills by political party

This means, that probably they are going to update soon, but didn't made the cut on April 11, 2017. Also, the PDF-to-TXT converter used within R left a total of 78 bills with no text, although in this case it appears to be a bias towards the Independents (22 bills). This is, we dropped from now on 2.5% of the total corpus, analyzing a total of 4251 bills presented in Congress.

3.3 Pre-processing specification and strategy

Firstly, as opposed to the general convention, we are not planning to convert text to lower case. The reason behind this, mainly, is because every title of each bill is in uppercase, and it may have information about the name of the bill that will help us discriminate between parties. This formality also applies to cited legislation or international treaties on the main text of the bill.

Secondly, we also decided to remove Spanish stop-words included in the `quanteda` package. This comprises a list of 308 words, composed mainly of pronouns and prepositions. We also removed a dictionary composed of 70 words, edited from the Parliamentary Terms Dictionary of the system *SIL*. Lastly, we removed all numbers, arabic and roman, from 1 to 100. This is mainly because bills make reference to other laws they are aiming to change/modify, and usually they refer to chapters (Roman Numbers) or articles (arabic numbers). By themselves, on a bag-of-words, this is not informative at all. Punctuations were also removed.

Thirdly, no stemming was done. Although this particular point is debatable, we think that some words, central to topic modeling, can be confused if we stemmed them. Take, for example, the word *pobreza* (poverty), which is stemmed as `pobre` (poor). This last word can be widely used not only to describe poverty, but also to describe a "poor" performance of something or someone.

Finally, we have decided to use to regular bag-of-words (contrary to the TFIDF) with single words as features (contrary to bi-grams or more). The primary reason behind this is computational limitations.

No other pre-processing strategy was tried[3]. An extension to this project could explore how different are the results of the models with different pre-processing strategies.

3.4 Text data description: Similarity and Distance

As described above, all the parties have more than 70 bills presented during the LXIII Legislature. Merging all the bills per party, we were able to first approach the data by looking into the similarities and distances between their respective `dfm's`, while using MORENA (the party of interest) as our benchmark. As seen in Appendix A, the most similar party to Morena, using the Cosine Similarity criteria is in fact PAN (right wing), followed by PRD, the other party of interest. PRD also comes as the third more distant in the Manhattan Distance and Euclidean Distance measure (only after PRI and PAN, the center hegemonic party and the right wing party in Mexico respectively). This last result has more to do with the number of bills each party has presented rather than the similarity they share. A larger number of bills might be only a reflection of greater language used in it, which may be the reason why MORENA is more distant to the political parties with more bills introduced.

In our attempt to prove this point, we computed the Type-Token Ratio (TTR) for each political party using the same `dfm`, parallel to the number of bills presented by each. As seen in Figure 3, there is a strong relationship between TTR and the number of bills presented, indicating that the measures of distances

[3] Mainly because of computational restrains. The resulting DFM with which the model was fitted has

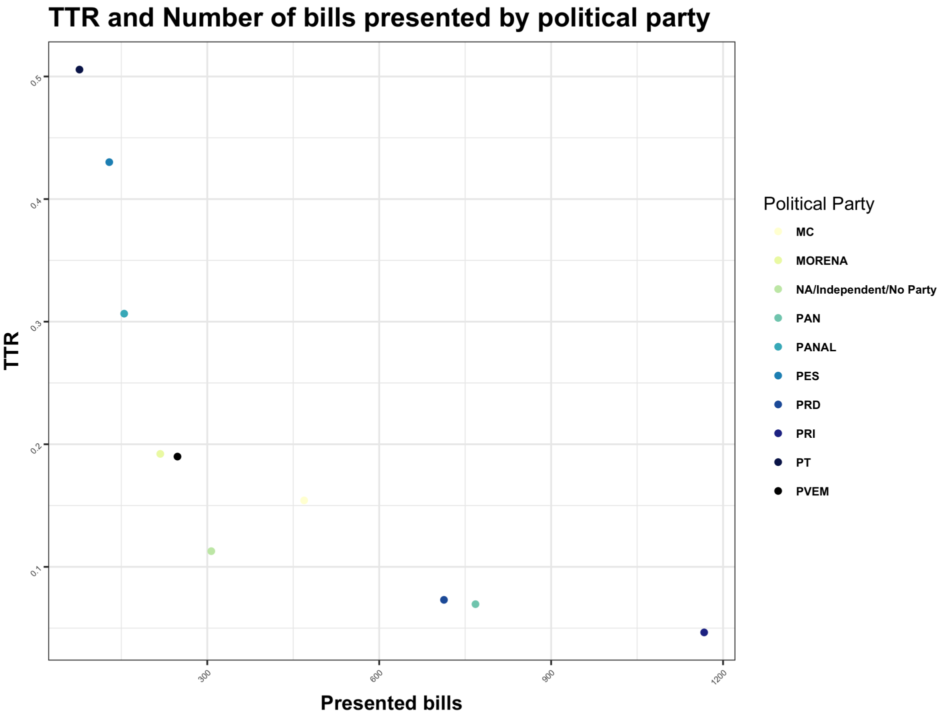

Figure 3: TTR and Number of bills presented

might be bias to a greater lexical diversity.

3.5 ZIPf's Law

Zipf's law, an empirically derived law, gives us the relation between types and tokens in the text. According to the law, the probability of encountering the i'th term in the corpus should be inversely proportional to the rank. It is also a good measure to check whether the corpus we are using is consistent or not. The corpus we are using clearly exhibits this relationship, as seen in the graph in Appendix B.

4 Method

Since our project involves identifying the difference in various topics distributions, we will be using Latent Dirichlet Allocation (LDA). The one interesting characteristic about LDA which makes it more appealing to use, is the fact that it assumes that each document consists of collection of words which belongs to a set of topics. It is very relevant to our project because any bill passed in the congress consists of topics related to current affairs, mostly present in the `Exposición de Motivos` section, where the lawmaker tries to motivate his or her proposal.

In LDA, each document is viewed as a mixture of various topics. It follows a probabilistic approach and assumes that every document-topic has a sparse Dirichlet prior. The sparse Dirichlet priors encode the intuition that documents cover only a small set of topics and that topics use only a small set of words frequently. The entire process is carried out in a generative way following these steps[4]:

1. Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$ where $\text{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

2. For every word in the document;

   (a) Draw a specific topic $z_{d,n} \sim \text{multi}(\theta_d)$ where $\text{multi}(\cdot)$ is a multinomial. Here $z_{d,n}$ is the topic assignment for the word in the nth position of the dth document.

   (b) Draw a word $w_{d,n} \sim \beta_{z_{d,n}}$. Here, $w_{d,n}$ is the word in the nth position of the dth document and it is being drawn from topic $\beta_{z_{d,n}}$.

5 Results

The Top 10 terms (untranslated) for each of the 30 topics run on the first model are presented in Appendix C. Almost all the topics can be clearly labeled, while others, as Topic-4, include contains majority of the words not included in the Stop words list and refer exclusively to Congressional lingo in Mexico, such as *rúbrica* or rubric, a formal word used on most documents[5]. All the topics were labeled to our best knowledge in Appendix D. Of the 30, 4 of them were unrecognizable, and 3 of them refer to the Congressional Lingo as discussed above. To answer the hypothesis, we next proceeded to analyze the topic distribution amongst the political parties.

[4] Steps copied from class slides. Author: Arthur Spriling

[5] For future works on the same corpus, the words included in this topic can be used as stop-words.

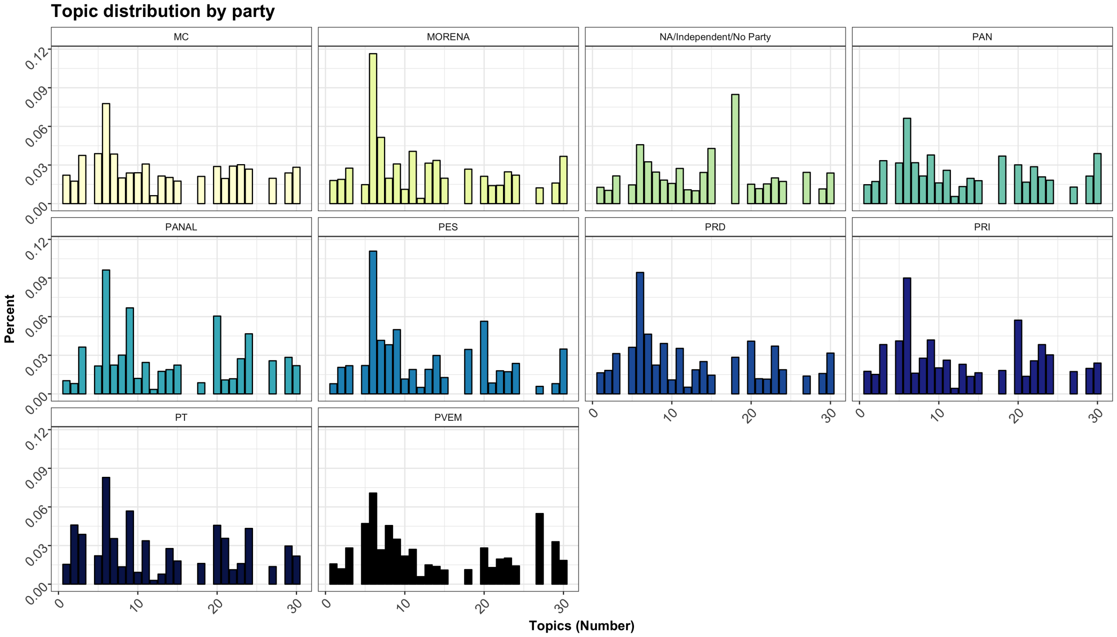

Figure 4: Topic distribution by Party, omitting "Unlabaled" and "Congressional Lingo"

As seen in Figure 4[6], there are some interesting differences regarding the new party, MORENA, in the topic distribution. First of all, MORENA seem to "score" very low regarding Women (Topic 20) in comparison to the other parties. While PRD's corpus speaks 4% about women, it drops to 2% with MORENA . But this is not the only topic where MORENA and PRD differ greatly.

As seen Figure 5, the two topics where they differ significantly, are topics 28 and 25. Interestingly, this topics were initially classified as "Congressional Lingo" which could signal a change between parties in the style they redact their bills[7]. Next to this, they differ greatly on topics 6 and 5, Health and Human Rights. The first one is more covered by PRD, while the second by MORENA. Other than this, they also differ on Indigenous People and and People with Disabilities/Youth, the first covered more by MORENA while the second more by PRD. It's worth noting that from the list of topics, these are perhaps the most associated with the left wing parties in Mexico.

[6] Surprisingly, the Congressional Lingo and unlabeled categories represent up to almost 20% on the topics distribution to some parties. This were omitted for clarity, but a complete Graph including all the topics is included in Appendix E.

[7] PRD was, from the beginning of the present *Legislatura*, an ally to the Federal Government. This could affect the presence of words as "Presidente", a word present on topic 29. In any case, this is an interesting matter but subject of another research paper.

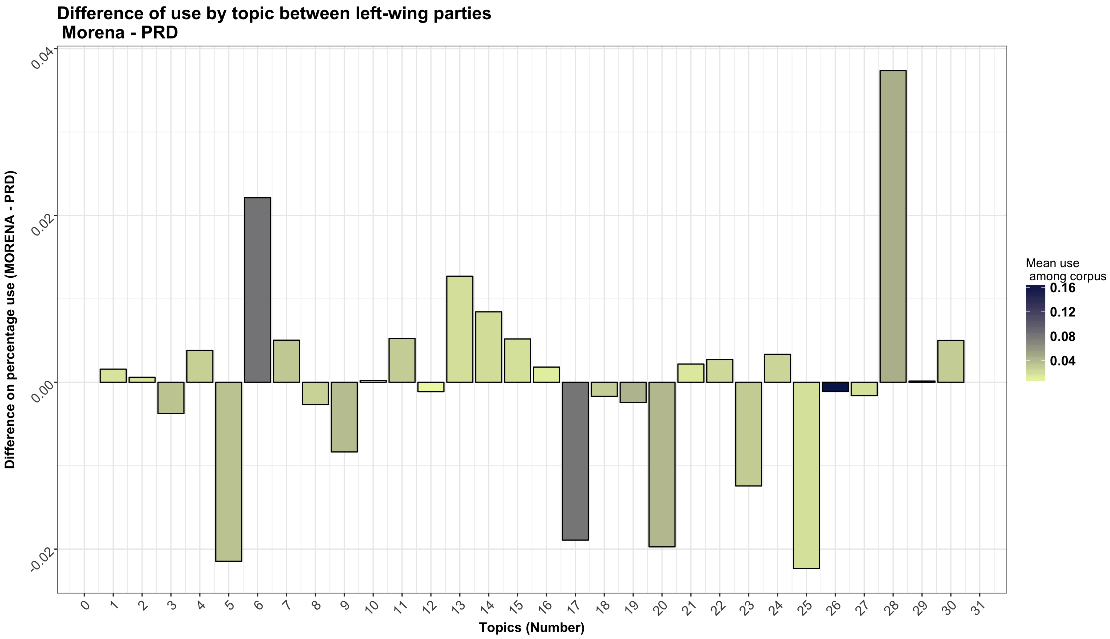

Figure 5: Topic distribution: MORENA vs PRD

5.1 Robustness Check: Number of topics

Using Nikita's Moor source code for in the package `ldatuning` (Moor n.d.), we adapted her functions to run them on already fitted models[8]. As seen in Appendix F, it seems that somewhere between k=30 and k=40 lies the correct number of topics maximizing Deveaud-2014. If we take into account the minimization criteria of Arun-2010 and CaoJuan-2009, the correct number of topics appears to be beyond well over 30.

This scores, although they partially justify the decision to pick k=30, it also arises a problem in the model: The stop word list initially used might be well underestimated.

5.2 Robustness Check: Topic stability

To asses whether our model was or not stable, we re-ran another `LDA()` model with k=30 and a different seed. The topics were matched between models using cosine similarity with the topic distribution over words. Later, the top-10 words for each topic for the two models were computed and the number of common words

[8] Because of the corpus size, each model takes between 2 to 3 hours to fit, so we decided to only fit the model, with the same seed, for 5 different k, save them as an ".RData" object and work over them. We were not able to replicate the `Griffiths2004` function

counted over the two models.

On average, the two models share 5.9 out of the top 10 terms per topic. Only two topics share all the 10 words (Social Programs and an Unlabeled topic), 5 topics share 9 words in common (Elections, Corruption and Indigenous People), and 2 topics share 8 words in common (a Congressional Lingo topic and Environment). On the other hand, one of the topics only has one word in common between the two models (Labour), and 5 topics only share 3 words in common (Education, Social Programs, Organ/Blood donation and 2 Unlabeled topics).

Only 25 topics out of the 30 from the original model were identified as similar to the 30 topics from the second model. The topics that didn't appear in the second model are Immigration, National Budget, Child Protection, Telecommunications and a Congressional Lingo topic.

6 Conclusions

In our work, we have examined how the political parties in Mexican Congress differ from each other, with a particular interest between MOREANA and PRD, the two main left-wing parties in Mexico. Since MORENA was created and the left-wing was more divided in the Mexican Congress, there has is no exploration on how this split reflected in policy proposal.

Running a Latent Dirichlet Allocation on text from all the bills presented in Mexican Congress for the *LXIII Legislatura* with a fix number of k=30 topics, we found evidence that the topics that are traditionally associated with the left-wing in Mexico were divided by these parties. For example, while MORENA is more associated with Human Rights and Indigenous people, PRD appears to be pushing for an agenda on women (presumably, women rights), health and people with disabilities and Youth.

We also found (weaker) evidence that the "Congressional Lingo" used by the two parties differ from each other. One hypothesis for this is that, while PRD has been a strong ally of the President, MORENA has greatly criticize it, which might reflect on the use of topics captured by the model that we were not able to label.

This work represents the first attempt to explore the policy differences between MORENA and PRD. It also opens a window for further analysis to be made. Future studies, for example, can use a bigger corpus, to explore with models such as burstiness how specific words and topics have been introduced in Congress since MORENA was created: Are Indigenous People rights a bursty topic since MORENA? Other works should also explore using other models as STM(), a topic that we were not able to test due to computational restrictions.

References

Diario Oficial de la Federación (2014). *Estatuto de MORENA*. URL: http://morena.si/wp-content/uploads/2014/12/Estatuto-de-MORENA-Publicado-DOF-5-nov-2014.pdf.

Excelsior (2012). *AMLO sale del PRD y apuesta por Morena*. URL: http://www.excelsior.com.mx/2012/09/10/nacional/858191.

Gerrish, Sean and David M. Blei (2011). "Predicting Legislative Roll Calls from Text". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Ed. by Lise Getoor and Tobias Scheffer. New York, NY, USA: ACM, pp. 489–496. URL: http://www.icml-2011.org/papers/333_icmlpaper.pdf.

Grimmer, Justin (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". In: *Political Analysis* 18.1, pp. 1–35. ISSN: 10471987, 14764989. URL: http://www.jstor.org/stable/25791991.

Gu, Yupeng et al. (2014). "Topic-factorized Ideal Point Estimation Model for Legislative Voting Network". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: ACM, pp. 183–192. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623700. URL: http://doi.acm.org/10.1145/2623330.2623700.

Mitofsky, Consutla (2017). *LAS PREFERENCIAS Y LOS ESCENARIOS PARA 2018*. URL: http://consulta.mx/index.php/estudios-e-investigaciones/elecciones-mexico/item/871-escenarios-rumbo-al-2018.

Moor, Nikita. "LDA tunning code". In: URL: https://github.com/nikita-moor/ldatuning.

Appendices

A Distances measures between Parties

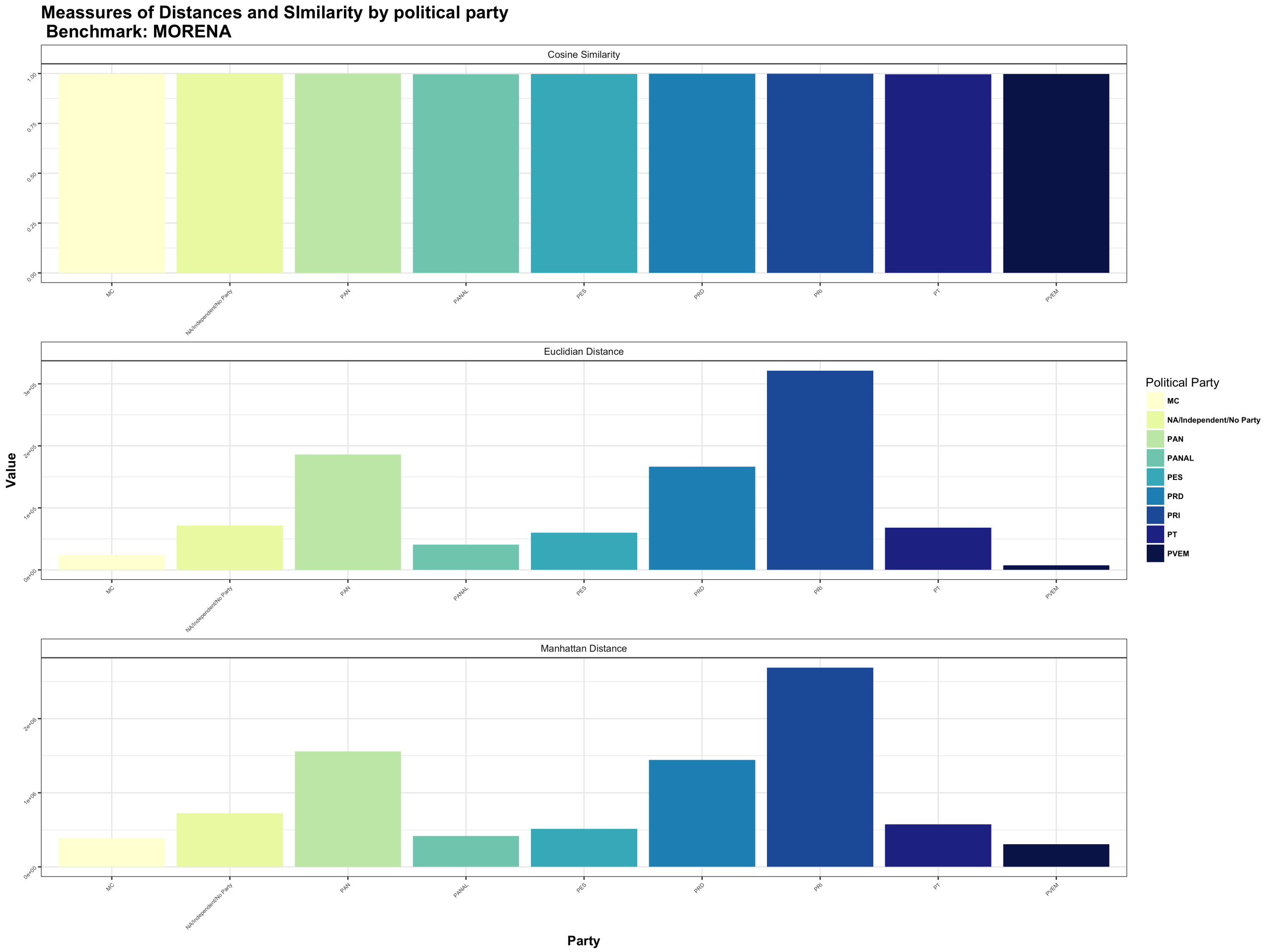

Figure 6: Measures of distance and Similarity Between Parties

B ZIP's Law - Figure and Top10 words in Corpus

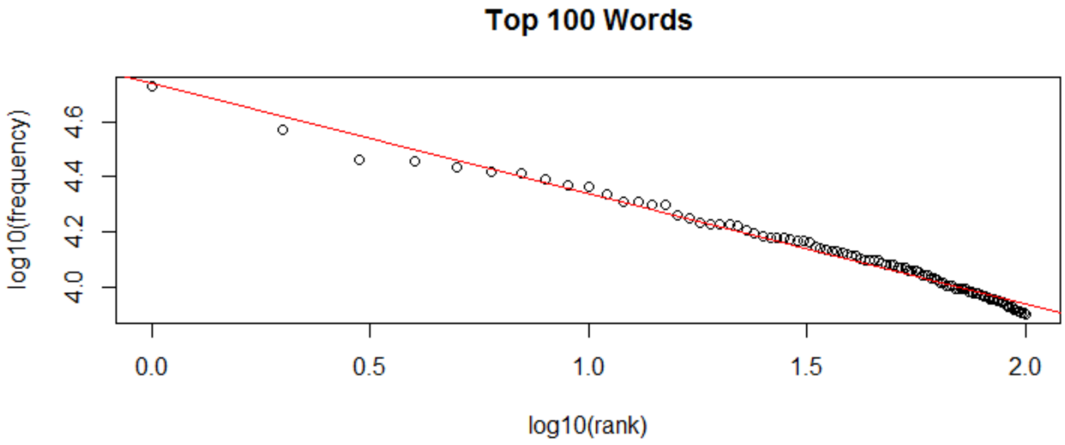

Figure 7: Zipf's Law

Top Features (10):

Word	Count
personas	24724
derechos	24021
ser	20016
derecho	19764
presente	19400
caso	18412
siguiente	17677
materia	17019
desarrollo	16977
país	16582

C Top 10 terms for LDA model with k = 30

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**
consumo	migrantes	trabajo	rúbrica	salud	derechos
alimentos	país	trabajadores	José	Salud	derecho
productos	territorio	laboral	Jesús	atención	humanos
uso	migratoria	Trabajo	María	enfermedades	ser
cannabis	personas	Social	Diputados	servicios	constitucional
producción	extranjeros	social	Congreso	cáncer	Derechos
Secretaría	Migración	derecho	presente	prevención	debe
sustancias	internacional	salario	Álvarez	tratamiento	protección
salud	mil	trabajador	Torres	médica	parte
venta	situación	Instituto	González	enfermedad	Política

Topic7	**Topic8**	**Topic9**	**Topic10**	**Topic11**	**Topic12**
recursos	niños	desarrollo	población	personas	Programa
Presupuesto	niñas	sector	desarrollo	delito	donación
gasto	adolescentes	Desarrollo	zonas	delitos	órganos
Fondo	menores	empresas	urbano	víctimas	Desarrollo
pesos	edad	social	ciudades	penal	Social
Secretaría	niño	económico	movilidad	derechos	PRESIDENCIA
entidades	animales	economía	transporte	persona	programas
ejercicio	protección	producción	Desarrollo	víctima	Salud
ingresos	derechos	país	turismo	prisión	sangre
público	animal	crecimiento	infraestructura	Penal	trasplante

Topic13	**Topic14**	**Topic15**	**Topic16**	**Topic17**	**Topic18**
indígenas	públicos	información	mil	materia	políticos
cultural	servidores	comunicación	medida	Secretaría	partidos
comunidades	corrupción	acceso	multa	entidades	electoral
pueblos	público	datos	actualización	programas	elección
cultura	Auditoría	uso	Unidad	acciones	electorales
culturales	Superior	medios	unidades	presente	representación
indígena	Pública	identidad	vigente	Ciudad	diputados
Cultura	recursos	internet	Medida	federativas	político
lenguas	administrativas	Información	Actualización	Sistema	política
patrimonio	entidades	telecomunicaciones	veces	deberán	candidatos

Topic19	**Topic20**	**Topic21**	**Topic22**	**Topic23**	**Topic24**
caso	personas	seguridad	servicio	mujeres	educación
deberá	discapacidad	Seguridad	transporte	género	Educación
podrá	población	pública	servicios	violencia	superior
procedimiento	social	armas	deporte	igualdad	escolar
autoridad	jóvenes	instituciones	vehículos	mujer	educativo
ser	mayores	Pública	carga	hombres	escuelas
días	acceso	fuerzas	usuarios	discriminación	educativa
persona	derechos	interior	pasajeros	derechos	educativos
plazo	condiciones	paz	seguridad	sexual	formación
proceso	edad	fuerza	Secretaría	Mujeres	educativas

Topic25	**Topic26**	**Topic27**	**Topic28**	**Topic29**	**Topic30**
Senado	país	ambiental	Congreso	agua	fiscal
SEN	ciento	ambiente	Comisión	uso	impuesto
ÚNICO	siguiente	naturales	Unión	aguas	ingresos
REFORMAN	http	especies	Diputados	residuos	Impuesto
MOTIVOS	mil	conservación	Poder	contaminación	pago
164	presente	forestales	Presidente	energía	fiscales
EXPOSICIÓN	ser	actividades	Ejecutivo	ambiente	crédito
siguiente	manera	recursos	Política	emisiones	contribuyentes
H	2016	medio	presidente	medio	ciento
Soberanía	años	Ambiente	ser	climático	párrafo

D LDA K=30 Topic Labels

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
Consumption (Food and Cannabis)	Immigration	Labour	Congressional Lingo	Health	Human Rights

Topic7	Topic8	Topic9	Topic10	Topic11	Topic12
National Budget	Child Protection	Economic Dev.	Crime	Social Programs	Organ/ blood donation

Topic13	Topic14	Topic15	Topic16	Topic17	Topic18
Indigenous People	Corruption	Telecommunications	UNLABALED	UNLABALED	Elections

Topic19	Topic20	Topic21	Topic22	Topic23	Topic24
UNLABALED	People with Disabilities/Youth	Peace/Security	Transportation	Women/Gender	Education

Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
Congressional Lingo	UNLABALED	Environment	Congressional Lingo	Water/Waste	Fiscal policy

E Complete Graph for Figure 4

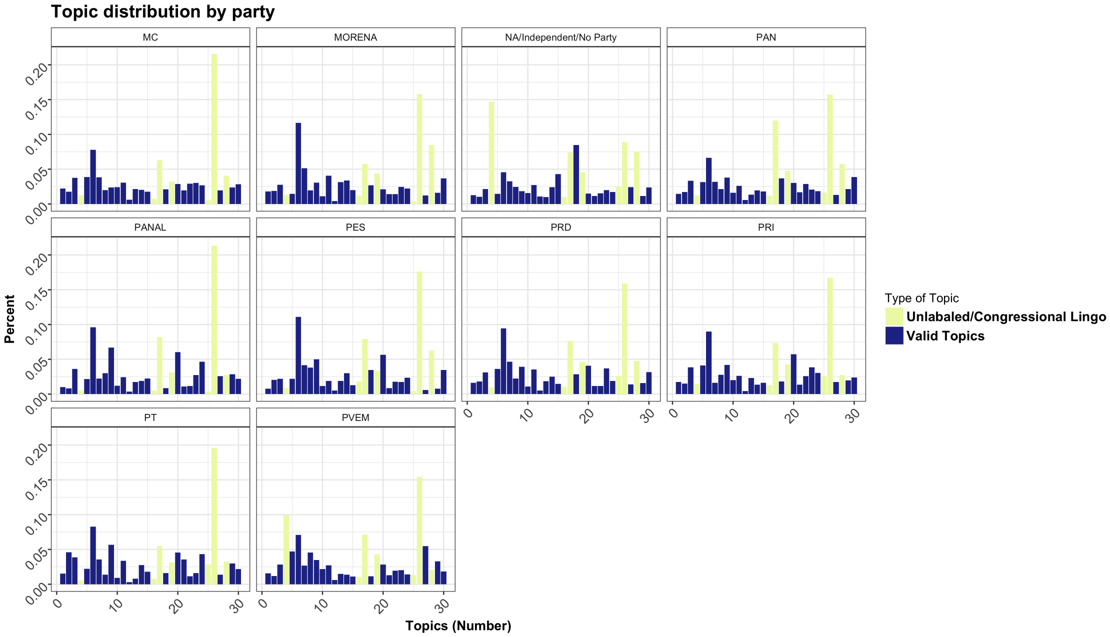

Figure 8: Topic distribution by Party

F Optimal number of topics: Scores

Number of topics(K)	CaoJuan2009	Arun2010	Deveaud2014
5	0.4095822	1212.8261	1.550213
10	0.3039988	1086.6989	1.783467
20	0.1577518	929.8224	2.168318
30	0.1047884	861.7938	2.275370
40	0.0716956	798.0841	2.345985