Diego Díaz H.
diegodiaz@udd.cl

4ta entrega Proyecto Diplomado
Data Science - Visualización de datos.
Profesor: Eduardo Graells

2018-03-31

1. Introducción y definición del problema

El crecimiento económico se explica por la acumulación de factores de producción como capital y trabajo, y mejoras en la tecnología disponible, lo que permite un uso más eficiente de los factores de producción (Romer, 1990). Sin embargo, en los últimos 20 años ha surgido un cuerpo importante de literatura que enfatiza la importancia de instituciones como protección de la propiedad privada, estado de derecho, certeza jurídica y restricciones al poder ejecutivo, en el crecimiento económico de largo plazo (Acemoglu & Robinson, 2012). En general, estas son instituciones asociadas a un tamaño de gobierno limitado, las que crean un marco de incentivos que permiten que los individuos exploten las oportunidades del mercado e innoven para mejorar sus condiciones de vida.

Por otro lado, para que existan estas reglas se requiere la presencia de una cultura favorable, y liderazgos que las promuevan (Mokyr, 2016). En este aspecto, la ideas juegan un papel importante. Algunos autores incluso han llegado a mostrar una relación entre tendencia política del gobierno con crecimiento económico (Bjørnskov, 2005).

Este trabajo persigue un doble objetivo. Por un lado busca estudiar el efecto de las ides en el crecimiento económico, en linea a lo que hace Bjørnskov (2005) pero longitudinalmente, es decir, a lo largo de la historia de Chile. Por otro lado, busca medir el efecto de la evolución de ciertos conceptos económico/culturales en el crecimiento económico.

2. Metodología

En este trabajo se utilizarán herramientas de análisis de texto para ambos objetivos, pero para esta sección de visualización se avanzó en la presentación de la evolución de tópicos a lo largo de la historia de Chile. Esto se determina por *topic modeling* usando un método estadístico generativo llamado *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003) en conjunto con la herramienta de visualización *ldavis*. LDA es un método de clasificación de texto, que asume que cada documento es una mezcla de un determinado número de categorías (o tópicos), y que la aparición de cada palabra se debe a una de las categorías a las que el documento pertenece.

Para avanzar en la visualización, la que en la entrega anterior consistía simplemente en la determinación de los tópicos y en graficar la evolución de algunos de estos a lo largo de la historia, lo que se puede hacer ya que se tiene una distribución de tópicos a lo largo de la historia de Chile desde que tenemos discursos presidenciales. Ahora se extiende incluyendo una app web que permite hacer una rápida exploración de los resultados obtenidos por LDA.

3. Datos necesarios y disponibilidad

Los datos a usar son los discursos presidenciales de estado de la nación de toda la historia de Chile, los que comenzaron en 1832. Tomando hasta el año 2015, se tiene un *corpus* de 183 documentos, de una cantidad de palabras muy variable (entre 5.000 y 30.000).

Acerca de la disponibilidad de datos, todos los discursos están a libre disposición en la página de la biblioteca del congreso nacional (bcn.cl). Estos están en formato *pdf*, por lo que conviene cambiar su formato a *txt* para facilitar el procesamiento en *python*. Esta etapa ya se hizo para los discursos desde 1926. Los más antiguos se encuentran en peor estado, lo que significa que se requiere una cierta cantidad de trabajo humano de transcripción.

3.1. Análisis exploratorio

En el *proof of concept* de la primera entrega se procesaron 26 discursos (1990-2015), y como análisis exploratorio, el texto fue separado por palabras (*tokenized*) para determinar el número de palabras, y ver la desviación estándar de la lista. Se calcularon algunos estadísticos del conteo de palabras de los 26 discursos mencionados, los que se muestran a continuación.

- Min: 9,714 en año 2001
- Max: 32,450 en año 1997
- Desviación Estándard: 5,324.2

También se hicieron nubes de palabras para los 26 discursos, para lo que fue conveniente usar los paquetes *WordCloud* y *nltk*. El script produce y guarda cada nube como un archivo de imagen con el año del discurso como nombre. En la figura 1 se muestra el resultado para el discurso de 1990, pronunciado por el presidente Patricio Aylwin. Es importante notar que los documentos tuvieron que ser procesados previamente para mostrar resultados interesantes, y esto consistió simplemente en remover las *stopwords* del *corpus*.

Figura 1: Nube de palabras discurso presidencial de Patricio Aylwin en 1990

En conjunto con la primera entrega se adjuntó un *jupyter notebook* llamado *Proof of concept* con un análisis exploratorio que produce los resultados mostrados más arriba y las nubes de palabras de todos los demás discursos en el intervalo 1990-2015.

4. Visualización de datos - Topic modeling, distancia vectorial y LDA-VIS

Después de hacer un análisis exploratorio de mayor complejidad en la sección del curso de Análisis de datos, en donde expandimos el conjunto de discursos, tomando todos los discursos de *State of the Union* en Chile en el periodo 1926-2015. El *jupyter notebook* adjunto, llamado *Díaz DS State of the Union* contiene muchas notas en *markdown* explicando lo que hago.

El *notebook* se adjunta en conjunto con los datos, que están separados en dos carpetas, *corpus*, que contiene los discursos en el periodo 1926-1989 y *1990-2015*, que contiene los discursos posteriores. Esta separación se hace debido a que los textos están en distinto *encoding*. También se adjunta una lista de *stopwords*, que corresponde a un archivo en formato *.txt* que contiene palabras que deben ser eliminadas de los discursos por no contener información relevante para la mayoría de los métodos, como artículos y preposiciones. Esto también ayuda a reducir la dimensionalidad del problema y mejorar la velocidad del análisis. También se incluye una lista con todos los presidentes del periodo analizado y un archivo *.csv* con datos de ingreso per cápita para Chile desde 1926.

Una parte fundamental de limpieza de datos es la creación de una *Document-Term Matrix (DTM)*, que contiene la frecuencia de aparición de cada palabra usada en cualquiera de los discursos, dado que tal palabra no pertenezca al conjunto de *stopwords*. Por lo tanto, la DTM principal es una matriz de 35096 x 90, ya que el vocabulario total es de 35 mil 96 palabras, y hay 90 discursos. También se creó una DTM secundaria, que tiene los discursos agregados por periodo presidencial, esta se llama *dtm-presidents* y permite hacer visualizaciones interesantes. Por ejemplo, podemos usar una medida de distancia vectorial para ver semejanzas respecto al uso de palabras entre discursos. La figura 2 muestra esta visualización, en la cual se ubican los discursos presidenciales entre ellos de acuerdo a una medida de distancia vectorial (similitud del coseno).

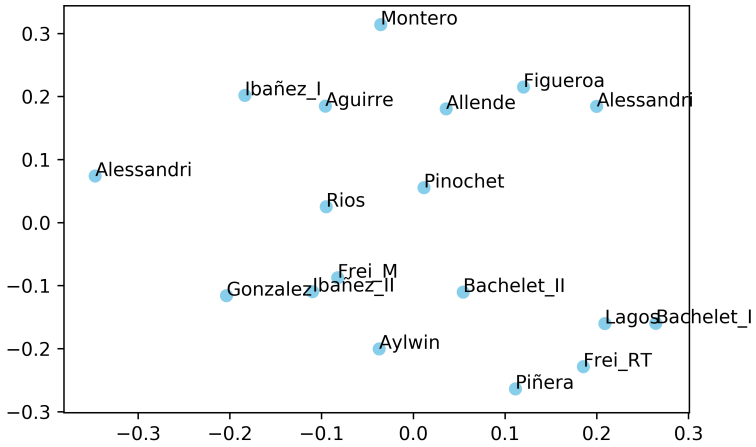

Figura 2: Distancia vectorial entre discursos por periodo presidencial. Nota: Es importante no considerar la posición respecto a un eje como una medida de tendencia política, lo único que nos dice el gráfico es la similitud entre presidentes según la frecuencia de las palabras usadas. Las que evidentemente pueden usarse en contextos muy distintos. Interesantemente, se puede ver un cluster de los presidentes desde la vuelta a la democracia en el sector sur este de la figura.

Se convierte la DTM a un *dataframe* de *pandas* para facilitar el análisis. Esto permite graficar de forma más directa la evolución de la frecuencia de ciertas palabras de interés político-social. Por ejemplo, en la figura 3 se grafica el uso de la palabra mujer en los discursos (para un análisis de otras palabras de interés ver el *jupyter notebook*), como se puede ver, parece aumentar la varianza desde alrededor de 1990, aunque no hay tendencia clara de largo plazo.

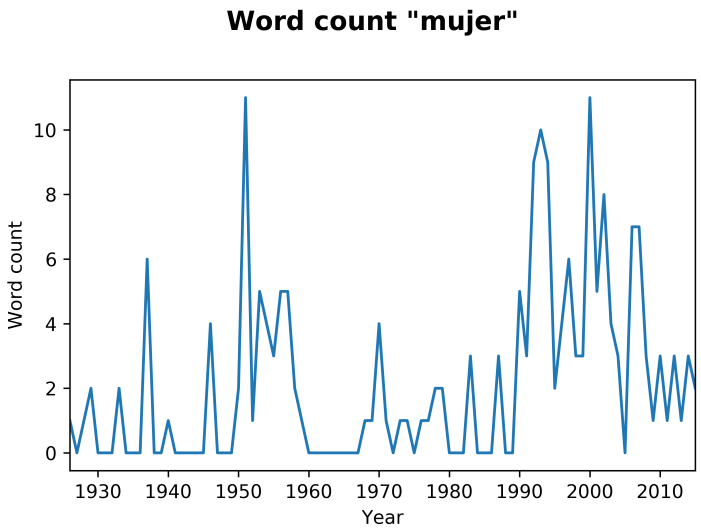

Figura 3: Frecuencia anual de uso de la palabra mujer en los discursos presidenciales.

A continuación hacemos un ejercicio de clusterización, el que entrega otra visualización acerca de la similitud

entre discursos. Esto consiste en un dendrograma para los discursos agregando por periodo presidencial (figura 4). En linea con el resultado de distancia vectorial, se puede observar un clúster con los últimos 4 presidentes.

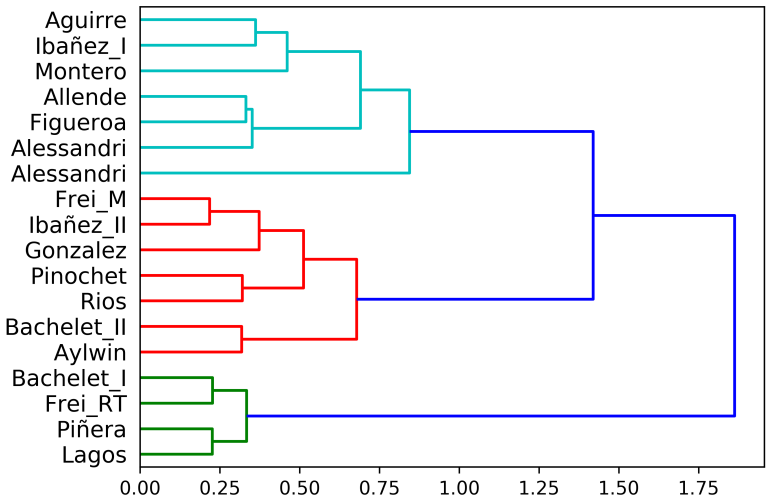

Figura 4: Dendrograma de discursos agregados por periodo presidencial.

El principal avance en la visualización de datos se hizo en el problema de *topic modeling*. el cual, como se explica en la metodología, se hace usando un método estadístico generativo llamado *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003). Este método recibe como input una DTM, o un corpus de textos para formar la DTM, y también el número de *topics* que se quiere encontrar. Los topics encontrados anteriormente fueron los siguientes (se muestran las 10 palabras más importantes para cada uno):

- Topic 0: pesos ley ano servicios trecientos produccion servicio numero cuatrocientos caja
- Topic 1: años educacion ley proyecto congreso sociedad estado plan paises publica
- Topic 2: año desarrollo social recursos estado sistema politica democracia futuro empresas
- Topic 3: nacional sector desarrollo programa sectores inversion participacion nivel coma proyectos
- Topic 4: gobierno politica republica accion naciones vida nacion social fuerzas forma
- Topic 5: gobierno progreso condiciones necesario atencion medios especial comercio actual especialmente
- Topic 6: produccion sistema politica gobierno dolares economico traves pueblo cobre escudos
- Topic 7: ano obras construccion desarrollo numero aumento doscientos labor plan forma
- Topic 8: ano calidad quiero familias sistema mundo programa salud oportunidades materia
- Topic 9: estado ley nacional fuerza administracion problemas materia ejecutivo medidas economia

El modelo calculado entrega la evolución de cada uno de estos topics, ya que cada discurso debe estar compuesto en su totalidad por ellos, es decir, la suma de topics en cada discurso debe sumar 1. Supongamos queremos estudiar si la transición a la democracia se vió reflejada en el discurso. Podemos tomar el topic 2 por mencionar la palabra 'democracia', junto contras palabras como 'futuro', 'desarrollo', y 'empresas'. La figura 5 muestra la evolución de este topic en el tiempo:

Para hacer una visualización más descriptiva, se usó la herramienta de visualización LDAvis (Sievert & Shirley, 2014), una herramienta web interactiva que permite visualizar los tópicos estimados usando LDA a través de una combinación de R y D3. A pesar de que hay una libreria disponible para python, se tuvieron demasiados problemas para implementarla, por lo que decidí finalmente implementarla en R, para lo cual se hizo todo el proceso de limpieza de datos en R. Se adjunta código de replicación y archivos con la visualización web json, javascript, y html.

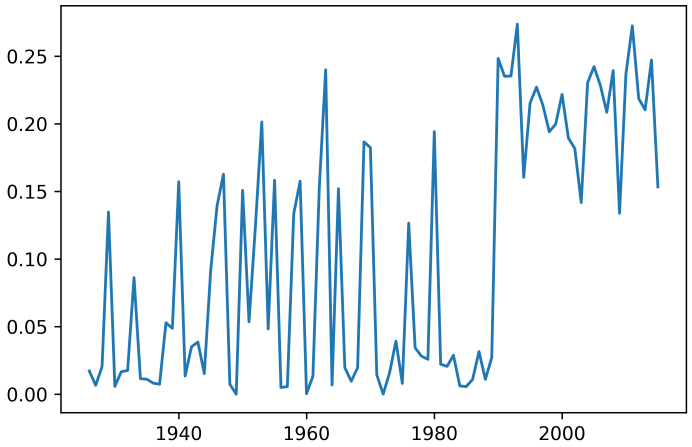

Figura 5: Evolución de topic 2 encontrado por LDA en 1926-2015.

La visualización permite una vista global de los tópicos y como difieren entre ellos, mientras al mismo tiempo permitir una inspección detallada de los términos con mayor asociación a un determinado tópico. En la figura 6 se agrega un screenshot de esta visualización, como es interactiva, permite una exploración muy rápida del modelamiento. Encontrando que el tópico 1, el cual es el más esparcido a lo largo de los discursos junto con el tópico 2 (cada uno ocupa alrededor del 10 % de la distribución marginal) está representado por palabras como Gobierno, año, país, proyecto, y vida, lo que está alejado de la mayoría de los tópicos encontrados (20 tópicos fueron modelados en la iteración entregada), esto puede saberse rápidamente debido a que la visualización hace un mapa de distancia entre tópicos usando un escalamiento multidimensional. De esta forma también podemos determinar que los tópicos entre 6-20 con excepción del 13 son similares entre si ya que forman un cluster en la sección superior del mapa.

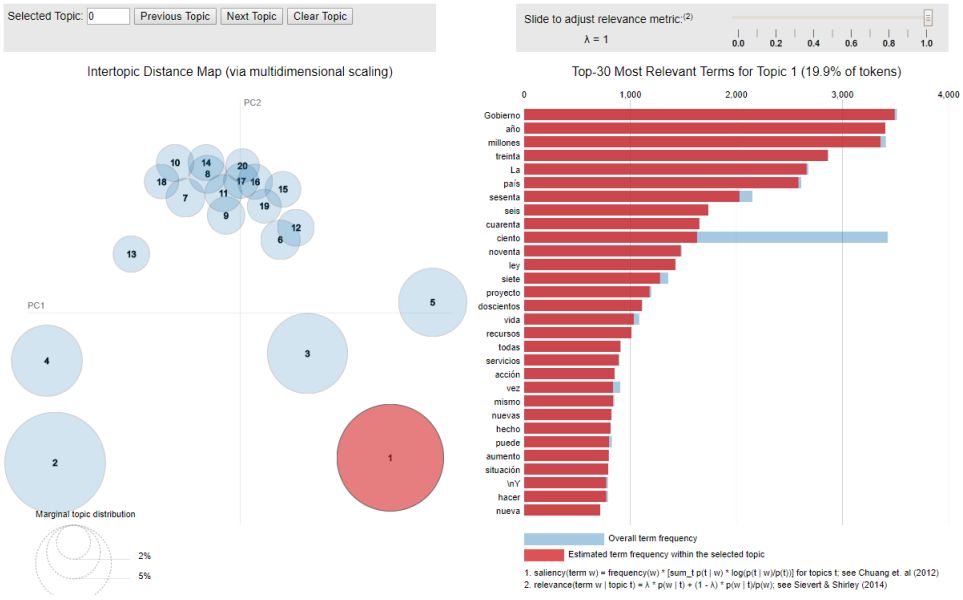

Figura 6: Screenshot de web app creada por ldavis

5. Bibliografía

Acemoglu, D. & J.A. Robinson. (2012). Why nations fail: The origins of power, prosperity, and poverty. Crown Business.

Bjørnskov, C. (2005). Does Political Ideology Affect Economic Growth? Public Choice, 123(1/2), 133-146. Retrieved from http://www.jstor.org/stable/30026794

Blei, David M., Ng, Andrew Y., Jordan, Michael I (2003). Lafferty, John, ed. ”Latent Dirichlet Allocation”. Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.

Mokyr, J. (2016). A Culture of Growth: The Origins of the Modern Economy. Princeton: Princeton University Press.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review, 97(2), 311-331.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Slapin JB, & Proksch S. (2008). A scaling model for estimating time-series party positions from texts. American Journal of Political Science 52(3):705–722.

Romer, Paul M. (1990). Endogenous Technological Change. Journal of Political Economy, 98(5): S71-S102.