Análisis de Discursos Presidenciales con LSA

Marcos Pietto *, **Sergio Romano** ** and **Noelia Rugna** ***

*CONICET-UNC,**CONICET-UBA, and ***UBA

Trabajo Práctico del Seminario Introducción a la Neurociencia Computacional

En este trabajo utilizamos "Latent semantic analysis" (LSA) para caracterizar discursos de distintos presidentes de latinoamérica a lo largo de la historia. En ellos buscamos conceptos que nos permitan distinguir elementos típicos de la construcción del "discurso populista" para evaluar la efectividad de LSA como herramienta para el análisis de discursos y para verificar la importancia del eje discursivo en el populismo.

lsa | discursos presidenciales | populismo

LSA es una técnica de procesamiento del lenguaje natural para analizar la relación entre un conjunto de documentos y las palabras que lo componen, mediante la creación de una matriz que representa un espacio semántico vectorial en el que los documentos y sus palabras están representados por medio de vectores que contienen sólo la *información principal* para la formación de conceptos [1, 2].

Este trabajo busca experimentar sobre la efectividad de esta técnica para el análisis de discursos presidenciales, con el fin de que pueda utilizarse para obtener métricas cuantitativas sobre distintos conceptos de las ciencias políticas tales como: hegemonía, liberalismo, populismo, etc. En particular este estudio pretende analizar el concepto de populismo.

El populismo es parte constitutiva de la política y de la democracia en América Latina. Durante muchos años, la literatura lo utilizó como un término peyorativo para designar a una vasta variedad de experiencias políticas mundiales, caracterizándolo por su vaguedad, irracionalidad y vacío ideológico [3–7]. Sin embargo, Laclau [8–10] introduce un cambio en esta visión planteando que el populismo no es una ideología de contenido específico, sino un método de construcción política que cohesiona una variedad de demandas y logra unificarlas en torno a una identidad popular común (un significante vacío que condensa las demandas) y a una frontera antagónica que divide al campo social. De aquí, es entonces que al análisis del discurso se vuelve central para entender y caracterizar al populismo.

Por tanto, proponemos una metodología de análisis de contenido a través de LSA sobre los discursos presidenciales dada su importancia en la construcción populista, y examinaremos métricas cuantitativas sobre dimensiones identificadas como propias de un discurso populista para evaluar la efectividad de la técnica de LSA para el ámbito de conocimiento de las ciencias póliticas.

Particularmente, nos centraremos en los discursos de apertura de congreso que los líderes de gobierno realizan cada año de su mandato ya que estos suelen ser transcriptos y además permiten un análisis comparativo en contextos políticos similares. Debido a que no sería apropiado comparar discursos de diferentes contextos como ser, un discurso de inauguración, de una expropiación, o de una cena con empresarios, pues se tendría un error aleatorio mucho mayor.

Método

Generación de LSA. Para generar la matriz LSA el primer paso es representar el corpus de 10.000 documentos extraidos al azar de la Wikipedia en Español en una matriz que contiene la cantidad de apariciones de cada palabra en cada artículo o documento. Previamente, a cada artículo se le quitó el código de sintáxis (markup) de la Wikipedia, una lista de palabras vacías (stopwords) y una lista de palábras en otros idiomas que pueden consultarse en el apéndice.

Luego, quitamos las palabras excesivamente frecuentes (aquellas que aparecen más del 50 % en todo el corpus de documentos) y las palabras que aparecen muy poco frecuente (menos de cinco veces en todo el corpus), con la idea de que las palabras demasiado frecuentes y aquellas muy rara vez utilizadas, no sirven para discriminar bien la información relevante.

El siguiente paso es someter esta matriz a un algoritmo llamado Descomposición del Valor Singular (SVD) para reducir el número de dimensiones sin que se pierda la información sustancial de la matriz original. En nuestro caso, redujimos la matriz a 400 dimensiones (ya que un valor entre 200 y 500 es el recomendado como un buen estandard [11]).

El código para generar la matriz de LSA se encuentra disponible en *generadorLSA.py* y utiliza la librería gensim [12], y en *article_fetcher.rb* se encuentra el código para descargar los artículos al azar de la Wikipedia en Español.

Análisis de textos a partir de LSA. A partir de la matriz LSA construida sobre el corpus de la Wikipedia en Español, comparamos en cada corpus de discursos tanto la distancia entre los demás corpus como la distancia a un concepto seleccionado. Para eso, transformamos el concepto y el corpus de los discursos al espacio vectorial de la matriz LSA, y luego medimos el grado de similitud mediante el coseno entre los dos vectores. Valores cercanos a 1 representan una gran similitud, valores cercanos a -1 que son muy disímiles, y valores cercanos a 0 que no hay correlación. Las funciones auxiliares utilizadas en el análisis pueden consultarse en *analizador.py*

El análisis comparativo entre cada corpus de discursos presidenciales consistió en analizar dos dimensiones pertenecientes al *discurso populista*. La construcción de un significante vacío, que por simpleza y tradición identificamos con el concepto de **pueblo** y la construcción de una frontera antagónica que asociamos al concepto de **enemigo**. Esto nos permitió obtener un índice de acercamiento de los discursos hacia fenómenos que suelen caracterizar al populismo: el nivel *antágonico* del discurso y el hecho de apelar a una identidad popular común.

Para este análisis utilizamos sólo los discursos de líderes de gobierno con mandato presidencial ocurrido en las últimas tres décadas (Raúl Alfonsín, Michelle Bachelet, Hugo Chávez, Rafael Correa, Fernando de la Rúa, Cristina Fernández de Kirchner, Néstor Kirchner, Ricardo Lagos, Mauricio Macri, Nicolás Maduro, Carlos Menem, Evo Morales, Augusto Pinochet, Tabaré Vázquez) y calculamos para cada uno su distancia a los conceptos de **enemigo** y **pueblo**. A fin de obtener un cálculo de la distancia de cada presidente respecto al promedio del grupo entero, convertimos los puntajes medios LSA de cada presidente según condición en puntajes Z.

Una vez efectuado dicho estudio, comparamos el valor LSA medio de los presidentes que superaban 1 desvío estándar respecto a la media (z-score > 1; casos únicos) con la muestra de presidentes por abajo de 1 desvío estándar (z-score < 1; grupo control). Para ello, utilizamos un t-test modificado [13–17] que permite evaluar la significancia a través de la comparación de múltiples puntuaciones individuales con valores normativos

derivados de pequeñas muestras (5 sujetos). Este test estádisitco resulta más robusto que las distribuciones no-normales. Este controla efectivamente errores de tipo I y demuestra robustez respecto a otros métodos [15].

Fuentes de datos. El LSA fue generado a partir de artículos extraidos de la Wikipedia en Español. Los discursos presidenciales son todos correspondientes a la apertura de sesiones legislativas con el fin de minimizar que los discursos traten sobre coyunturas particulares de cada país.

Resultados

Relación directa entre discursos. Un primer objetivo es evaluar si la matriz LSA obtenida a partir de los artículos de la Wikipedia y con la reducción de dimensionalidad propia de la técnica, mantiene de todas manera la información necesaria para encontrar relacciones entre los discursos presidenciales. Para eso, en la figura 1 realizamos una relación entre los discursos de presidencia con los del resto en el espacio vectorial de la matriz LSA.

En la relación entre ellos podemos ver cómo los discursos se relacionan no simplemente por el país al que pertencen, sino también por época. De esta manera, los discursos de Juan Manuel de Rosas, Justo José de Urquiza, Bartolomé Mitre y Domingo Faustino Sarmiento guardan una mayor relación entre sí, que los de Raúl Alfonsín, Carlos Saul Menem, Fernando De la Rúa, Eduardo Duhalde, Néstor Kirchner y Cristina Fernández de Kirchner, pese a que todos pertenecen a la República Argentina.

Por otro lado, también se destaca claramente en la figura 1 los cluster de presidentes chilenos (Salvador Allende, Augusto Pinochet*, Eduardo Frei, Ricardo Lagos, Michelle Bachelet y Sebastián Piñera), los de los presidentes uruguayos (Gregorio Álvarez*, Julio Sanguinetti, Luis Lacalle, Jorge Batlle, Tabaré Vázquez, José Mujica) y los presidentes venezolanos (Hugo Chávez y Nicolás Maduro).

Es interesante destacar el caso de los tres presidentes uruguayos más antiguos del cuerpo de datos: Juan María Bordaberry*, Óscar Gestido y Aparicio Méndez*, que pese a las referencias a Uruguay en sus discursos, guardan muy poca relación con sus compatriotas.

También cabe señalar que la relación no es exclusiva del país. Así como notamos diferencias entre presidentes de los mismos países que no guardan relación entre sí, podemos ver cómo los discursos de Rafael Correa (presidente de Ecuador) guardan una relación con los discursos de Chávez, Alfonsín, Menem, De La Rúa, Duhalde, Kirchner y Morales. O cómo el discurso de Cristina Fernández de Kirchner guarda mayor relación con el de su esposo por sobre el resto.

Relación a partir de conceptos. Respecto a la relación del concepto **enemigo** con los discursos, se puede ver como los discursos de Nicolás Maduro y Rafael Correa muestran los valores más altos, mientras que los discursos de Fernando de la Rúa y Michelle Bachelet presentan alta variabilidad (Fig. 2). Cuando consideramos el concepto **pueblo** se destacan Evo Morales y nuevamente Nicolás Maduro con los valores más altos, mientras que Tabaré Vázquez muestra alta variabilidad (Fig. 3).

*Presidente de facto

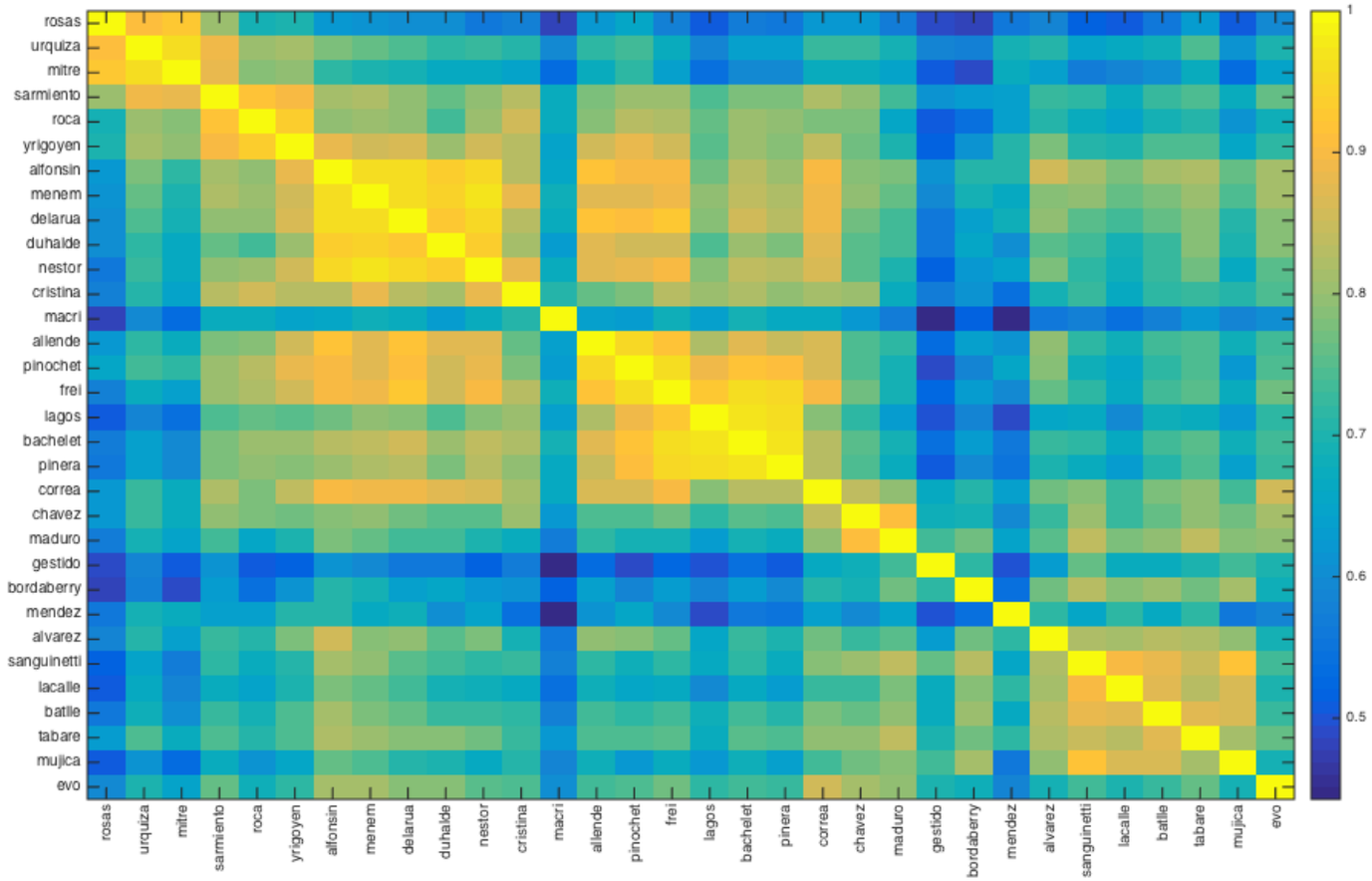

Figura 1: Matriz de relación entre discursos

A partir del examen de los Z Scores (Fig. 4) es posible observar que Nicolás Maduro, Hugo Chávez, Rafael Correa y Tabaré Vázquez para el concepto **enemigo** y tanto Nicolás Maduro como Evo Morales para **pueblo** se distancian mayormente del promedio respecto al resto del grupo.

Para el análisis de comparación utilizamos como casos únicos solamente aquellos presidentes con valores Z mayores a 1 (Fig. 5) ya que nuestro interés era evaluar los discursos que estaban principalmente asociados con los conceptos seleccionados. Los resultados son reportados en la tabla 1, en síntesis estos muestran diferencias significativas entre cada caso único y el grupo de control.

Discusión

Lo primero que nos gustaría destacar es que si bien la transformación de un discurso al espacio semántico de LSA implica una pérdida de información, igual logra mantener información relevante para analizar los discursos. Esto puede notarse en la matriz de relación entre discursos (Fig. 1) que, como mencionábamos en la sección de resultados, muestra un clustering razonable para los presidentes que va más allá del país (y por ende la mera aparición del nombre de cada país).

Sin embargo, la poca cantidad de discursos con los que contamos de cada presidente y de cada país (las aperturas de sesiones son una vez por año, no existen mandatos muy largos de un mismo presidente y no todos los discursos de apertura estaban disponibles), hace que el análisis pueda estar fuertemente influenciado por variantes dialécticas de la época y del país del discurso que hagan difícil su uso efectivo. Pese a eso,

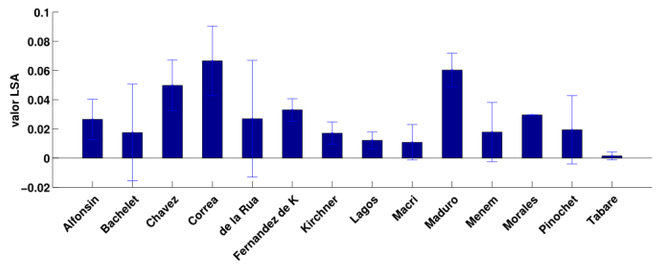

Figura 2: Distribución del valor medio de la distancia del discurso al concepto enemigo en LSA

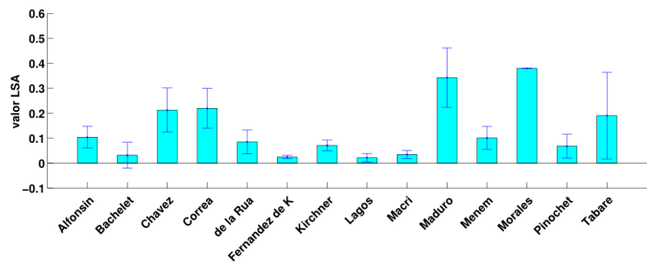

Figura 3: Distribución del valor medio de la distancia del discurso al concepto pueblo en LSA

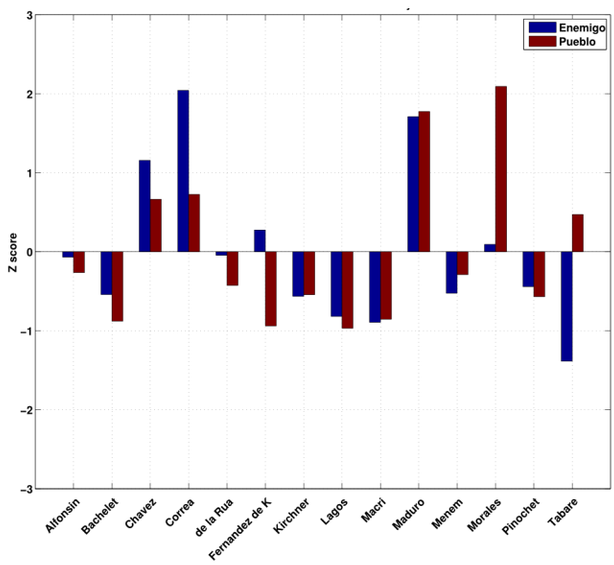

Figura 4: Z Scores de los conceptos enemigo y pueblo por presidente

creemos que los resultados muestran indicios claros de que el análisis de LSA puede ser una herramienta efectiva para el análisis de conceptos de las ciencias políticas en los discursos presidenciales.

Para el caso del populismo hemos dicho que no es esta una categoría teórica estricta de la que exista un consenso en la literatura. Sin embargo, sí existe un consenso en el peso que

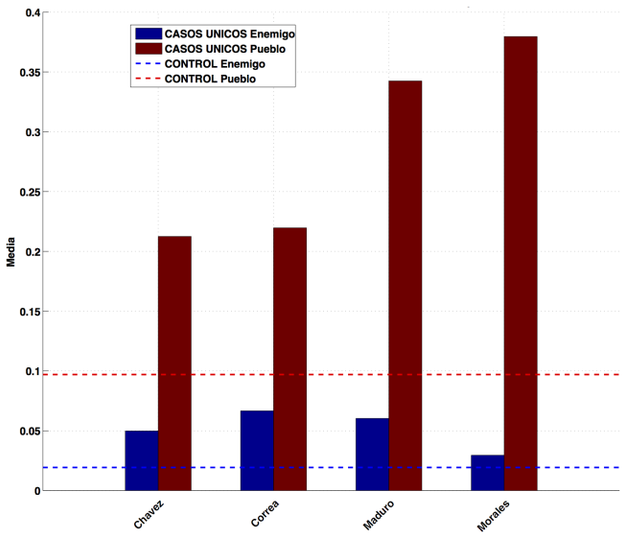

Figura 5: Media grupo control vs presidentes

los discursos tienen en la práctica populista. Por tal motivo, LSA resulta una herramienta natural para su análisis.

El caso más paradigmático de los resultados fue el de Nicolás Maduro que obtuvo una media alta en sus discursos respecto del resto de los presidentes tanto para el concepto de **enemigo**, como para el concepto de **pueblo**. Y en menor medida, esto también se ve para el caso de Rafael Correa, Hugo Chavez y Evo Morales. Lo interesante es que, pese a que no siempre suele haber un consenso claro sobre qué presidentes son considerados populistas, sin dudas estos presidentes son caracterizados de esta manera tanto por especialistas como Laclau, como por la mayoría de la literatura.

Por otro lado, presidentes como Néstor Kirchner o Cristina Fernández de Kirchner, no muestran valores altos para los conceptos. Esto no significa que sus gobiernos puedan o no ser caracterizados como populistas sino que permite entender los alcances y limitaciones de esta técnica para el análisis del populismo. Por un lado, el populismo implica la construcción de un significante vacío y una frontera antagónica distinta en cada contexto que adquiere su propio significado en la disputa hegemónica (pueblo-oligarquía, kirchnerismo-corporaciones, revolución ciudadana - pelucones, pueblo-costa). Distinguir estos conceptos con una única denominación a través de distintos períodos y países, no permite distinguir de manera efectiva la presencia o ausencia de dichos significados como para caracterizar las construcciones políticas.

Nuevamente, los principales inconvenientes resultan ser la escasa cantidad de discursos para poder analizar en cada presidente la presencia o ausencia de estos antagonismos; el poco consenso que existe en torno a la pertenencia de cada líder político en la categoría; la propia práctica populista de la construcción hegemónica que implica cierto juego con los significantes y las cadenas de equivalencia.

Pese a eso, creemos que el trabajo (además del objetivo principal de familizarnos con la técnica de LSA), nos dió indicios claros de que el método puede resultar efectivo para analizar la presencia de elementos típicos de la construcción populista en los discursos.

Presidentes	Enemigo (Media; Desviación estándar (máx y mín))				Pueblo (Media; Desviación estándar (máx y mín))			
	M = 0.0193; DE = 0.0092 (0.0014-0.0330)				M,= 0.0970; DE = 0.0725 (0.0214-0.2196)			
	LSA	**t**	**p**	**z-cc**	**LSA**	**t**	**p**	**z-cc**
Chavez	0.0498	3.16	0.01	3.30	-	-	-	-
Correa	0.0666	4.90	<0.01	5.12	-	-	-	-
Maduro	0.0603	4.25	<0.01	4.44	0.3424	3.25	<0.01	3.39
Morales	-	-	-	-	0.3796	3.75	<0.01	3.90

Tab. 1. Comparaciones (Crawford) entre CTL y presidentes > 1 desvío estándard

Appendix: Código fuente

El código utilizado para el trabajo se encuentra disponible en https://github.com/sromano/lsatp

Appendix: Palabras removidas

Palabras vacías: a, al, algo, algunas, algunos, ante, antes, como, con, cual, cuando, de, del, desde, donde, durante, e, en, entre, era, erais, eran, eras, eres, es, esa, esas, ese, eso, esos, esta, estaba, estabais, estaban, estabas, estad, estada, estadas, estado, estados, estamos, estando, estar, estaremos, estará, estarán, estarás, estaré, estaréis, estaría, estaríais, estaríamos, estarían, estarías, estas, este, estemos, esto, estos, estoy, estuve, estuviera, estuvierais, estuvieran, estuvieras, estuvieron, estuviese, estuvieseis, estuviesen, estuvieses, estuvimos, estuviste, estuvisteis, estuviéramos, estuviésemos, estuvo, está, estábamos, estáis, están, estás, esté, estéis, estén, estés, fue, fuera, fuerais, fueran, fueras, fueron, fuese, fueseis, fuesen, fueses, fui, fuimos, fuiste, fuisteis, fuéramos, fuésemos, ha, habida, habidas, habido, habidos, habiendo, habremos, habrá, habrán, habrás, habré, habréis, habría, habríais, habríamos, habrían, habrías, habéis, había, habíais, habíamos, habían, habías, han, has, hasta, hay, haya, hayamos, hayan, hayas, hayáis, he, hemos, hube, hubiera, hubierais, hubieran, hubieras, hubieron, hubiese, hubieseis, hubiesen, hubieses, hubimos, hubiste, hubisteis, hubiéramos, hubiésemos, hubo, la, las, le, les, lo, los, me, mi, mis, mucho, muchos, muy, más, nada, ni, no, o, os, para, pero, poco, por, porque, que, quien, quienes, qué, se, sea, seamos, sean, seas, seremos, será, serán, serás, seré, seréis, sería, seríais, seríamos, serían, serías, seáis, sido, siendo, sin, sobre, sois, somos, son, soy, su, sus, suya, suyas, suyo, suyos, sí, también, tanto, te, tendremos, tendrá, tendrán, tendrás, tendré, tendréis, tendría, tendríais, tendríamos, tendrían, tendrías, tened, tenemos, tenga, tengamos, tengan, tengas, tengo, tengáis, tenida, tenidas, tenido, tenidos, teniendo, tenéis, tenía, teníais, teníamos, tenían, tenías, ti, tiene, tienen, tienes, todo, tu, tus, tuve, tuviera, tuvierais, tuvieran, tuvieras, tuvieron, tuviese, tuvieseis, tuviesen, tuvieses, tuvimos, tuviste, tuvisteis, tuviéramos, tuviésemos, tuvo, un, una, uno, unos, y, ya, éramos.

Palabras extranjeras: des, du, est, sciences, star, the, of, commons, and, wikimedia, fishes, wikispecies, wikipedia, www, gnu, unported, you, university, insee, music, school, org, love, sub, wars, science, one, king, live, list, society, angry, emelec, et, dans, au, pour, dee, page, to, in, creative, oil, polish, revenge, strikes, blu, phantom, bloom, agent, fellowship, avengers, zone, tails, pass, bear, heads, google, challenger, wii, nintendo, elementary, xbox, keys, pride, renault

1. Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. Discourse processes, 25(2-3):259–284, 1998.
2. Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2):211, 1997.
3. Aníbal Viguera. "populismoz"neopopulismo.[en] américa latina. Revista Mexicana de Sociología, pages 49–66, 1993.
4. Carlos Vilas. Entre la democracia y el neoliberalismo: los caudillos electorales de la posmodernidad. Socialismo y participación, 69:31–43, 1995.
5. Carlos María Vilas. La democratización fundamental: el populismo en América Latina. Consejo Nacional para la Cultura y las Artes, 1995.
6. Aboy Carlés. Repensando el populismo. Política y gestión, 4, 2003.
7. Gerardo Aboy Carlés. Populismo y democracia en la argentina contemporánea. entre el hegemonismo y la refundación. Estudios sociales, 28(1):125–149, 2005.
8. Ernesto Laclau. Populismo:¿ qué nos dice el nombre? In El populismo como espejo de la democracia, pages 51–70. Fondo de Cultura Económica, 2009.
9. Ernesto Laclau. La razón populista. Fondo de cultura Económica, 2012.
10. Ernesto Laclau and Chantal Mouffe. Hegemonía y estrategia socialista. Siglo Veintiuno de España ed., 1987.
11. Roger B. Bradford. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, pages 153–162, New York, NY, USA, 2008. ACM.
12. Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May 2010. ELRA. .
13. John R Crawford and David C Howell. Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist, 12(4):482–486, 1998.
14. John R Crawford, Paul H Garthwaite, and Sara Porter. Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. Cognitive Neuropsychology, 27(3):245–260, 2010.
15. John R Crawford and Paul H Garthwaite. Single-case research in neuropsychology: a comparison of five forms of t-test for comparing a case to controls. Cortex, 48(8):1009–1016, 2012.
16. John R Crawford, Paul H Garthwaite, and David C Howell. On comparing a single case with a control sample: An alternative perspective. Neuropsychologia, 47(13):2690–2695, 2009.
17. John R Crawford, Paul H Garthwaite, and Kevin Ryan. Comparing a single case to a control sample: testing for neuropsychological deficits and dissociations in the presence of covariates. Cortex, 47(10):1166–1178, 2011.