XXII Seminario de Ingeniería Biomédica
Núcleo de Ingeniería Biomédica
Facultades de Medicina e Ingeniería
UdelaR

Introducción a LSA (Latent Semantic Analysis)

Gabriel Slomovitz

gslomo@gmail.com

Resumen. Latent Semantic Analysis es un método novedoso para extraer los componentes principales de un cuerpo de texto que ha sido utilizado para categorización y búsqueda de información. Sin embargo, por los resultados sorprendentes en lo que refiere a similitud con procesos realizados por humanos se convierte en una teoría promisoria para explicar procesos cognitivos. La presente introducción al tema detalla el principio de funcionamiento del método, basado en la técnica algebraica conocida como descomposición en valores singulares y resume algunos estudios para su aplicación en detección de esquizofrenia.

1. Introducción

LSA (Análisis Semántico Latente) es un método que permite analizar texto y extraer relaciones de similitud de significado entre palabras o entre pasajes de palabras. LSA utiliza algunas operaciones algebraicas para representar a las palabras en un espacio multidimensional. En particular, aplica una técnica de descomposición matricial denominada descomposición en valores singulares (SVD por sus siglas en inglés) para extraer los componentes principales de la representación del texto. Estas representaciones permiten simular algunos procesos cognitivos que ocurren en el cerebro, tales como: reconocimiento de vocabulario, comprensión de lenguaje, evaluación de composición de relato [1].

Las estimaciones producidas por LSA son inferidas a partir de operaciones matemáticas que logran extraer relaciones subyacentes que no son detectables utilizando métodos más directos como análisis de co-occurrencia , contigüidad, correlacionamiento de palabras u otros métodos estadísticos.

LSA utiliza como entrada solamente el texto a analizar, y si bien tiene limitaciones ya que no utiliza el orden de las palabras ni relaciones sintácticas o lógicas, consigue resultados que sorprenden por su similitud con análisis realizados por humanos.

2. Funcionamiento

LSA representa el texto a analizar en una matriz que se conforma del siguiente modo. Cada fila de la matriz corresponde a una palabra en particular, y cada columna corresponde al pasaje o contexto en el que aparece. Las celdas de la matriz contienen la cantidad de veces que la palabra aparece en el pasaje. Luego se aplica el método de descomposición en valores singulares (SVD) que permite descomponer una matriz rectangular $\boldsymbol{X}$ (de $\boldsymbol{t}$ columnas por $\boldsymbol{p}$ filas) en el producto de tres matrices:

$$\boldsymbol{X} = \boldsymbol{T} * \boldsymbol{S} * \boldsymbol{P}^{\boldsymbol{T}} \qquad (1)$$

Donde ***X*** es la matriz original, ***T*** es una matriz de ***t*** x ***r*** de columnas ortonormales, ***P*** es una matriz ***p*** x ***r*** de columnas ortonormales y ***S*** es una matriz de ***r*** x ***r*** diagonal ordenada de manera decreciente. Los componentes de la matriz ***S*** son los valores singulares, y las matrices ***T*** y ***P*** corresponden a los vectores de palabras y de pasajes, respectivamente. Puede demostrarse que es posible descomponer cualquier matriz rectangular de esta manera [1].

Se puede reducir la dimensionalidad de la matriz original utilizando para su reconstrucción solo algunos coeficientes de la matriz diagonal ***S***, tal como se describe en (2), donde se utilizan los primeros ***k*** valores singulares.

$$X' = T_k * S_k * P_k^T \qquad (2)$$

Esto es equivalente a aproximar por mínimos cuadrados [2]; ***X'*** es la mejor aproximación a ***X*** por mínimos cuadrados. Es justamente en esta reducción de dimensiones que reside la gran capacidad del método para lograr extraer los componentes principales de los datos a analizar.

Se describe a continuación una ejemplo simple de aplicación del método [1], utilizando pasajes de texto que corresponden a nueve artículos técnicos sobre "Human Computer Interaction" y cuatro sobre "Graph theory". La *Figura 1* muestra el texto original, los títulos de ambos temas aparecen agrupados en **c** y **m** respectivamente.

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Figura 1 - Texto original

La matriz ***X*** se conforma según se muestra en *Figura 2.*

$\{X\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Figura 2 - Matriz palabra-pasaje

En la *Figura 3* se presenta el resultado de la descomposición en valores singulares (SVD) y en la *Figura 4* se muestra la reconstrucción de la matriz ***X*** utilizando solamente dos dimensiones, es decir utilizando solo las dos primeras columnas de las matrices.

$$\{X\} = \{W\}\{S\}\{P\}'$$

$\{W\} =$

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

$\{S\} =$

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

$\{P\} =$

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

Figura 3 - Matrices correspondientes a la descomposición en SVD

$\{\hat{X}\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Figura 4- Matriz X reconstruida

En términos generales, la matriz reconstruida permite describir un segmento de texto a partir de la cuantificación de las palabras que lo componen. A su vez, cada palabra puede verse como la suma de sus aportes en cada segmento. A partir de estas cuantificaciones, pueden estimarse relaciones de similitud entre palabras o entre pasajes de texto. Esta idea se describe con más detalle a continuación comparando las filas correspondientes a las palabras *human, user* y *minors.* En la matriz original se observa que la palabra *human* nunca aparece en el mismo pasaje que *use*r y *minor.* El coeficiente de correlación Spearman [1] entre *human* y *user* es -0.38 y entre *human* y *minors* 0.29. Sin embargo, en la matriz reconstruida los coeficientes de correlación se transforman en 0.94 para *human-user* y -0.83 para *human-minor*. Pese a que *human* y *user* nunca aparecen en el mismo pasaje, el procedimiento de reducción de dimensiones permite descubrir relaciones subyacentes. Como los contextos en los que aparecen las palabras tienen un significado similar, LSA las representa como similares y lo opuesto ocurre para *human* y *minor*.

En lo que refiere a comparaciones de similitud de palabras o pasajes, la manera más simple es calcular el coseno entre los vectores que resultan de la descomposición en valores singulares y su posterior reducción de dimensiones. Adicionalmente, pueden implementarse otras medidas de similitud como distancia Euclideana o “city block distance” [2].

3. Aplicaciones

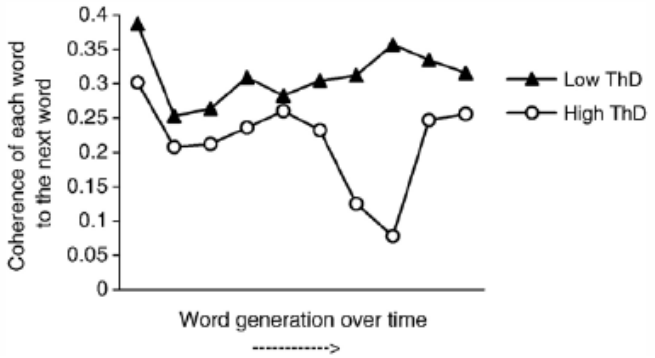
LSA ha sido utilizado inicialmente como método de búsqueda de información y categorización de texto. Sin embargo, su habilidad para descubrir estructuras escondidas y asociar objetos similares lo convierten en una técnica adecuada para modelar ciertos procesos que ocurren en el cerebro [5]. Esto abre un enorme campo de investigación en el ámbito de ciencias cognitivas.

En lo que refiere a aplicaciones médicas en particular, LSA ha sido propuesto como método para evaluar desórdenes del lenguaje en pacientes con esquizofrenia. Este análisis permite cuantificar el grado de coherencia del discurso del paciente, constituyendo una herramienta de evaluación objetiva del nivel de enfermedad.

En [6] se establece que pacientes con desórdenes de pensamiento (ThD por sus siglas en inglés) y en particular con esquizofrenia, muestran anormalidades en el uso del lenguaje. Los métodos tradicionales de evaluación de estos pacientes, comprenden análisis clínicos basados en entrevistas, incluyendo cuestionarios y preguntas de final abierto. Existen estándares para calificar el desempeño del paciente por parte del médico que realiza el diagnóstico, según se establece en [7]. De estas evaluaciones se puede derivar una calificación global que indica el grado de desorden en las habilidades de comprensión, lenguaje y comunicación (TLC por sus siglas en inglés). El valor 0 significa ausencia de desorden, y el valor 4 indica desorden extremo.

En [6] se realizan tres tipos de experimentos con pacientes catalogados con diferentes niveles de esquizofrenia, así como con un grupo de control. El primer experimento tiene que ver con asociación de palabras. El paciente tiene que expresar la primera palabra que le viene en mente, luego de leer 10 palabras consecutivas (“Dios”, “comida”, “niño”, “oscuro”, “difícil”, “alto”, “rey”, “mesa”, “lento” y “hombre”). Se espera que pacientes con ThD muestren una asociación de palabras menos usual que el grupo de control. El grado de similitud entre las palabras propuestas y la respuesta del paciente es medido utilizando LSA, en particular usando los algoritmos desarrollados en [8]. Se compara el resultado de los cosenos entre los vectores, haciendo análisis de varianza (Anova). Los resultados muestran que LSA es capaz de evaluar diferencias sutiles en grado de ThD. Adicionalmente, estos resultados se correlacionan significativamente con los obtenidos por evaluación clínica de ThD.

El segundo experimento busca medir la fluidez verbal. Se le pide al paciente que exprese la mayor cantidad de “animales” que se le ocurra, en un minuto. Esta es una prueba muy utilizada en pacientes con esquizofrenia, ya que habitualmente muestran una capacidad reducida en las respuestas esperadas. El experimento mide, utilizando LSA, la similitud entre palabras consecutivas. Los resultados pueden observarse gráficamente en la *Figura 5*. El grupo de pacientes con ThD bajo muestra un grado de coherencia más alto que el grupo de ThD alto. El método logra tener una mejor sensibilidad para este tipo de pruebas, que las técnicas tradicionales.



El tercer experimento consiste en hacer que el paciente hable por algunos minutos, respondiendo ciertas preguntas o contando una pequeña historia. Se utiliza una ventana deslizante de unas pocas palabras y se mide, utilizando LSA, el grado de similitud de estas palabras con la pregunta realizada. A medida que la ventana avanza en el tiempo, se espera que la coherencia disminuya. Asimismo, a medida que se agranda la ventana, el grado de coherencia disminuye tal como se muestra en la *Figura 6*.

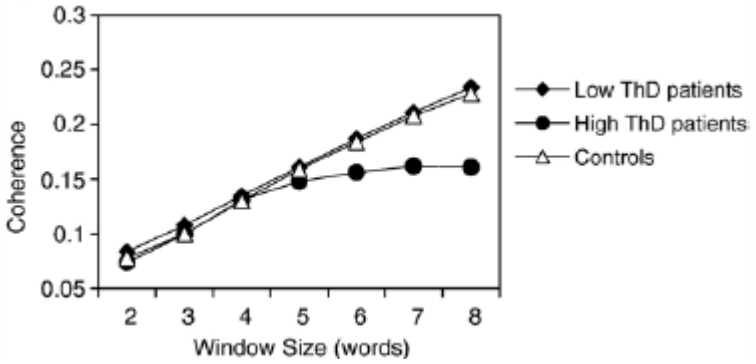


Figura 6 - Coherencia en texto con ventana deslizante

Como conclusión de estos estudios puede afirmarse que LSA es capaz de detectar pacientes con esquizofrenia con una exactitud razonable, así como evaluar la severidad de incoherencia en el uso del lenguaje de estos pacientes.

En [9] se analiza otra orientación del tema que destaca la importancia de detectar desviaciones o desórdenes sutiles en el discurso de los pacientes pero en etapas tempranas, previas a la aparición de los primeros síntomas. Para ello, se estudian grupos de pacientes con esquizofrenia pertenecientes a una misma familia, grupos de pacientes no familiares y grupos de control. Se utiliza LSA en conjunto con otras técnicas, con el objetivo de demostrar que los experimentos pueden diferenciar de manera

exitosa discursos pertenecientes a pacientes y familiares de esos pacientes, de discursos pertenecientes a otros grupos. De esta manera se busca establecer que es posible detectar diferencias psicológicas sutiles que se expresan en el lenguaje, correspondientes al carácter hereditario de la enfermedad.

En [10] se plantea la posibilidad de utilizar LSA para realizar evaluaciones de narrativa psiquiátrica que simulen un nivel de experto. Esto es, realizar un sistema que automáticamente pueda extraer los conceptos clínicos del texto, sin la necesidad de un médico experto. Para ello, se usa LSA para generar un espacio semántico, del que se pueden extraer asociaciones entre términos psiquiátricos. Los resultados obtenidos, comparados contra evaluaciones de médicos expertos, muestran que se aproximan a los que se obtendrían por evaluación médica.

4. Conclusiones

Las operaciones algebraicas realizadas por LSA, en particular SVD, pueden verse como un método para identificar y ordenar las dimensiones sobre las cuales los datos exhiben mayor variación. Una vez identificadas las dimensiones, pueden aproximarse los datos por su proyección sobre este espacio. SVD permite reducir los datos a un espacio dimensional más pequeño que expone la subestructura de los datos originales y además los ordena de mayor a menor variación. LSA, simplemente ignora las variaciones por debajo de cierto umbral, reduciendo drásticamente la cantidad de datos pero preservando los principales componentes.

Las mejoras en las últimas décadas de la capacidad de procesamiento computacional han abierto un enorme campo para la aplicación de LSA en diversos ámbitos. Sorprendentemente, los resultados obtenidos son similares a los que se obtendrían si esas tareas fueran realizadas por humanos. Además, se asume que para realizar esas tareas las personas deben “entender” las palabras y pasajes de texto analizadas. Es por esto último que LSA es visto no solamente como un método para categorización o búsqueda de texto, sino como una teoría computacional para análisis del lenguaje y la comprensión. Las aplicaciones para detección de síntomas de esquizofrenia, son una interesante muestra de las capacidades de LSA y sus relaciones con los procesos cognitivos que se desarrollan en el cerebro. Existe un vasto campo de trabajo futuro en esta área, por ejemplo en detección de anormalidades en el uso del lenguaje y la memoria causados por Alzheimer, desorden bipolar, o demencia semántica.

5. Bibliografía

[1] T. K. Landauer, P. W. Foltz, D. Laham, ”Introduction to Latent Semantic Analysis,” Discourse Processes, 25, pp .259-284, 1998.
[2] T. K. Landauer, S. Dumais (2008), Scholarpedia, 3(11):4356.http://www.scholarpedia.org/article/Latent_semantic_analysis
[3] T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsh, "Handbook of Latent Semantic Analysis,” 2011.
[4] K. Baker, “Singular Value Decomposition Tutorial,” 2005.
[5] J. C. Valle-Lisboa, E. Mizraji, “The uncovering of hidden structures by Latent Semantic Analysis,” Information Sciences, 177, pp. 4122–4147, 2007.
[6] B. Elvevag, P. W. Foltz, D. R. Weinberger, T. E. Goldberg, “Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia,” Schizophrenia Research, 93, pp. 304–316, 2007.
[7] N. C. Andreasen, “Scale for the assessment of thought, language and communication (TLC),” Schizophrenia Bulletin, 12, pp. 474–482, 1986.

[8] http://lsa.colorado.edu/
[9] B. Elvevag, P. W. Foltz, M. Rosenstein, L. E. DeLisi, ”An automated method to analyze language use in patients with schizophrenia and their first-degree relatives,” Journal of Neurolinguistics, 23, pp. 270–284, 2010.
[10] T. Cohen, B. Blatter, V. Patel, “Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative,” Journal of Biomedical Informatics, 41, pp. 1070–1087, 2008.