Procesamiento de Lenguaje Natural o Minería de textos
Tema 6: Aprendizaje no Supervisado para análisis de textos.

Objetivo: El participante identificará la técnica de modelado de tópicos para la identificación y análisis de temas en una colección de textos, apoyado en el enfoque de modelado Latent Dirichlet Allocation (LDA) y su implementación en Python.

Temario:

1. Modelado de tópicos: LDA
2. Agrupamiento de documentos

Lecturas:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Ponweiser, M. (2012). Latent Dirichlet allocation in R. Obtenido de https://epub.wu.ac.at/3558/

Priyantina, R., & Sarno, R. (2019). Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM. International Journal of Intelligent Engineering and Systems, 12(4), 142-155. Obtenido de http://www.inass.org/2019/2019083114.pdf

Introducción

El **aprendizaje automático** o aprendizaje automatizado o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es **desarrollar técnicas que permitan que las computadoras aprendan**.

Los modelos de **aprendizaje automático** se pueden dividir en dos grandes familias: **aprendizaje supervisado** y **aprendizaje no supervisado**. La principal diferencia entre estas dos familias se encuentra en los datos de entrenamiento. En el **aprendizaje supervisado** los resultados que se desean obtener del modelo **son conocidos previamente**. Siendo utilizados para guiar su entrenamiento. Por otro lado, en el **aprendizaje no supervisado** el resultado deseado no se utiliza durante el entrenamiento. En la mayoría de los casos tampoco se conoce previamente, siendo descubierto durante el proceso aprendizaje.

Aprendizaje supervisado: En el entrenamiento de los algoritmos de aprendizaje supervisado, además de los datos necesarios para realizar la predicción, es necesario disponer de una característica objetivo para cada una de las instancias.

- **Clasificación**: se desea **obtener una categoría**.
- **Regresión**: en estos se intenta **predecir un valor continuo**.

Aprendizaje no supervisado: A diferencia de los algoritmos de aprendizaje supervisado, en los no supervisados no es necesario **disponer de la respuesta correcta** en los datos de entrenamiento. Ya que no se busca la reproducción de un resultado conocido, sino el **descubrimiento de nuevos patrones o resultados.**

Estos problemas aparentan ser más complejos que los anteriores. Ya que se espera que el modelo aprenda sin decirle el qué. Los problemas más habituales en este tipo de aprendizaje son los de **clúster. En estos se busca grupos de registros que son similares entres si y, al mismo tiempo, diferentes del resto**. Una vez obtenidos los grupos se le ha de asignar una clasificación a cada uno, la cual puede ser conocida o no antes de entrenar el modelo. Lo que muchas veces lleva al descubrimiento de patrones desconocidos.

capaces de dibujar una nube de palabras y quieren saber si se pueden agrupar por temas porque esos 4 nombres que aparecen recurrentemente dan la pista de que podría tratarse de 4 capítulos de una obra en los que se habla de esos autores[1].

Existe una técnica procedente de la IA, del subcampo del aprendizaje automático (*machine learning*), que puede ser de gran ayuda para clasificar estos textos. Es el llamado *topic modeling*, que lo que pretende es identificar, sin la ayuda de ningún diccionario, los temas (*tópicos*) principales que encierra un texto.

Supongamos que en un archivo se ha encontrado un legajo (Conjunto de papeles archivados, generalmente atados, que tratan de un mismo asunto) con más de 450 páginas de texto que parecen tratar de pensadores porque al trascribirlas han aparecido recurrentemente los nombres Freud, Voltaire, Chomsky y Maquiavelo y términos como *lenguaje, lingüística, política, revolución, sociedad, crítica, análisis, social, historia, príncipe, moral, ideas, psicoanálisis*, etc. Tan solo han sido

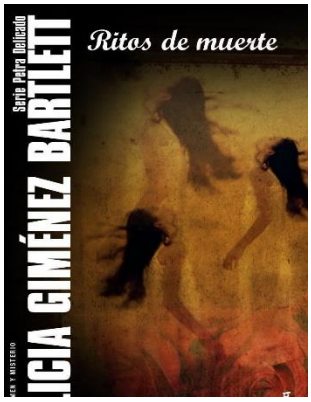
[1] http://www.aic.uva.es/cuentapalabras/topic-modeling.html#topic-modeling-1

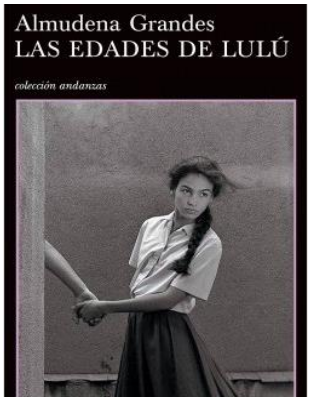
Examina las nubes de palabras de la figura. Cada una de ella tiene las palabras semánticas más frecuentes de 4 novelas españolas e intenta establecer de qué trata cada una de ellas.

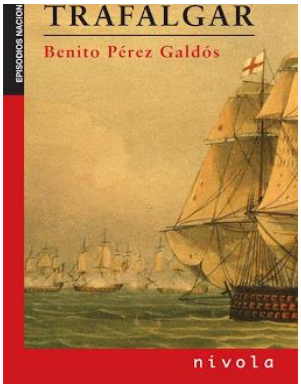
La primera (de arriba hacia abajo y de izquierda a derecha) **es una novela policíaca**, como muestra la ocurrencia de términos como **comisario, subinspector, policía, caso y que el delito parece ser una violación por la aparición de la palabra violador**. En la **segunda** abundan las referencias a diversas partes del cuerpo como **piernas, lengua, brazos, ojos, boca, dedos, cabeza, labios, cara**…; estas palabras **por sí solas no constituyen un tópico** puesto que pueden aparecer en muchos otros tipos de textos; **cabeza y ojos, por ejemplo, aparecen en las cuatro nubes**. Sin embargo, **la ocurrencia de cama y sexo permiten restringir el tema y, podría ser una novela rosa (o erótica). La tercera** parece que se **trata de una batalla naval**, como **lo delatan las palabras navío, escuadra, buque, barcos, combate, muerte, guerra, mar, artillería, cañones y marineros**, y, además, en ella **participan los ingleses. La última** parece situarse también en **un ambiente militar**, **pero** infinitamente **más tranquilo, en la vida de cuartel (capitán, uniforme, reclutas, botas, ejército, militares, sargento, campamento) durante la llamada mili, es decir, el servicio militar obligatorio**.

Lo mismo que acabas de hacer para ver de qué tratan esas novelas, pero has tenido que jugar con tu conocimiento del mundo y con un amplio repertorio léxico, puede hacerlo una máquina que no sabe nada de español, o para el caso de ninguna lengua, pues para ella todo son ceros y unos.

Las cuatro novelas procesadas y representadas en las cuatro nubes de palabras de la figura son:







1. Ritos de muerte, de Alicia Gimenez-Bartlet (1996),
2. Las edades de Lulú, de Almudena Grandes (1989),
3. Trafalgar, de Benito Pérez Galdós (1873) y
4. Ardor guerrero, de Antonio Muñoz Molina (1995).

El proceso matemático que hay tras el modelado de tópicos, como en casi todo lo que estás viendo, es tremendamente complejo, pero el procedimiento, a grandes rasgos es bastante sencillo de entender.

Todo texto presenta un abanico de **tópicos** y esos tópicos se expresan por medio de **palabras**, en especial sustantivos, lo único que tiene que hacer la máquina es contar las palabras y ver cuáles coocurren con

cuáles, y después el investigador debe decidir cuáles son los verdaderos tópicos, pues no todos son tan sencillos de decidir como los que te he mostrado en las nubes de la figura.

¿Qué es modelado de tópicos?

El modelado de tópicos (del inglés, *topic modelling*), básicamente consiste en identificar tópicos o temas en textos, es decir, realizar un análisis de lo que hay en una colección de texto. Se asume que un documento es una mezcla de temas y, por otra parte, los temas se representan como una distribución de palabras.

Un tópico en el contexto de modelado de tópicos es una distribución de probabilidades de palabras para un conjunto, e indica la probabilidad que una palabra aparezca en un documento sobre un tópico en particular.

Existen diferentes enfoques de modelado de temas:

- Análisis semántico latente probabilístico (PLSA) [Hoffman ’99]
- **Latent Dirichlet Allocation (LDA) [Blei, Ng y Jordan, ’03]**
- Basada en aprendizaje profundo (LDA2VEC) [Moody, ‘16]

El modelado de Latent Dirichlet Allocation (LDA) asume que los documentos se producen a partir de una mezcla de temas. Esos temas luego generan palabras basadas en su distribución de probabilidad. Dado un conjunto de datos de documentos, LDA realiza un seguimiento e intenta averiguar qué temas crearían esos documentos en primer lugar. En resumen: este modelo genera tópicos proponiendo una cierta distribución de todas las palabras del corpus, y calcula la distribución de estos tópicos en cada documentos.

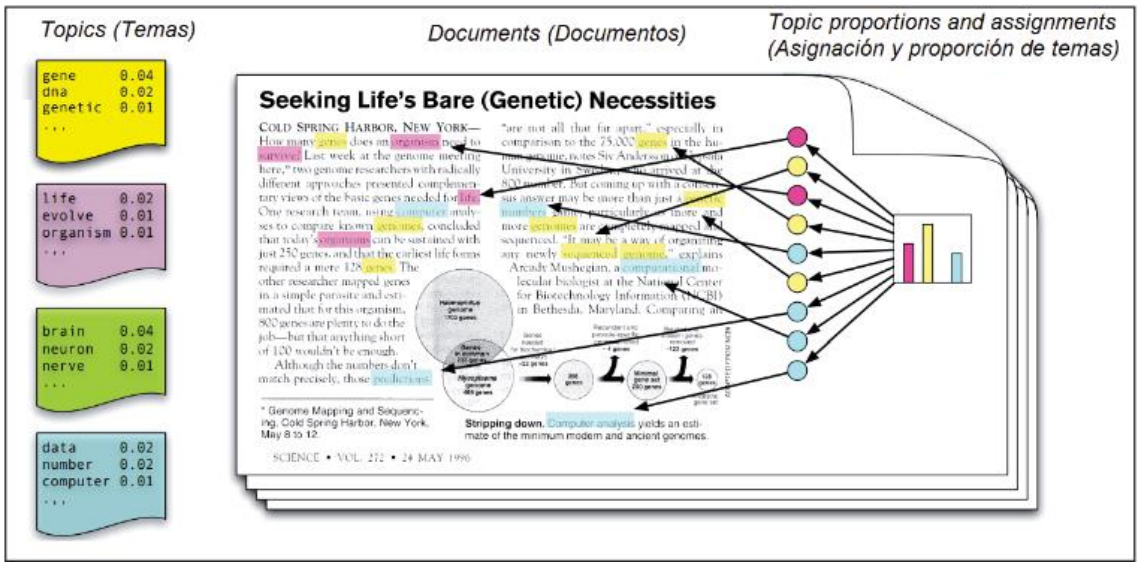
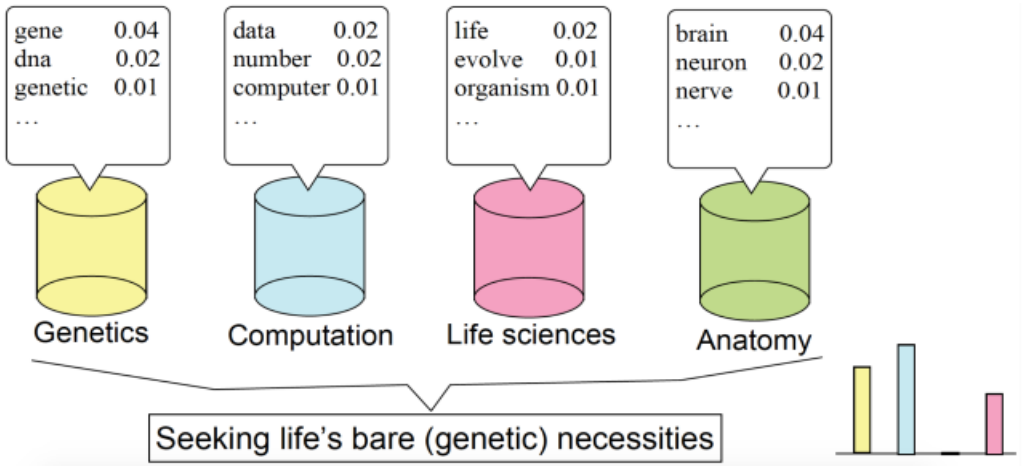


Fig. La intuición detrás de la técnica LDA

Fuente: http://opac.pucv.cl/pucv_txt/txt-5000/UCD5100_01.pdf

La técnica LDA es un modelo estadístico para colección de documentos que intenta capturar esta intuición. En esta se define un tema para pasar a ser una distribución sobre un diccionario de palabras fijado. En la figura se muestra un documento donde se pueden separar y distinguir las palabras que tienen que ver **con análisis como "computer" y "prediction",** las palabras que tienen que ver con **biología evolutiva como "life" y "organism"** y las palabras que hablan **sobre genética como "sequenced" y "genes"**. Si se tomara más tiempo y se resaltara todas las palabras que tienen que ver con análisis, biología evolutiva y genética, se daría cuenta que todo el documento posee estos tres temas mezclados, pero en diferentes proporciones.

Lo interesante de esta manera de operativizar los temas, es que cada tópico puede ser entendido como un campo semántico, un conjunto de palabras que suelen correlacionar en distintos documentos. Luego, en el momento del análisis de estos resultados, buscaremos inferir un tema a partir de las palabras que más contribuyen a cada tópico. Por ejemplo, en el documento analizado en la figura el tema "genética" tiene un vocabulario de palabras que poseen una alta probabilidad de pertenecer al tema "genética". Según uno de los autores del modelo, la interpretabilidad de la mayoría de los temas es el resultado de "la estructura estadística del lenguaje y cómo interactúa con los supuestos probabilísticos específicos de LDA" (D. Blei, 2012, p. 79).



Intuición: documentos como una mezcla de tópicos

(la búsqueda de necesidades básicas (genéticos) de la vida)

LDA es **una técnica de factorización matricial**; convierte la Matriz de Término del Documento (Fig. a) en dos matrices de dimensiones inferiores: $M1$ y $M2$, utilizando técnicas de muestreo para mejorar estas matrices.

- $M1$ es una matriz de temas de documentos de dimensión (N, K) (Fig. b).
- $M2$ es una matriz de temas de dimensión (K, M) (Fig. c)

donde N es el número de documentos, K es el número de temas y M es el tamaño del vocabulario.

	w_1	w_2	w_3	w_m
D_1	0	2	1	3
D_2	1	4	0	0
D_3	0	2	3	1
D_n	1	1	3	0

a. Matriz de Término Documento

	K_1	K_2	K_3	K_n
D_1	1	0	0	1
D_2	1	1	0	0
D_3	1	0	0	1
D_n	1	0	1	0

b. M1: Matriz de temas de documentos

	w_1	w_2	w_3	w_m
K_1	0	1	1	1
K_2	1	1	1	0
K_3	1	0	0	1
K_n	1	1	0	0

c. M2: Matriz de temas con dimensiones

Fig. Factorización matricial del LDA

Itera a través de cada palabra "w" para cada documento "D" e intenta ajustar el tema actual. Se asigna un nuevo tema "K" a la palabra "w" con una probabilidad P que es producto de dos probabilidades p_1 y p_2. Para cada tema, se calculan dos probabilidades p_1 y p_2.

- $p_1 = p(\frac{T}{D})$: la proporción de palabras en el documento y que actualmente están asignadas al tema T.
- $p_2 = p(\frac{w}{T})$: la proporción de asignaciones al tema t sobre todos los documentos que provienen de esta palabra w.

Después de varias iteraciones, se logra un estado estable donde el tema del documento y las distribuciones de los términos del tema son bastante buenos.

Modelado de tópicos en la práctica

¿Cuántos temas? Encontrar o incluso adivinar la cantidad de temas es difícil

Interpretar tópicos

- Los temas son solo distribuciones de palabras.
- Dar sentido a las palabras / generar etiquetas es subjetivo

Es importante tener en cuenta que no siempre todos los tópicos presentarán un campo semántico coherente: en muchos casos pueden referir a regularidades propias del tipo de comunicación que estamos analizando (e.g., palabras que remiten a una interacción por parte del usuario, si es que estamos trabajando con contenido tomado de páginas interactivas), o una mixtura de palabras tal que, en lugar de permitirnos inferir un campo unívoco, nos resulte incoherente.

Luego, debemos organizar nuestros tópicos:

- ***¿Descartamos tópicos irrelevantes?:*** Más allá de los tópicos incoherentes o para los que un campo semántico no es tan evidente, podemos decidir filtrar otros tópicos en vistas de su (ir)relevancia para nuestra pregunta teórica.
- ***¿Agrupar tópicos?:*** Generalmente, en una codificación cualitativa, el proceso se repite iterativamente, haciendo inferencias cada vez más generales (mayor abstracción) y coordinadas

(mayor coherencia), lo que nos permite pasar de los códigos a los temas y argumentos. El modelo LDA no tiene esa estructura jerárquica, pero nosotros podemos agrupar o colapsar tópicos en temáticas más generales. Esto es casi siempre necesario cuando trabajamos con un K elevado.

Se debe tener en cuenta que el algoritmo no tiene información externa sobre los temas encontrados y los artículos no están etiquetados con los temas o palabras claves (es un proceso de aprendizaje no supervisado). La distribución de temas encontrados surge mediante el cálculo de la estructura de temas ocultos que probablemente generan la colección de documentos observados.

¿Por qué es útil el modelado de tópicos?

- **Clasificación de texto**: el modelado de temas puede mejorar la clasificación al agrupar palabras similares en temas en lugar de usar cada palabra como una característica.
- **Sistemas de recomendación**: utilizando una medida de similitud podemos construir sistemas de recomendación. Si nuestro sistema recomendase artículos para lectores, recomendará artículos con una estructura de temas similar a los artículos que el usuario ya ha leído.
- **Descubrir temas en textos**: útil para detectar tendencias en publicaciones en línea, por ejemplo.

Trabajando con LDA en Python

- Muchos paquetes disponibles, como: gensim[2], sklearn[3]

```
conda install -c anaconda gensim
conda install -c anaconda scikit-learn
conda install -c anaconda nltk
```

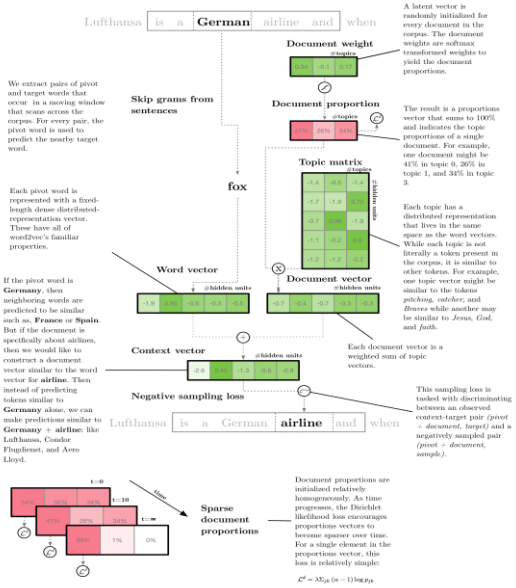
- Pre-procesamiento de texto
  Tokenizar[4], normalizar[5] (minúsculas[6])
  Eliminar palabras cerradas[7]
  Stemming[8] / Lematizar[9]

[2] https://radimrehurek.com/gensim/models/ldamodel.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html
[4] La tokenización es la forma de separar el texto en palabras comúnmente llamadas tokens, usando algunos caracteres como referencia para dividir
[5] Eliminación de signos de puntuación, ya que no agrega ninguna información adicional al tratar los datos de texto; eliminar todos los casos ayuda a reducir el tamaño de los datos de entrenamiento y prueba
[6] evita tener múltiples copias de las mismas palabras
[7] Las palabras vacías es un listado de términos (preposiciones, determinantes, pronombres, etcétera) considerados de escaso valor semántico, que cuando se identifican en un documento se eliminan, sin considerarse términos índices para la colección de textos a analizar. No aportan ningún significado al texto. La supresión de todos estos términos evita los problemas de ruido documental y supone un considerable ahorro de recursos, ya que, aunque se trata de un número relativamente reducido de elementos tienen una elevada tasa de frecuencia en los documentos
[8] Los algoritmos de stemming intentan reducir las palabras flexionadas y derivadas en su forma raíz, es decir, extraer la raíz de una palabra, la raíz lingüística a la que pertenece
[9] Lematización de los términos, es una parte del procesamiento lingüístico que trata de determinar el lema de cada palabra que aparece en un texto. Su objetivo es reducir una palabra a su raíz, de modo que las palabras clave de una consulta o documento se

- Convertir documentos tokenizados a una matriz de termino–documento
- Construir modelos LDA en la matriz termino-documento

Ejercicio4(es)-Aprendizaje Supervisado.ipynb

Tendencias



LDA2VEC: Aprende simultáneamente representaciones de tópicos y representaciones de documentos también.

El modelo lda2vec intenta mezclar las mejores partes de word2vec y LDA en un solo marco. word2vec captura poderosas relaciones entre palabras, pero los vectores resultantes son en gran parte ininterpretables y no representan documentos. LDA, por otro lado, es bastante interpretable por humanos, pero no modela relaciones de palabras locales como word2vec. Creamos un modelo que construye temas tanto de Word como de documentos, los hace interpretables, convierte los temas en clientes, tiempos y documentos, y los convierte en temas supervisados.

https://arxiv.org/abs/1605.02019

Código de experimentos, software de investigación en Python: https://github.com/cemoody/lda2vec

Conclusiones

- El modelado de tópicos es una herramienta exploratoria, frecuentemente utilizada para extracción de textos.
- LDA es un modelo generativo, utilizado extensivamente para modelar grandes corpus de texto
- LDA también se puede utilizar como una técnica de selección de características ,para clasificación de textos y otras tareas

representen por sus raíces en lugar de por las palabras originales. El lema de una palabra comprende su forma básica más sus formas declinadas (formas básicas de las palabras, sin género ni conjugación).