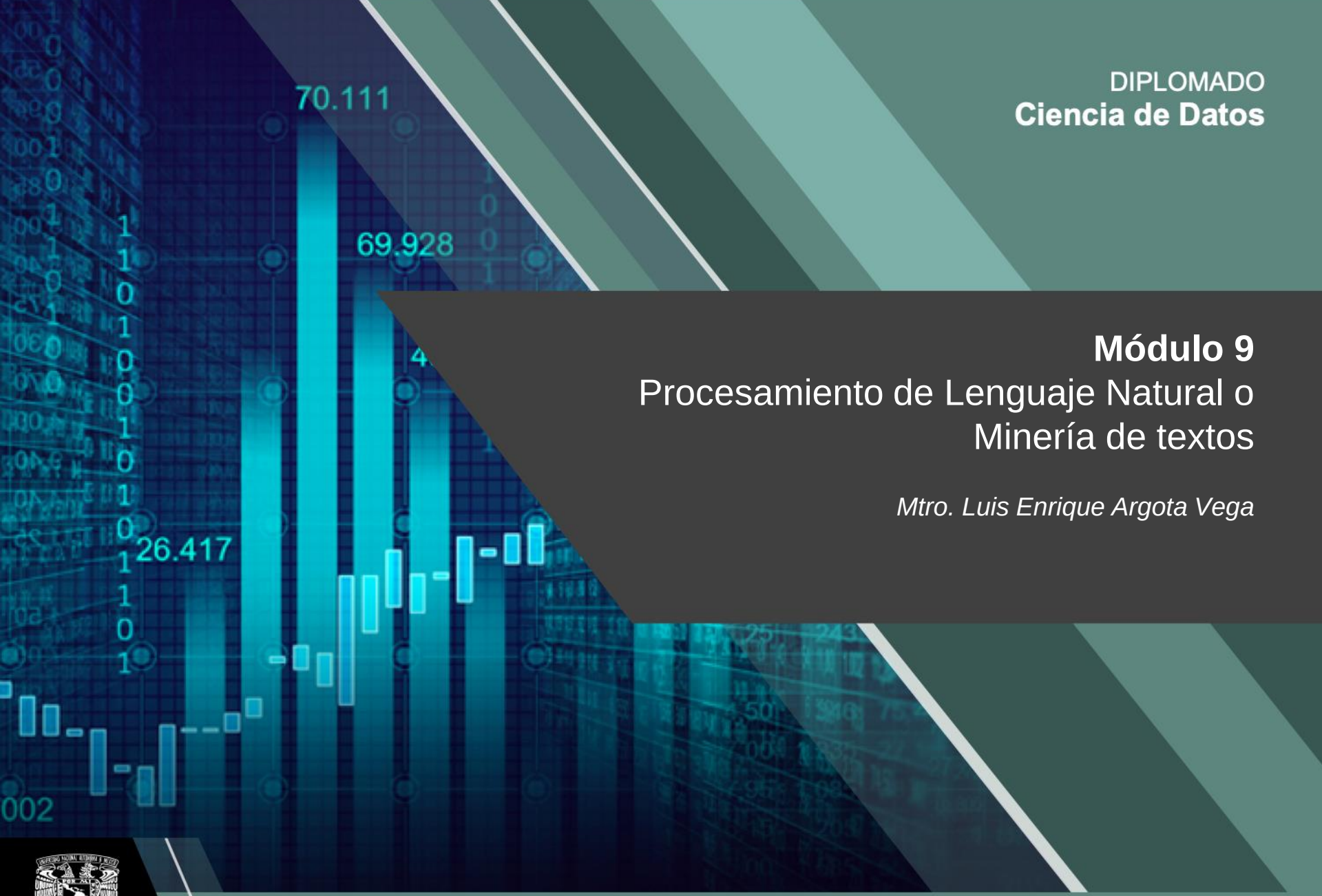


DIPLOMADO
Ciencia de Datos

Módulo 9
Procesamiento de Lenguaje Natural o Minería de textos

Mtro. Luis Enrique Argota Vega

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Tema 5: Modelo de espacio vectorial

Objetivo

El participante identificará cómo mejorar los resultados en una tarea de similitud semántica de textos o ser de gran ayudar para distintas tareas relacionadas con el procesamiento de textos en lenguaje natural, a partir de los modelos basados en la obtención de vectores de palabras y documentos.

Contenido

1. Vectores de palabras (*word embeddings*)
2. Similitud de palabras
3. Vectores de documentos
4. Similitud de documentos
5. Visualización y PCA (Análisis de Componentes Principales)

Introducción

Similitud a nivel de palabras

Similitud léxica de palabras

- Dos palabras son similares a nivel léxico si están compuestas por secuencias parecidas de caracteres.
- Para determinar la similitud léxica de palabras se suelen utilizar distintas métricas basadas en comparación de cadenas de caracteres.

Similitud semántica de palabras

- Nos permite medir si dos palabras tienen significados parecidos o se usan en contextos parecidos.
- Hay dos grandes *grupos de técnicas para calcular la similitud semántica de palabras*: técnicas basadas en conocimiento y técnicas basadas en corpus.

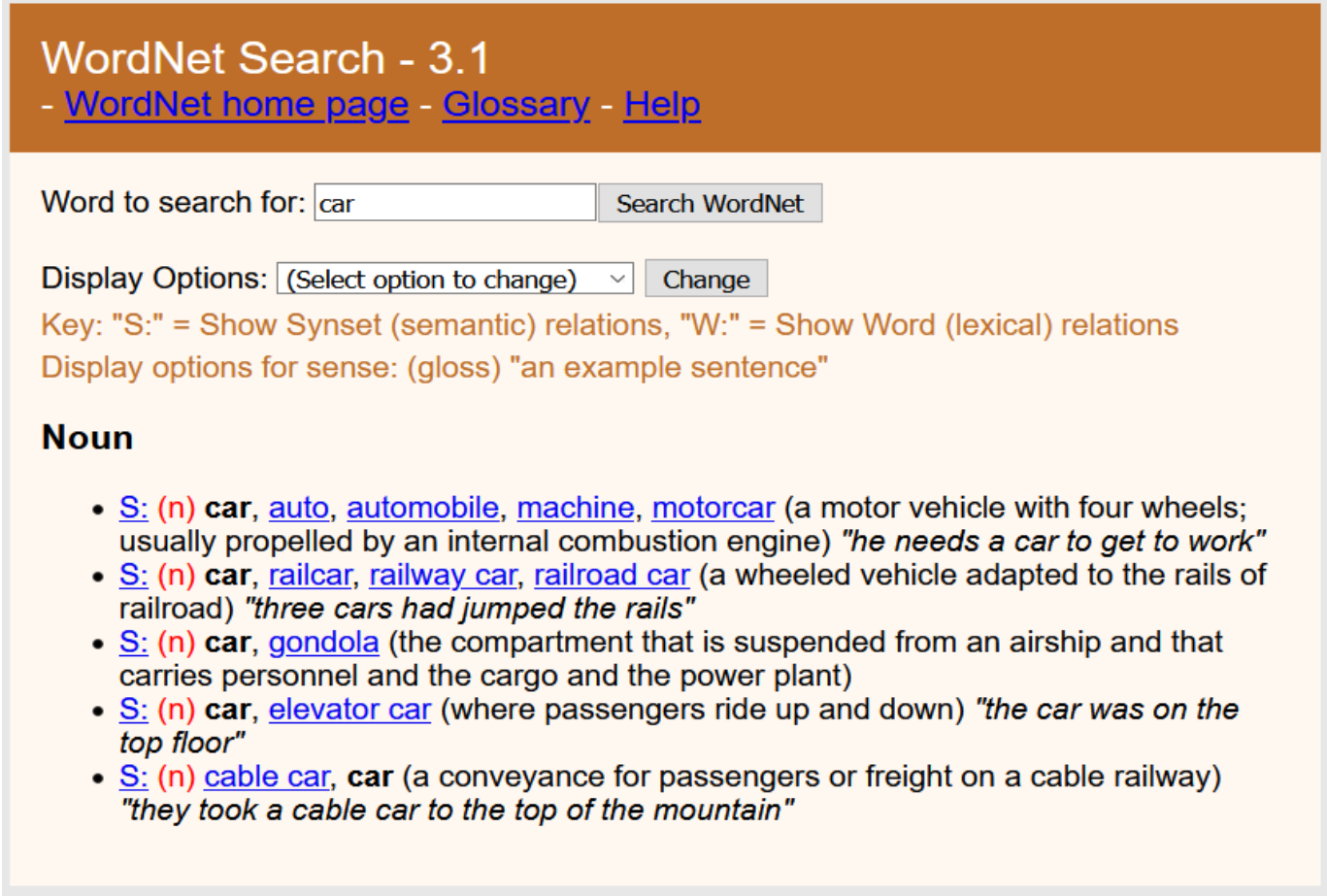
Similitud de palabras basadas en conocimiento

- ✓ Miden el grado de parecido entre palabras apoyándose en algún tipo de recurso lingüístico que proporcione información sobre el significado de las palabras.

- ✓ El recurso por excelencia es ***WordNet*** (Miller,1995), una base de datos léxica organizada en torno a varias relaciones entre palabras.

- ✓ La relación de sinonimia es la más importante y en ella se apoya el concepto de synset (grupo de sinónimos), que permite definir de forma implícita el significado de las palabras a través del conjunto de synsets en los que aparece.

WordNet Online



WordNet online. Obtenido de: http://wordnetweb.princeton.edu/perl/webwn

Meronimia - Sustantivos

- ✓ La relación entre la parte y el todo se mantiene entre los synsets. Las partes se heredan de sus superiores: si una silla tiene patas, entonces un sillón también tiene patas.

- ✓ Las piezas no se heredan "hacia arriba", ya que pueden ser características solo de tipos específicos de cosas, en lugar de la clase en su conjunto: las sillas y los tipos de sillas tienen patas, pero no todos los tipos de muebles tienen patas.

Verbos

- ✓ Los synsets de verbos también se organizan en jerarquías; Los verbos hacia el fondo de los árboles (tropónimos) expresan maneras cada vez más específicas que caracterizan un evento, como en {comunicarse} - {hablar} - {susurro}.

- ✓ Volumen (como en el ejemplo anterior) es solo una dimensión a lo largo de la cual se pueden elaborar verbos. Otros son la velocidad (mover-trotar-correr) o la intensidad de la emoción (querer-amar-idealizar).

- ✓ Los verbos que describen eventos que necesaria y unidireccionalmente se relacionan entre sí están vinculados: {comprar} - {pagar}, {lograr} - {intentar}, etc.

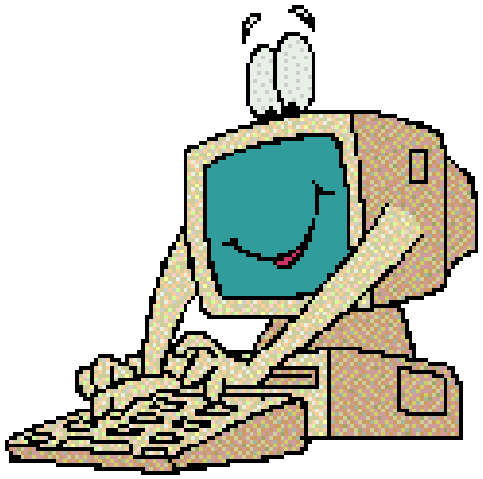
Adjetivos

- ✓ Los adjetivos están organizados en términos de antonimia.
- ✓ Los pares de antónimos "directos" como húmedo-seco y joven-viejo, reflejan el fuerte contrato semántico de sus miembros.
- ✓ Cada uno de estos adjetivos polares, a su vez, está vinculado a una serie de "semánticamente similares". Seco está vinculado a: áridos, desecados y húmedo a empapados, etcétera.
- ✓ Los adjetivos semánticamente similares son "antónimos indirectos" del miembro del polo opuesto.

Adverbios

- ✓ Hay pocos adverbios en WordNet (difícilmente, seguramente, realmente, etcétera), ya que la mayoría de los adverbios en inglés se derivan directamente de los adjetivos a través de la fijación morfológica (sorprendentemente, extrañamente, etcétera)

Práctica



Ejercicio5(es)-Modelo de Espacio Vectorial.ipynb

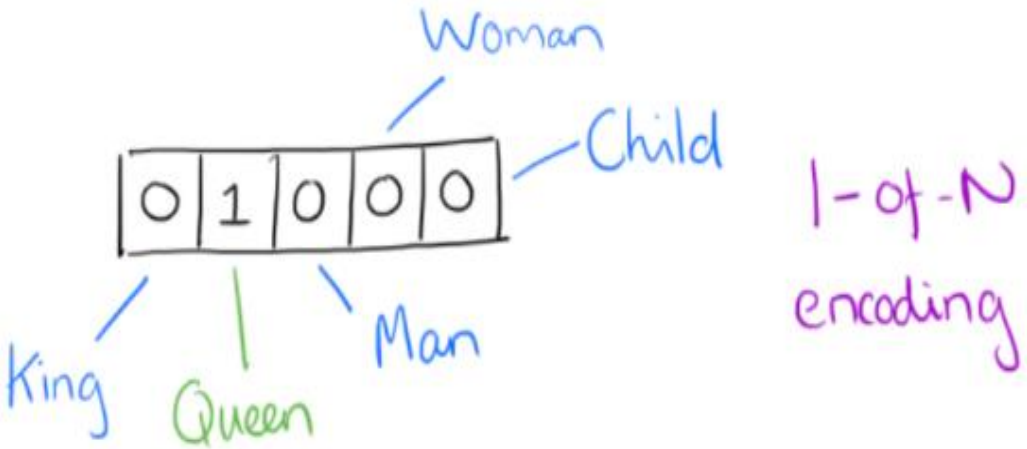
Similitud de palabras basadas en corpus

- ✓ Determinan el parecido semántico de dos palabras en función de los usos de esas palabras en una gran colección de textos.
- ✓ Por lo general, estas técnicas se basan en algún tipo de representación vectorial de las palabras en función de los distintos contextos en los que dichas palabras aparecen.
- ✓ Dentro de las técnicas basadas en corpus destacan las de ***latent semantic analysis*** **(LSA)** y las de ***word embedding***.

Vector de palabras *(word embedding)*

- Es simplemente un vector de pesos

Ejemplo: Codificación 1 de N. Cada elemento del vector es asociado con una del vocabulario.



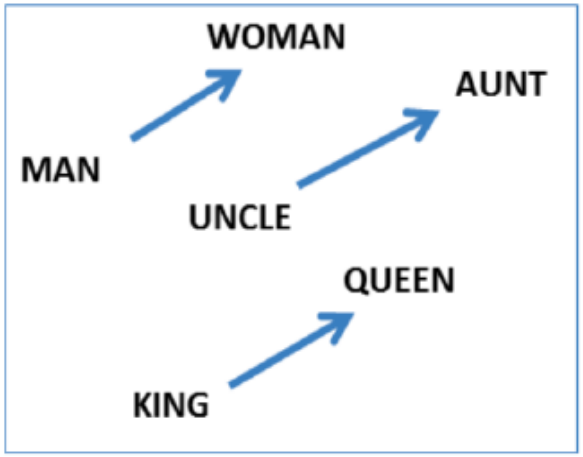
Vector de palabras *(word embedding)*

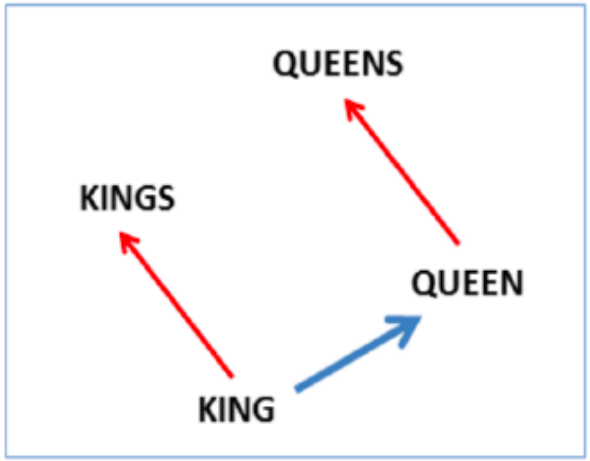
Los vectores son muy buenos para responder preguntas del tipo: A es a B como C es a…

La similaridad entre dos palabras viene en general dada por la distancia coseno entre los vectores.

Los resultados se prueban con listas de similaridad y relacionalidad elaboradas por humanos.

Ejemplo: hombre es a mujer como tío es a…

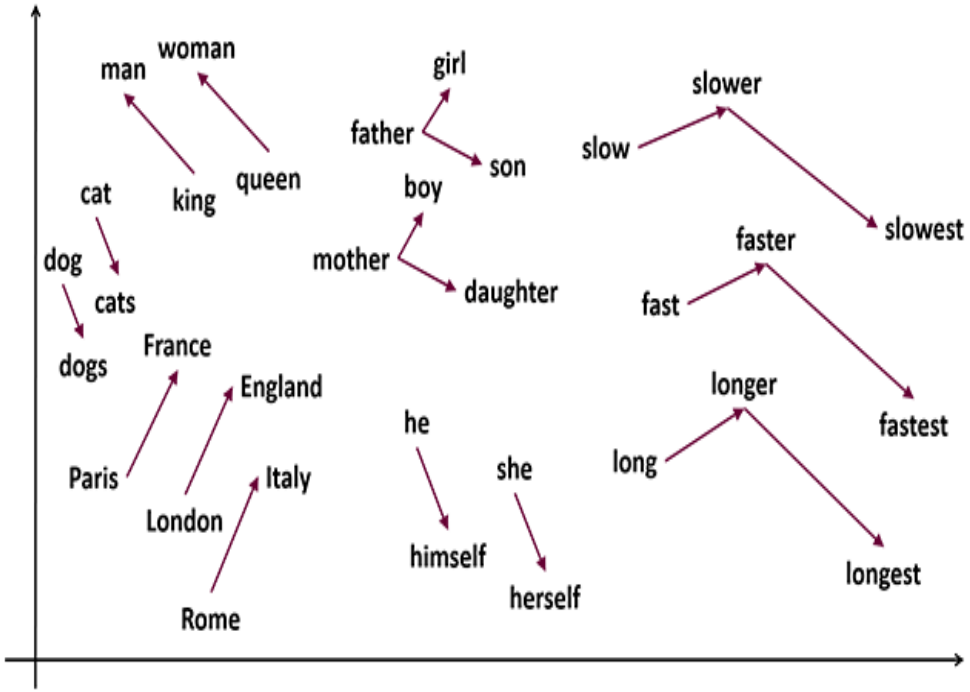




Vector de palabras *(word embedding)*

Las técnicas de *word embedding* han demostrado ser muy útiles en múltiples tareas del PLN, aparte de la similitud de textos, y en la actualidad gozan de gran popularidad.

- ✓ Sistemas de traducción automática neuronal
- ✓ Clasificación de textos
- ✓ Recuperación de información
- ✓ Sistemas de preguntas – respuestas



Métodos de generación de *word embeddings*

- ✓ **Word2Vec**
  https://code.google.com/archive/p/word2vec/
- ✓ **FastText**
  https://fasttext.cc/docs/en/english-vectors.html
- ✓ **GLOVE**
  https://nlp.stanford.edu/projects/glove/
- ✓ **LEXVEC**
  https://github.com/alexandres/lexvec
- ✓ **jointcHYB**
  http://ixa2.si.ehu.eus/ukb/bilingual_embeddings.html
- ✓ **PDC y HDC**
  http://ofey.me/projects/wordrep/
- ✓ **SketchEngine**
  https://embeddings.sketchengine.eu/static/index.html
- ✓ **UKB**
  http://ixa2.si.ehu.eus/ukb/
- ✓ **Numberbatch**
  https://github.com/commonsense/conceptnet-numberbatch
- ✓ **Context2Vec**
  https://u.cs.biu.ac.il/~nlp/resources/downloads/context2vec/

Word2Vec

- Modelo predictivo de generación de *word embeddings*.
- Implementa dos modelos neuronales: CBOW y Skip-gram.
  - CBOW: dado el contexto de la palabra objetivo, intenta predecirla.
  - Skip-gram: dada la palabra intenta predecir el contexto.
- Las capas internas de la red neuronal codifican la representación de la palabra objetivo, es decir, los *word embeddings*.

Vectores entrenados en el dataset Google News, disponible en: https://code.google.com/archive/p/word2vec/

- Cuenta con alrededor de 100 mil millones de palabras y contiene vectores de 300 dimensiones para 3 millones de palabras y frases.

FastText

- Es una extensión del modelo Word2Vec.
- Cada palabra es tratada como la suma de sus composiciones de caracteres llamados *ngrams*. El vector para una palabra está compuesto por la suma de sus ngrams.

Por ejemplo:

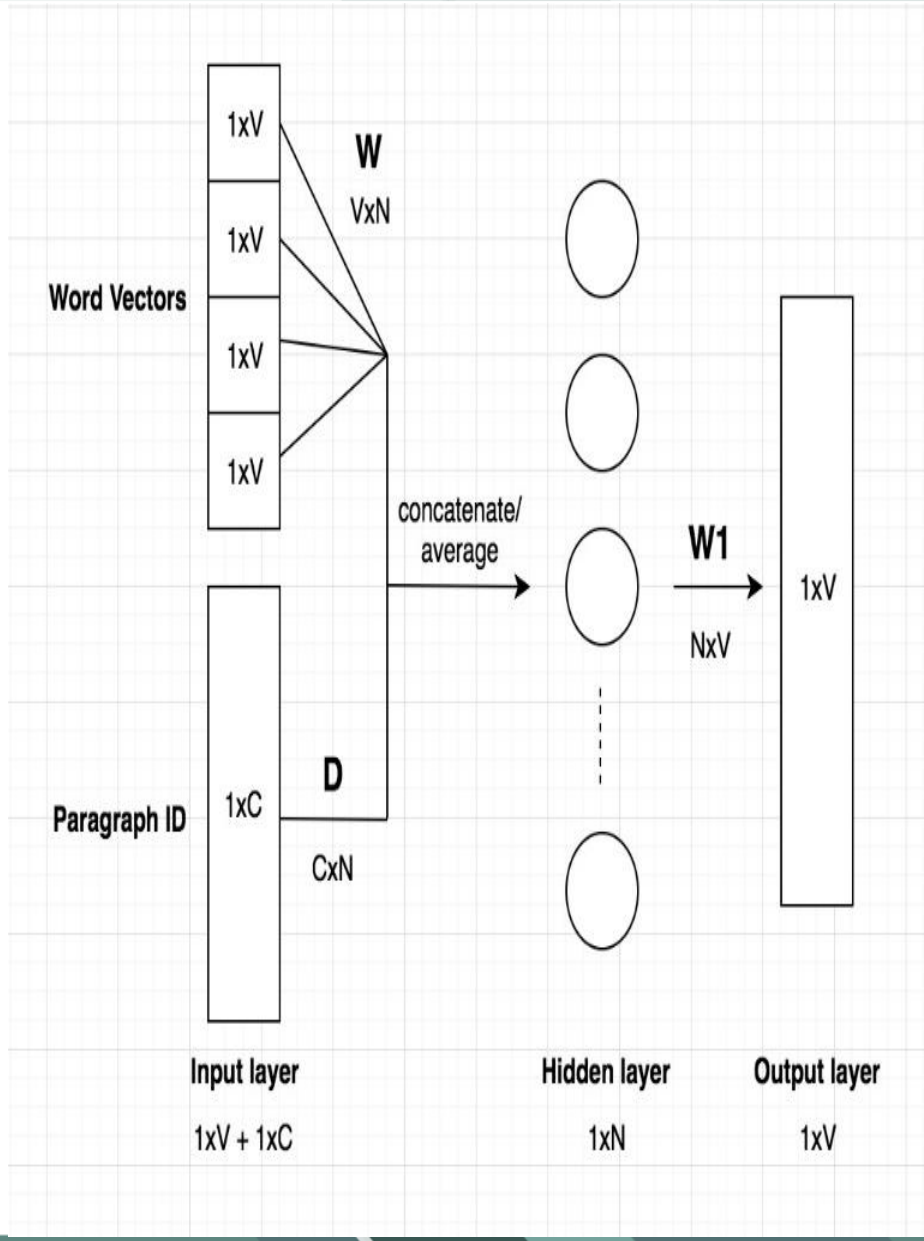
"apple" → "<ap, app, appl, apple, apple>, ppl, pple, pple>, ple, ple>, le>"

Vectores disponibles en: https://fasttext.cc/docs/en/english-vectors.html

- FastText (corpus *common crawl*) 600 mil millones de tokens; 300 dimensiones para 2 millones de palabras.
- FastText (corpus wikipedia, UMBC y statmt.org) 300 dimensiones para 1 millón de palabras; 16 mil millones de tokens.

Vector de documentos

El concepto de Doc2Vec es bastante simple, si ya está familiarizado con el modelo de Word2vec. El modelo Doc2vec se basa en Word2Vec, con solo agregar otro vector (ID de párrafo) a la entrada. La arquitectura del modelo Doc2Vec se muestra en la figura.

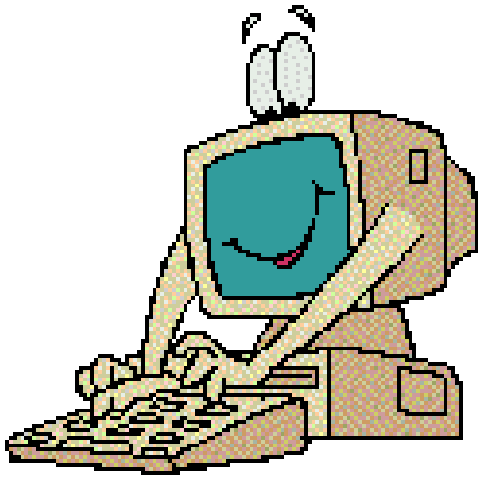


Vector de documentos

- El modelo anterior se llama ***Distributed Memory version of Paragraph Vector*** (PV-DM).

- Otro algoritmo de Doc2Vec que se basa en Skip-Gram, se llama ***Distributed Bag of Words version of de Paragraph Vector*** (PV-DBOW).

Práctica



Ejercicio5(es)-Modelo de Espacio Vectorial.ipynb

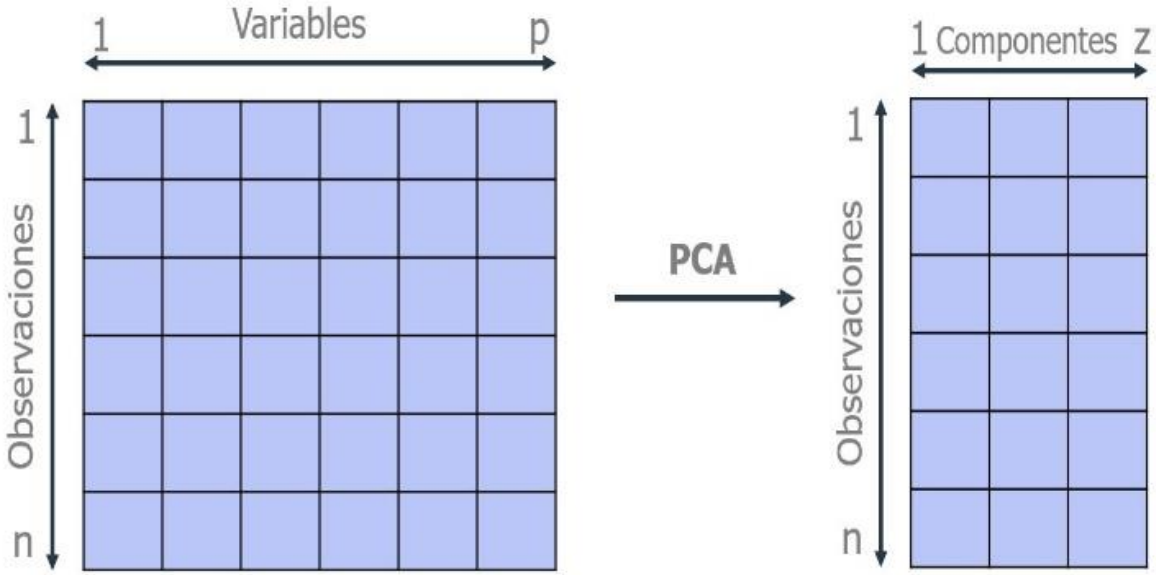
PCA (Análisis de Componentes Principales)

El análisis de componentes principales (*Principal Component Analysis* PCA) es un método de reducción de dimensionalidad que permite simplificar la complejidad de espacios con múltiples dimensiones a la vez que conserva su información.

n individuos

p variables $(X_1, X_2, \ldots, X_p)$

Espacio muestral de p dimensiones



número de factores subyacentes $(z < p)$ que explican aproximadamente lo mismo que las p variables originales

PCA (Análisis de Componentes Principales)

- El método permite "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Sin olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes.
- Dos de las principales aplicaciones son la visualización y el preprocesado de predictores previo ajuste de modelos supervisados.
- La librería *scikitlearn* contiene la clase *sklearn.decomposition.PCA.*

PCA (Análisis de Componentes Principales)

Con PCA se obtiene:

1. Una medida de como cada variable se asocia con las otras (matriz de covarianza)
2. La dirección en las que nuestros datos están dispersos (autovectores)
3. La relativa importancia de esas distintas direcciones (autovalores)

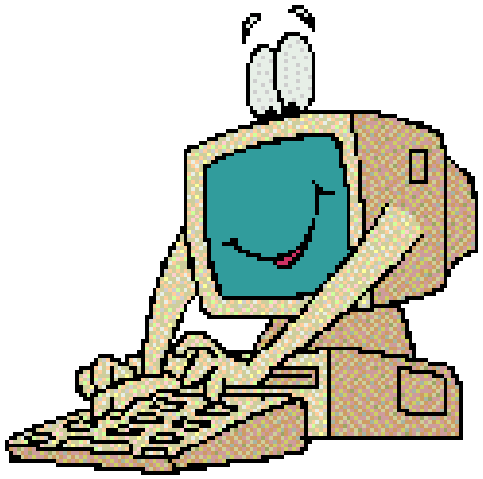
PCA combina los predictores y permite deshacernos de los autovectores de menor importancia relativa.

PCA (Análisis de Componentes Principales)

- Como contras, el algoritmo es muy influenciado por los *outliers* en los datos. Por esta razón, surgieron variantes de PCA para minimizar esta debilidad. Entre otros se encuentran: *RandomizedPCA*, *SparcePCA* y *KernelPCA*.

- Por último, citar que PCA fue creado en 1933 y ha surgido una buena alternativa en 2008 llamada *t-SNE* con un enfoque distinto.

Práctica



Ejercicio5(es)-Modelo de Espacio Vectorial.ipynb

Conclusiones

- Muchas de las tareas de recuperación de información como la búsqueda, agrupamiento o categorización de textos tienen como primer objetivo procesar documentos en lenguaje natural. El problema que surge es que los algoritmos que pretenden resolver estas tareas, necesitan representaciones internas explícitas de los documentos. En el área de recuperación de información normalmente se usa una expresión vectorial, donde las dimensiones del vector representan términos, frases o conceptos que aparecen en el documento. En este aspecto la representación más adoptada es la conocida como bolsa de palabras.

Conclusiones

- Si bien el rendimiento de un sistema de recuperación de información depende en gran medida de las medidas de similitud entre documentos, la ponderación de términos desempeña un papel fundamental para que esa similitud entre documentos sea más confiable. Así, por ejemplo, mientras que una representación de documentos basada solo en las frecuencias o apariciones de términos, no es capaz de representar adecuadamente el contenido semántico de los documentos, la representación de términos ponderados (Aplicación de métodos de normalización a la matriz documento-término) hace frente a errores o incertidumbres asociadas a la representación simple de documentos.

Conclusiones

- El modelo de espacio vectorial tiene las siguientes limitaciones:

1. Los documentos largos quedan poco representados ya que contienen pocos valores en común (un producto escalar menor y una gran dimensionalidad)

2. Las palabras de búsqueda deben coincidir con las palabras del documento; partes de una palabra pueden dar en falsos positivos.

3. Sensibilidad semántica, documentos con contextos similares, pero con diferente vocabulario no serán asociados, resultando en falsos negativos.

Lecturas

- García Ferrero, I. (2018). Estudio de word embeddings y métodos de generación de meta embeddings. Obtenido de https://addi.ehu.es/bitstream/handle/10810/29088/MemoriaTFG_IkerGarciaFerrero.pdf?isAllowed=y&sequence=3
- Justicia de la T., M. d. (2017). Nuevas Técnicas de Minería de Textos: Aplicaciones. Universidad de Granada. Tesis Doctorales. Obtenido de http://hdl.handle.net/10481/46975
- López Solaz, T., Troyano Jiménez, J. A., Ortega Rodríguez, F. J., & Enríquez de Salamanca Ros, F. (2016). Una aproximación al uso de word embeddings en una tarea de similitud de textos en español. Obtenido de http://rua.ua.es/dspace/handle/10045/57753

Contacto

Luis Enrique Argota Vega

Máster en Ciencia e Ingeniería de la Computación

luiso91@gmx.com

Tels: 5578050838

Redes sociales:

LinkedIn https://cutt.ly/ifPyTEH

Google Scholar https://cutt.ly/WfPtYZz