DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos

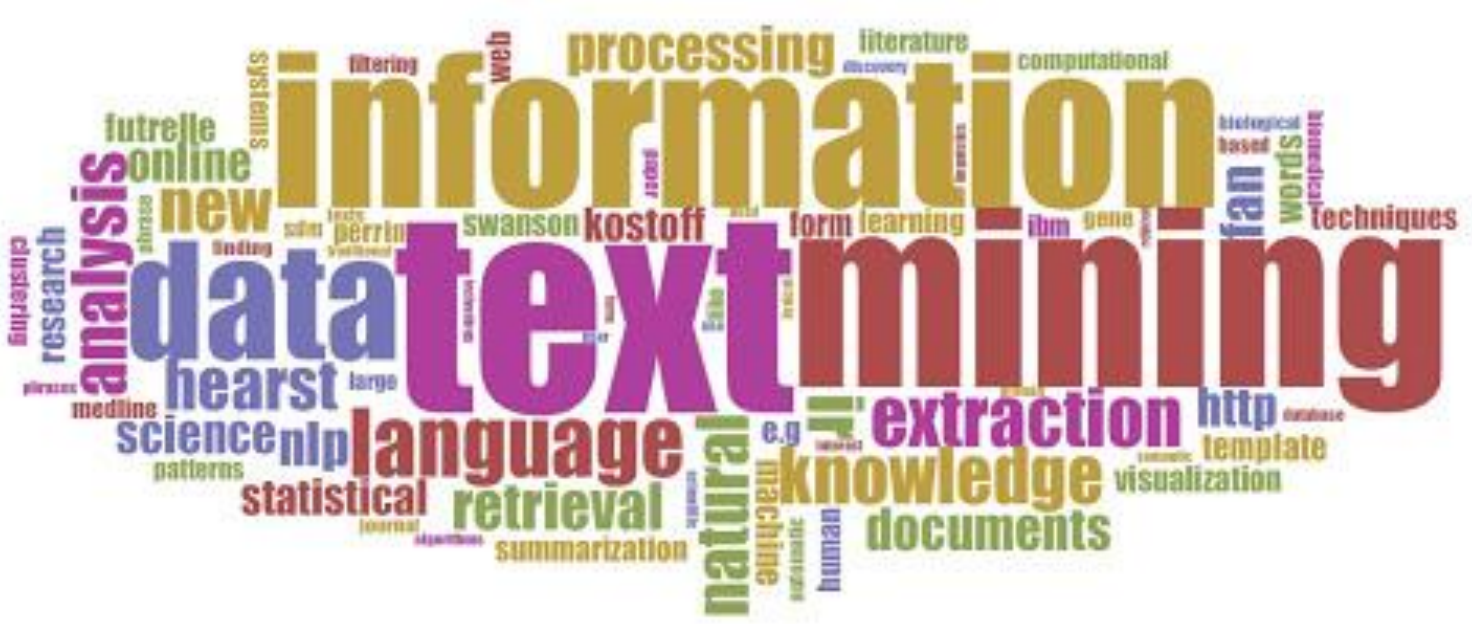
Universidad Nacional de Colombia

Contenido

- Introducción
- Algunas aplicaciones
- Pre-procesamiento y limpieza
- Representaciones de conteos
- Embedding
- Modelamiento

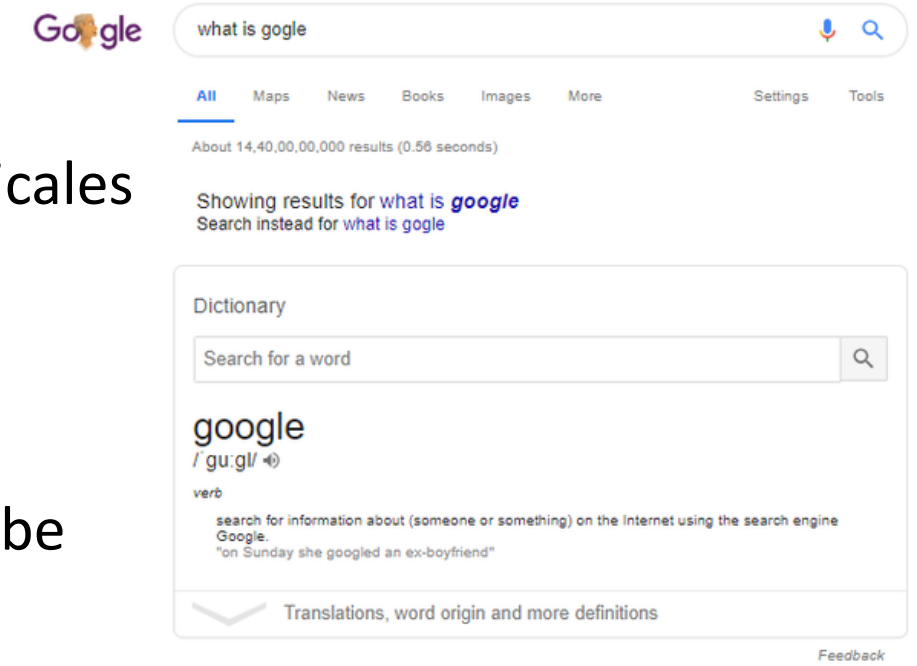
Hay muchas fuentes de texto

- Reseñas de productos, películas, videos
- Posts de Facebook
- Tweets
- Recomendaciones
- Declaraciones
- Discursos
- Leyes
- Emails
- …



NLP – Natural Language Processing - Procesamiento de lenguaje natural

Algunas de las aplicaciones de NLP

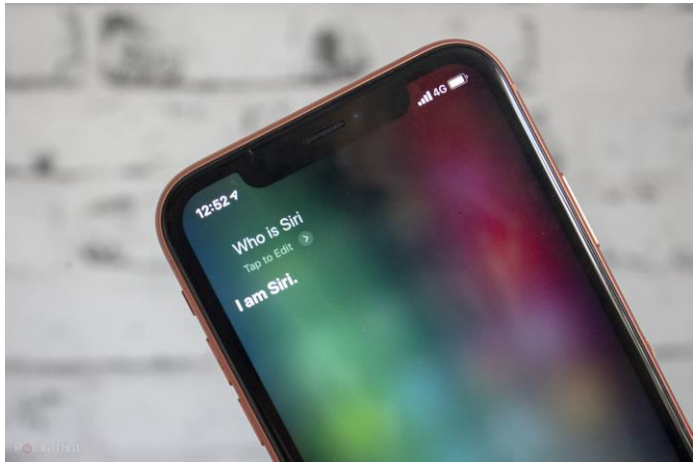


- Muy buenos resultados a búsquedas con errores gramaticales

- Traducciones automáticas de texto
https://www.blog.google/products/translate/higher-quality-neural-translations-bunch-more-languages/

- Generación automática de subtítulos en vídeos de YouTube
https://ai.googleblog.com/2009/12/automatic-captioning-in-youtube.html

- Siri



NLP tiene un rango tan diverso de aplicaciones que es difícil tener una definición para este

Los computadores utilizan conceptos matemáticos y lenguajes de programación

Los seres humanos utilizamos “lenguajes naturales”: Español, Inglés, Alemán, Francés, etc

Cuando intentamos hacer un puente entre el lenguaje de computadores y el de los seres humanos, hablamos de NLP

NLP tiene un rango tan diverso de aplicaciones que es difícil tener una definición para este

Para ayudar a los computadores a “entender” como las personas hablamos (NLP, procesamiento de audio)

Traducir automáticamente entre distintos lenguajes (Google Translate)

Poner etiquetas o categorías a textos (detección de correo spam)

Llevar audio a texto (YouTube captions)

NLP tiene un rango tan diverso de aplicaciones que es difícil tener una definición para este - Aplicaciones

Categorización de documentos

Análisis de sentimientos/opiniones

Análisis de historias médicas, documentos legales, aplicaciones laborales

Identificación de autores

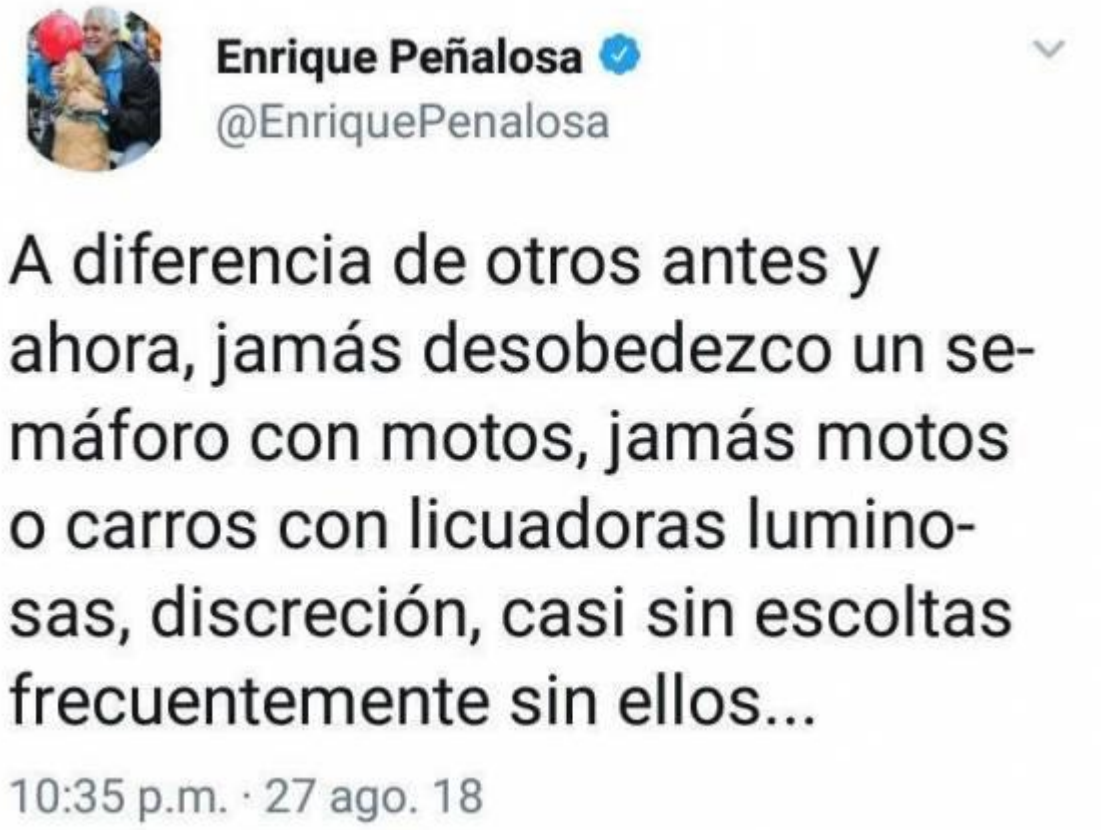
Clasificación automática de incidencias en centros de atención al cliente

Filtrado de documentos

Creación automática de resúmenes

Análisis de discurso

¿Cómo trabajar con textos?



Pero… los textos no tienen la estructura que hemos trabajado hasta el momento

No-estructurada o semi-estructurada

La información no se representa en variables-vectores

Tiene estructura lingüística: el lenguaje, relación entre palabras, importancia de palabras, negaciones

El texto requiere de mucha limpieza: gramática, ortografía, abreviaciones,…

El texto es comunicación entre personas

¿Cómo trabajar con textos?

Alta dimensionalidad

Cien años de soledad: 7 generaciones de personajes.
¿Cómo emplear los métodos de machine learning para analizar estos textos?

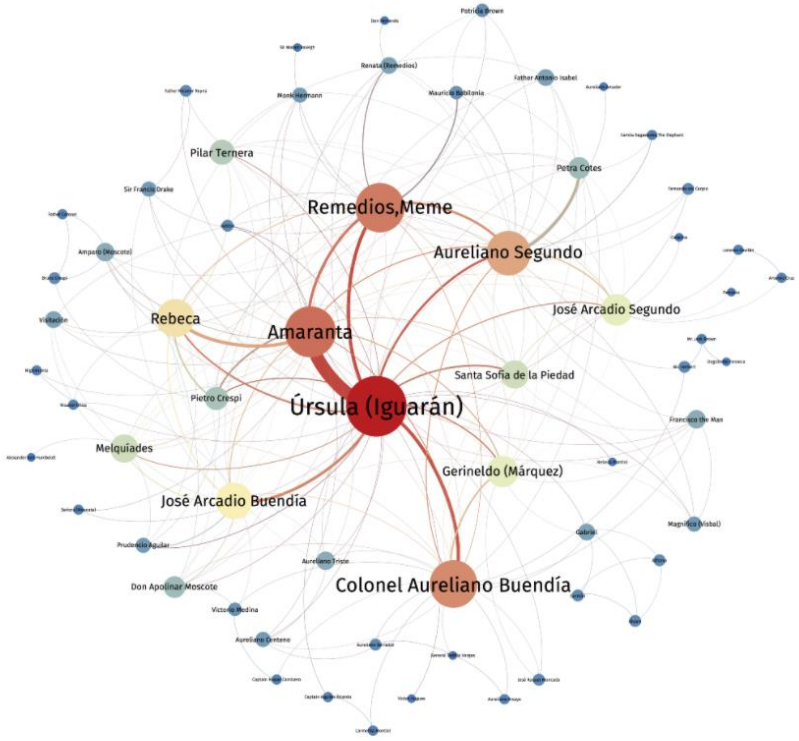
ML tienen un pésimo desempeño en estos espacios de tan alta dimensionalidad

Los enfoques usuales de ML no tienen en cuenta reglas importantes del lenguaje

Bus – Autobús

Se refieren a los mismo, pero se escriben diferente

NLP no es ML “usual” aplicado a textos.



https://medium.com/@finalfire/one-hundred-years-of-solitude-how-i-analyzed-my-favorite-book-6c20456480c8

¿Cómo trabajar con textos?

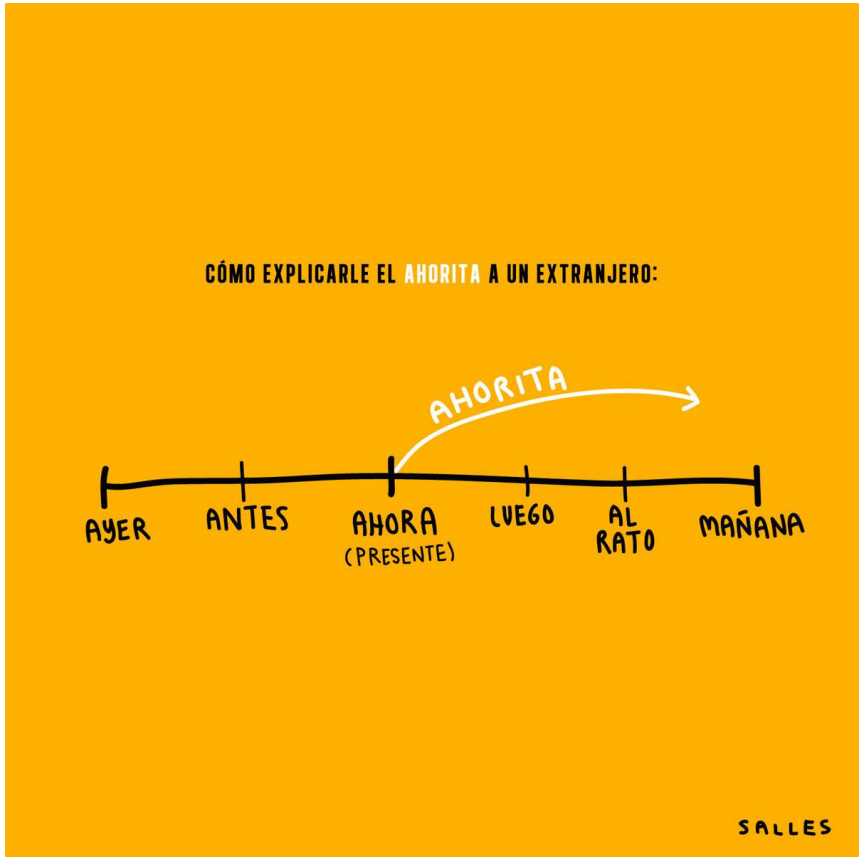
El Texto necesita de contexto

Una palabra puede tener significados diferentes dependiendo el contexto

Vino (bebida o verbo venir)

Ahorita

Rapar
Colombia: arrebatar, corte de pelo
Rep. Dominicana: Relación fuera del matrimonio



¿Cómo trabajar con textos?

¿Cómo trabajar con textos?

Pre procesamiento y estandarización

- Corregir errores sencillos. Formato, encoding

- Estandarización de errores de tipeo/ortográficos.
  Reemplazar xq con porque

- Transformación: capitalización, tokenización, lematización

- Crear variables, Ingeniería de características (feature engineering)
  Etiquetas POS (part of speech: verbo, adjetivo)
  Matrices de *documento x término*

Capitalización

Llevar todas las palabras a minúsculas o todas a mayúsculas

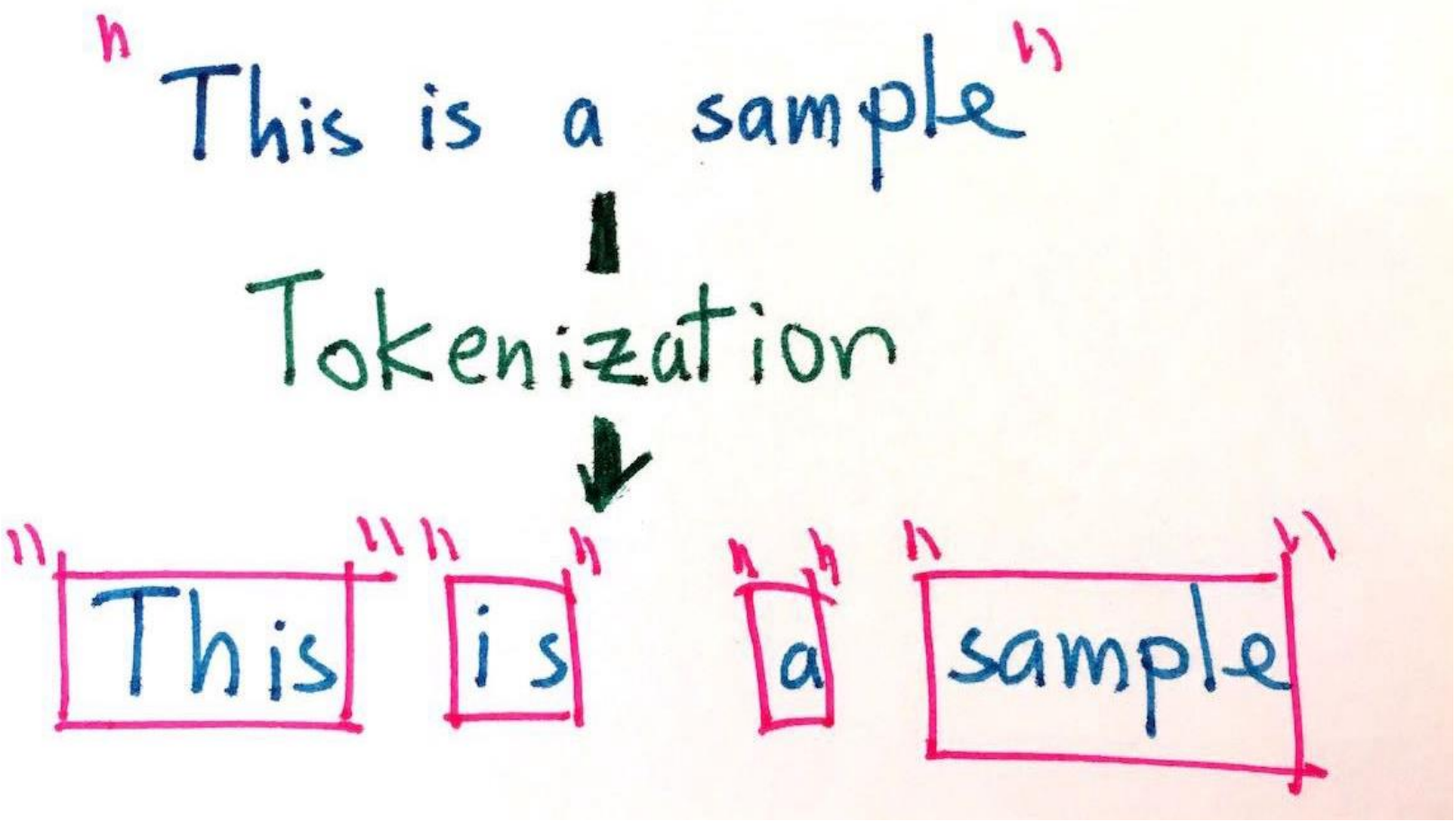
Ellos viajaban a Colombia

ellos viajaban a colombia

Tokenización

La data estructurada se compone de variables, los textos se componen de oraciones.

El primer paso es hacer **tokenización** de los textos: dividir un texto en sus oraciones, dividir las oraciones en palabras.



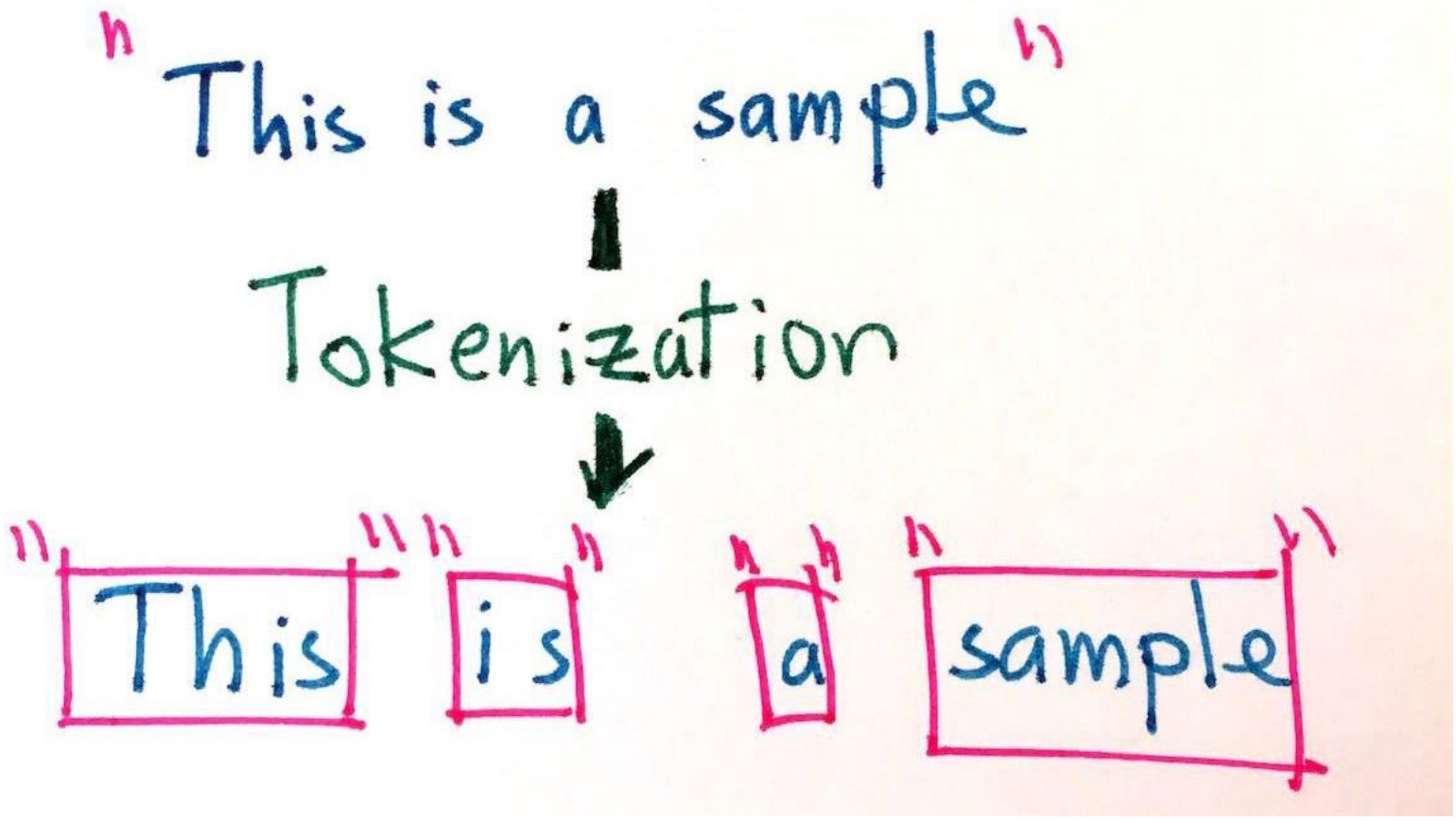
Tokenización

Token: sería la unidad de observación

¿Cómo dividirlos?

Las oraciones terminan con puntos (no siempre es el caso: tweets)

Un espacio o salto de línea indica la diferencia entre palabras



Tokenización

Token: sería la unidad de observación

¿Cómo dividirlos?

Las oraciones terminan con puntos (no siempre es el caso: tweets)

Un espacio o salto de línea indica la diferencia entre palabras

Ellos viajaban a Colombia

Ellos *viajaban* *a* *Colombia*

Lematización

Sacar la raíz (lema) de las palabras

Leíamos → leer

Así asociamos con un mismo lema diferentes palabras

Tenemos mayor flexibilidad en el análisis

Trabajó
Trabaja
Trabajando
Trabajamos

Trabajar

Lematización

Sacar la raíz (lema) de las palabras

Así asociamos con un mismo lema diferentes palabras

Ellos viajaban a Colombia

*Ellos **viajar** a Colombia*

Truncación (steamming)

Corta o trunca las palabras en lo que considera la raíz de la palabra, así se identifican palabras con la misma raíz.

Quita partes del inicio o final de las palabras, teniendo en cuenta listas de prefijos y sufijos

Es similar a la lematización, pero los resultados finales no necesariamente son palabras

Amamos → amam

La lematización si hace uso del vocabulario y el análisis morfológico del lenguaje (enfoque lingüístico)

Palabra original	**Lematización**	**Truncación**
Ver	Ve	V

Remover las stopwords

Las stop-words son aquellas palabras que son demasiado comunes en el lenguaje

Están definidas para cada idioma

Español: a, acá, ahí, ante, aquel, como, con, cual, cuando, de, del, el, en, es, eres, ha, ir, me, ni, no, se, un, …

Inglés: the, a, me, my, it, an, and, at, by, then, no, so, …

Son palabras que no aportan de más al significado de los textos en si

Remover las stopwords

Remover palabras que no aportan de más al significado de los textos en si

El niño comía un helado de chocolate mientras esperaba a su mamá

~~El~~ niño comía ~~un~~ helado ~~de~~ chocolate mientras esperaba ~~a~~ ~~su~~ mamá

Niño comía helado chocolate mientras esperaba mamá

Expresiones regulares

Buscar patrones de caracteres en nuestro texto para reemplazarlo o eliminarlo

- links : ‘https://…’
- @, #
- palabras o expresiones específicas (buscar todos los xq en un texto y transformarlo con porque)

En ocasiones pueden ser útiles para hacer tareas demandantes en texto rápidamente

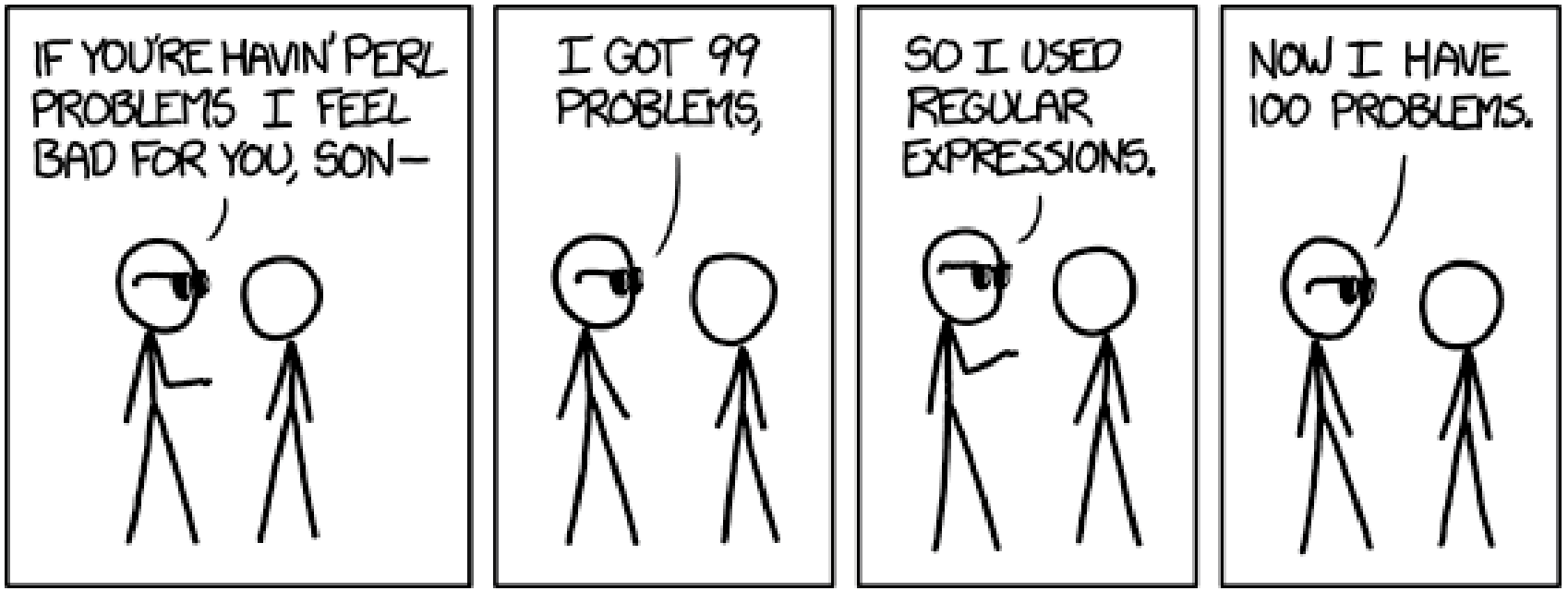
Expresiones regulares

Curso rápido de expresiones regulares

1. . Encuentra cualquier carácter diferente del salto de línea \n
2. \d = [0-9] encuentra cualquier dígito
3. \D = [^0-9] encuentra todo carácter no dígito (^ es negación)
4. \w = [a-zA-Z0-9_] encuentra cualquier caracter alfanumérico
5. \W = [^a-zA-Z0-9_] encuentra cualquier carácter no alfanumérico
6. [a-d]+ encuentra {a,b,c,d} una o más veces
7. [a-d]{3} encuentra {a,b,c,d} exactamente 3 veces
8. [a-d]* encuentra {a,b,c,d} cero o más veces

Expresiones regulares

A veces también pueden ser problemáticas



Otras transformaciones

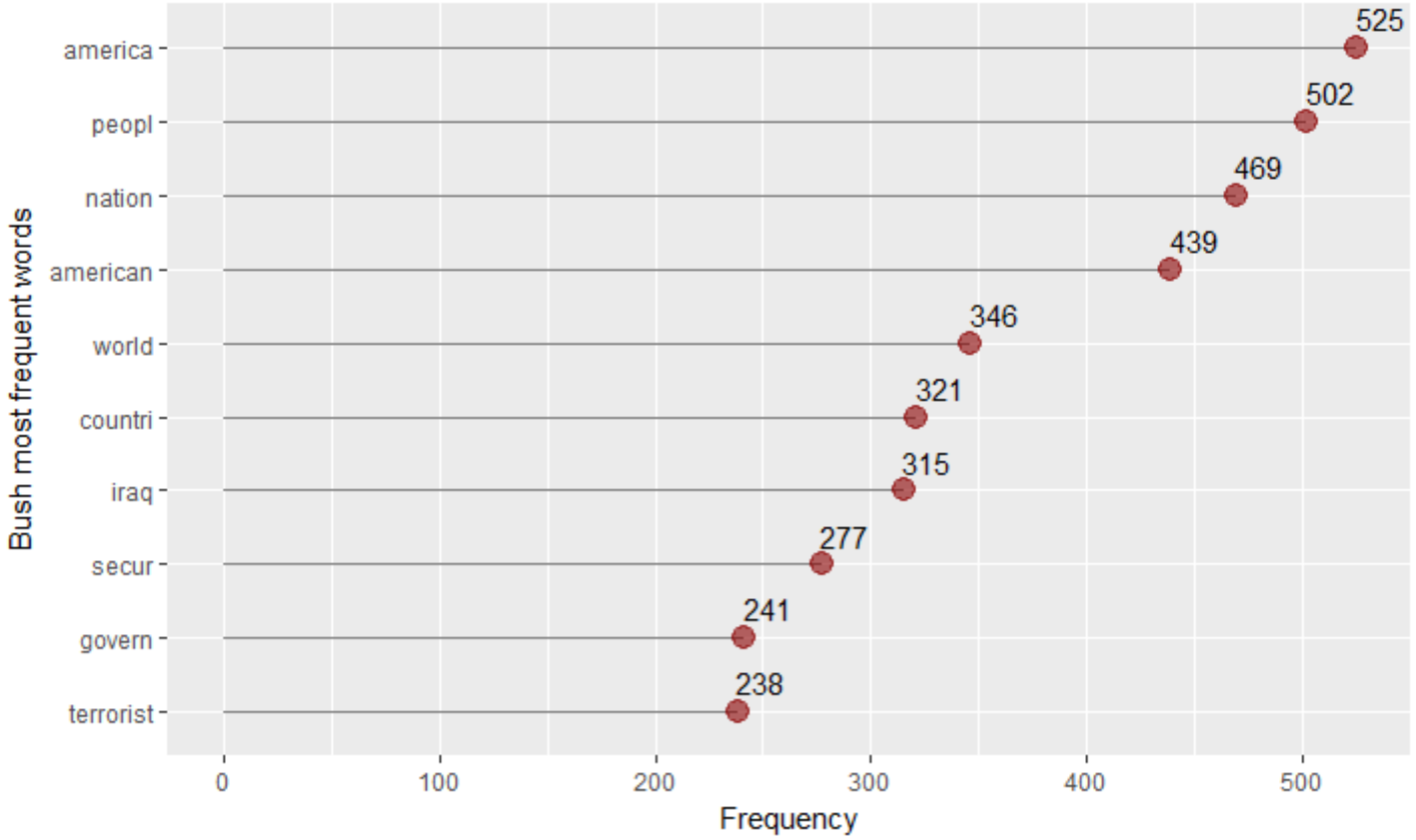
Remover signos de puntuación

Remover números

Remover caracteres o palabras que puedan generar ruido en el análisis

Representaciones de conteos

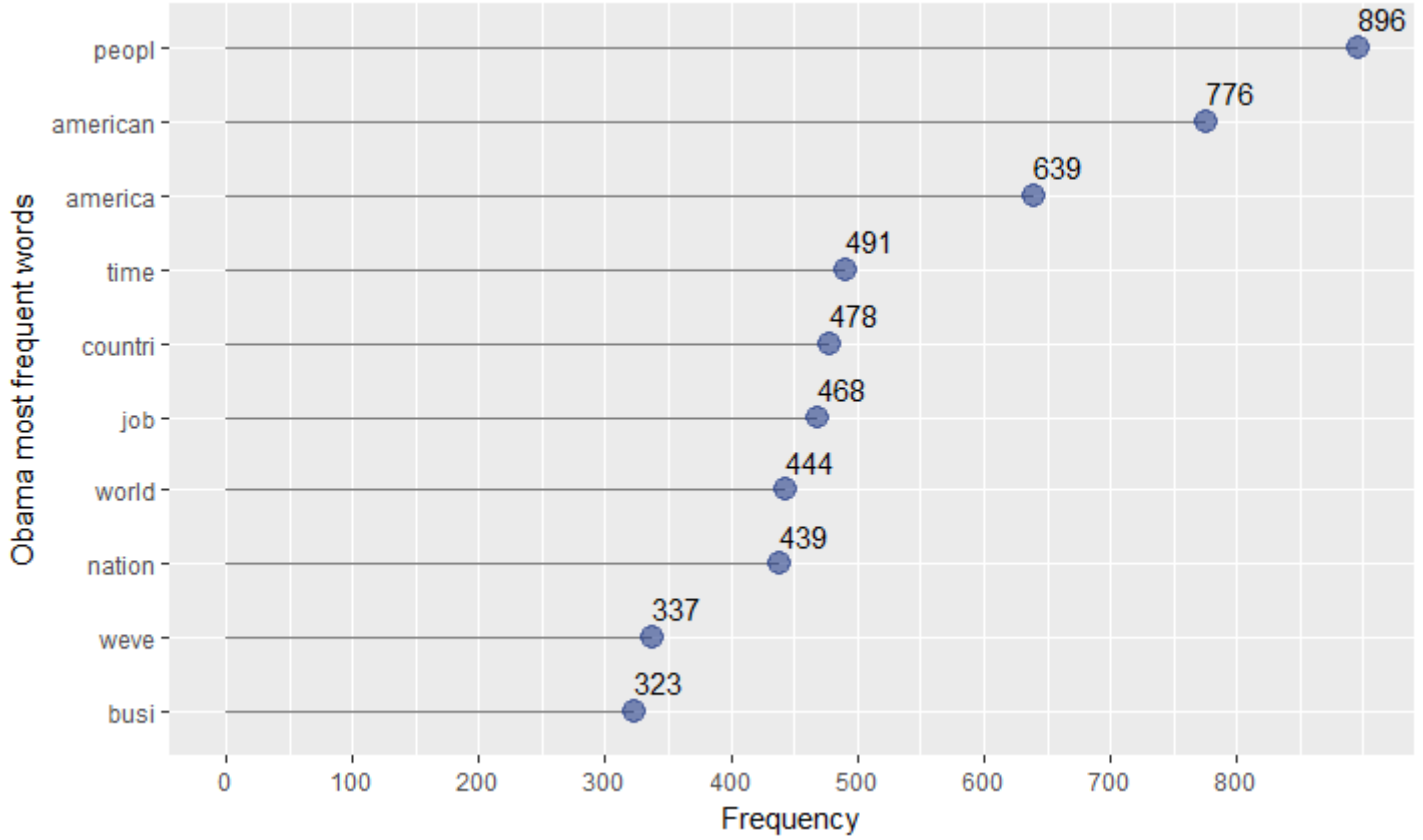
Palabras frecuentes
Identificar palabras que sean muy comunes o poco comunes



Se puede definir unos límites para las frecuencias (o fracciones) de documentos en los que una palabra ocurre

Representaciones de conteos

Palabras frecuentes
Identificar palabras que sean muy comunes o poco comunes



Se puede definir unos límites para las frecuencias (o fracciones) de documentos en los que una palabra ocurre

Representaciones de conteos

Las frecuencias deben normalizarse

Documentos de diferentes longitudes

Por el número de documentos revisados

- Dividir el conteo de una palabra en un documento por el número total de palabras en el mismo
- Teniendo el cuenta el total de veces que la palabra aparece en un documento relativo al total de documentos.
- Una palabra común en todos los documentos puede no ser tan relevante como una que aparezca mucho en solo algunos documentos

Representaciones de conteo

Term frequency (TF) – frecuencia de término

$$TF(t,d) = \frac{\#\ t\ aparece\ en\ d}{longitud\ de\ palabras\ en\ d}$$

Se calcula para cada término para cada documento

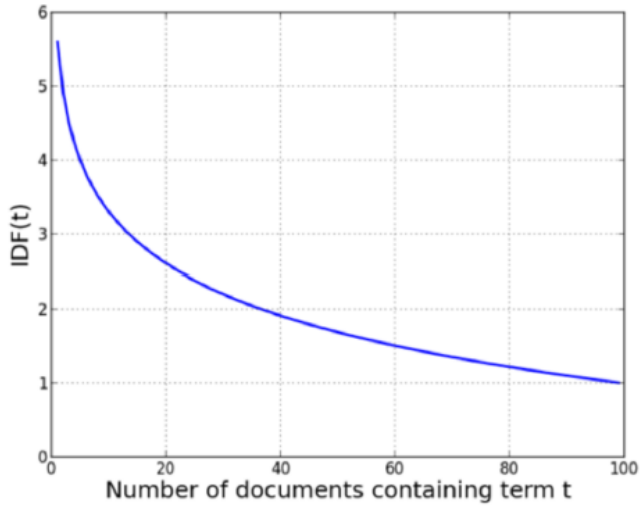
Podemos evaluar qué tan común es una palabra en un documento

Representaciones de conteo

Inverse document frequency (IDF) - Frecuencia de documento inversa

$$IDF(t) = 1 + \log\left(\frac{total\ documentos}{documentos\ que\ contienen\ t}\right)$$

Se calcula por cada término sobre todos los documentos (corpus). Entre más documentos contengan t, el IDF se va a cero



Representaciones de conteo

Combinarlas para tener un peso para cada término en cada documento

$$TFIDF(t,d) = TF(t,d) * IDF(t)$$

Prioriza términos que ocurren frecuentemente en un documento, pero penaliza aquellos que ocurren frecuentemente en todos los documentos

Medimos que tan ‘original’ o ‘interesante’ es una palabra considerando su frecuencia en un documento y en cuántos documentos está incluida

Se puede generar una matriz de vectores para cada documento

N-gramas

Conteos de secuencias ordenadas de n palabras que co-ocurren comúnmente

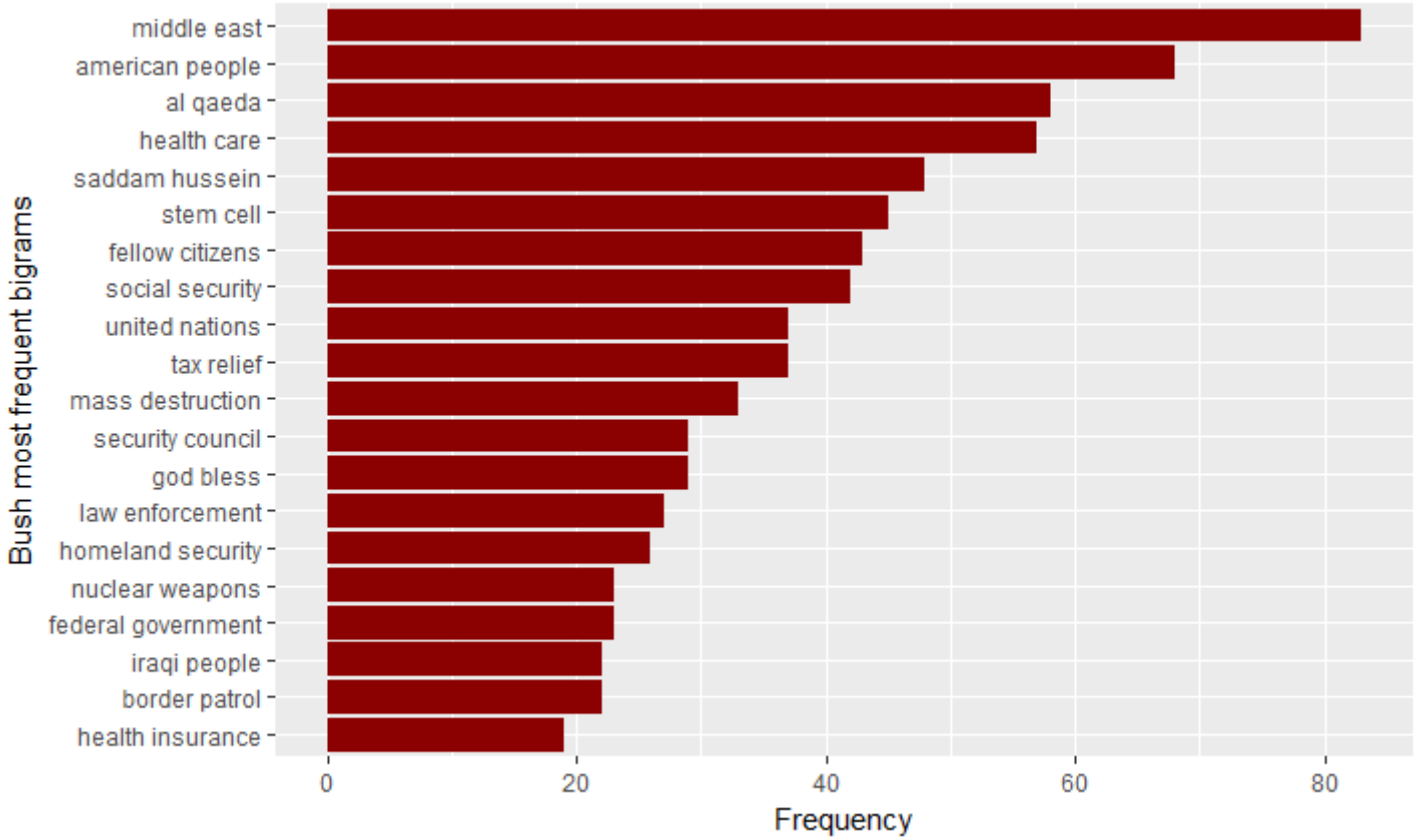
Ejemplo:

La rana verde saltaba alto

Se transforma en: {rana-verde, verde-saltaba, saltaba-alto}

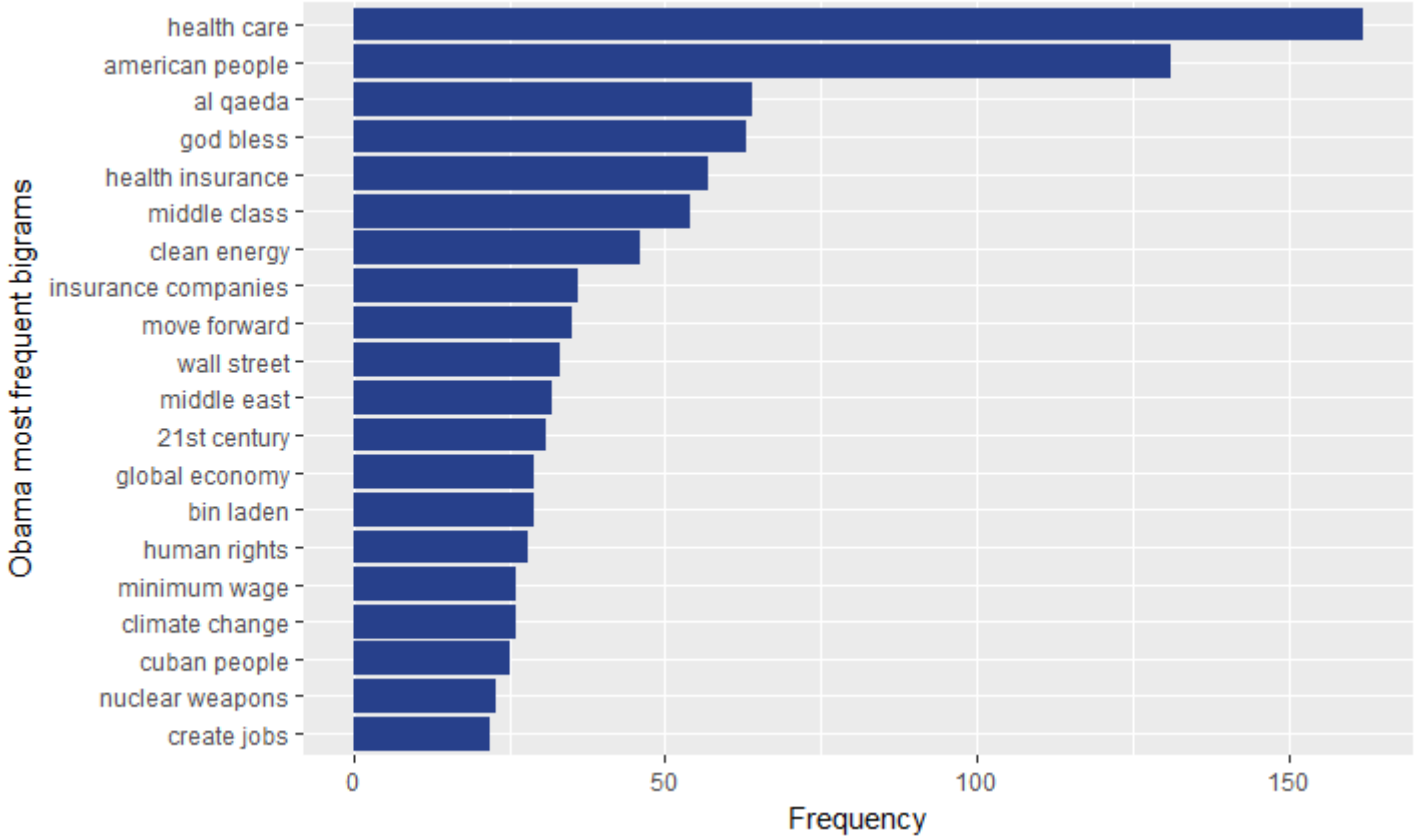
N-gramas

Conteos de secuencias de palabras adyacentes
Pares de palabras los llamamos bi-grams
Tríos de palabras los llamamos tri-grams



N-gramas

Conteos de secuencias de palabras adyacentes
Pares de palabras los llamamos bi-grams
Tríos de palabras los llamamos tri-grams



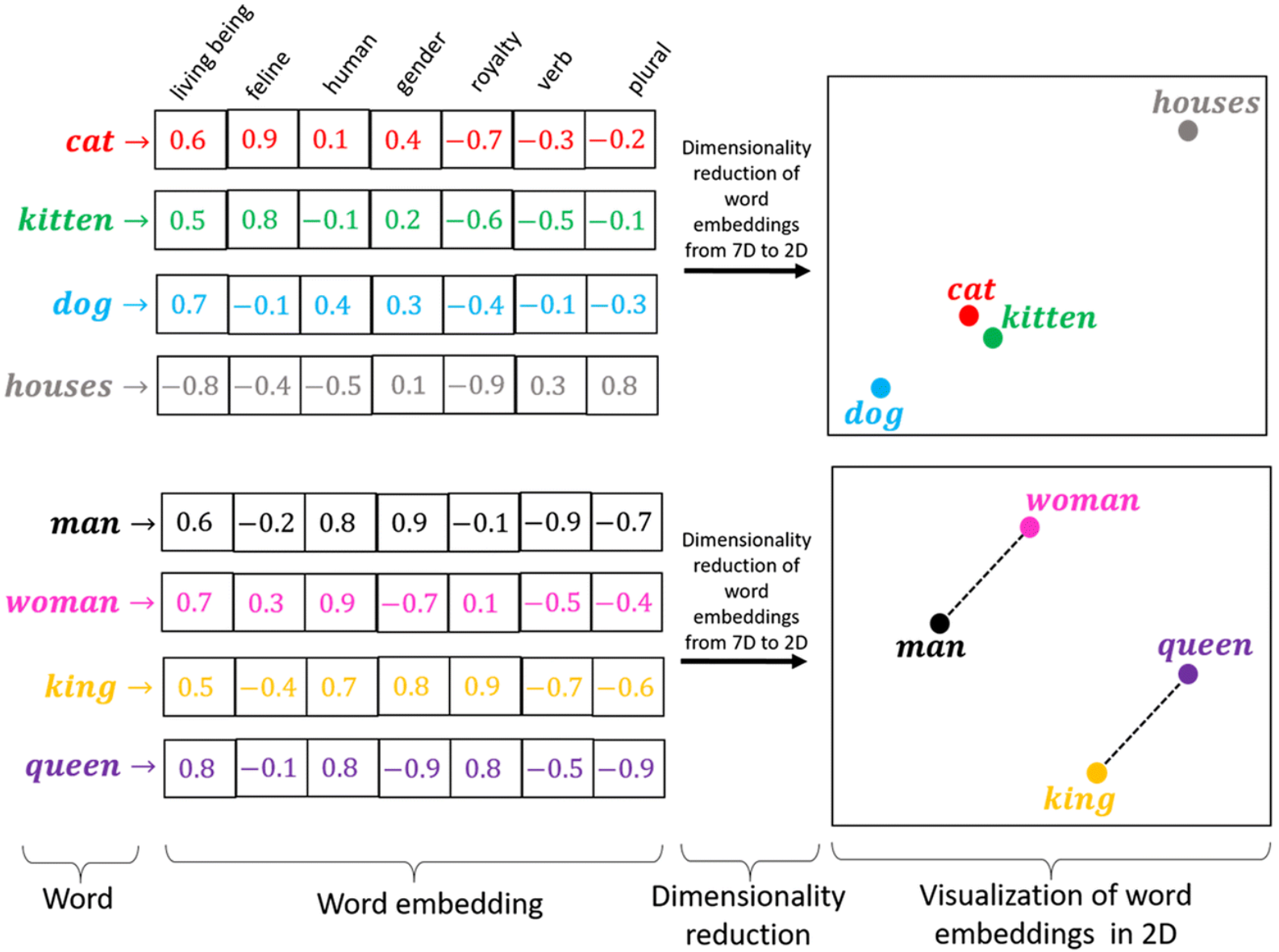
Bolsa de palabras

Representación vectorial de las palabras en los textos
"Transformar texto a número"

Se pueden definir métricas de distancia que muestren asociación entre términos

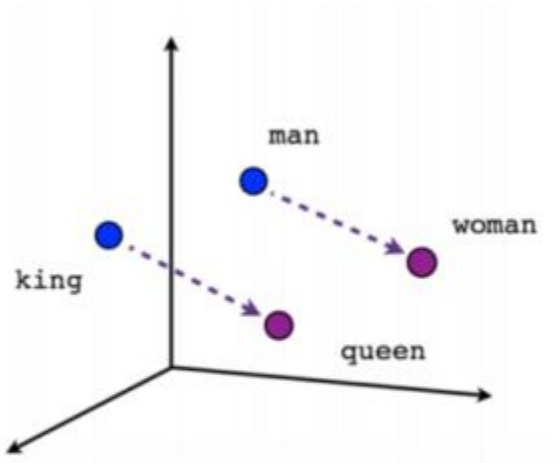
Word embedding

Representación vectorial de las palabras en los textos

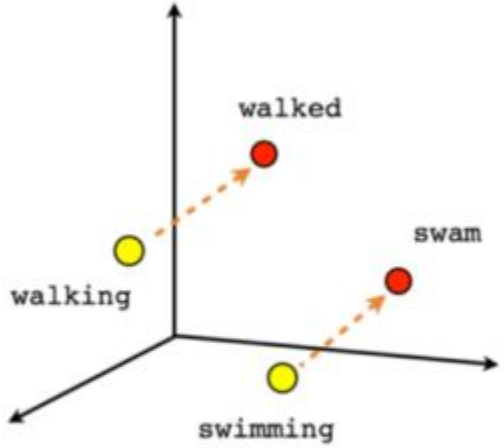


Embedding de palabras

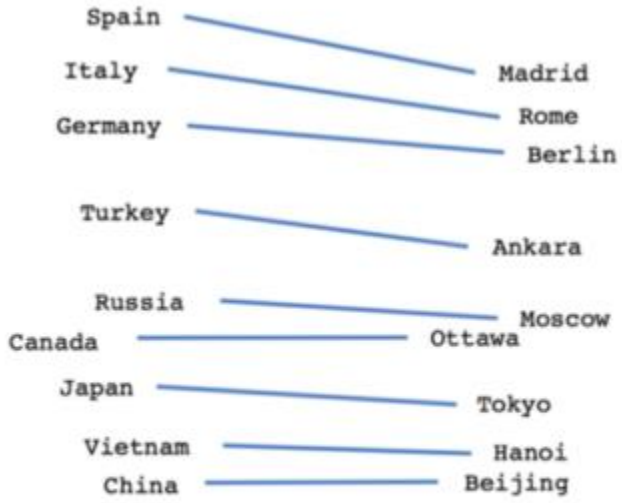
Hombre es a mujer como rey es a _____
China es a Beijing como Rusia es a _____



Male-Female



Verb tense



Country-Capital

https://embeddings.macheads101.com/

Embedding de palabras

Word2vec

Una palabra está correlacionada y definida por su contexto

Dos tipos de entrenamiento

CBOW: Predecir una palabra a partir del contexto

CSG: Enfocado en una palabra predecir el contexto

Embedding de palabras

Word2vec https://en.wikipedia.org/wiki/Word2vec

GloVe https://nlp.stanford.edu/projects/glove/

ELMo https://allennlp.org/elmo

BERT https://github.com/google-research/bert

Análisis de sentimiento

https://www.paralleldots.com/sentiment-analysis

Análisis de sentimiento

Estudio computacional de las opiniones/actitudes/emociones de las personas hacia entidades, individuos, aspectos, eventos o tópicos

Dos enfoques:

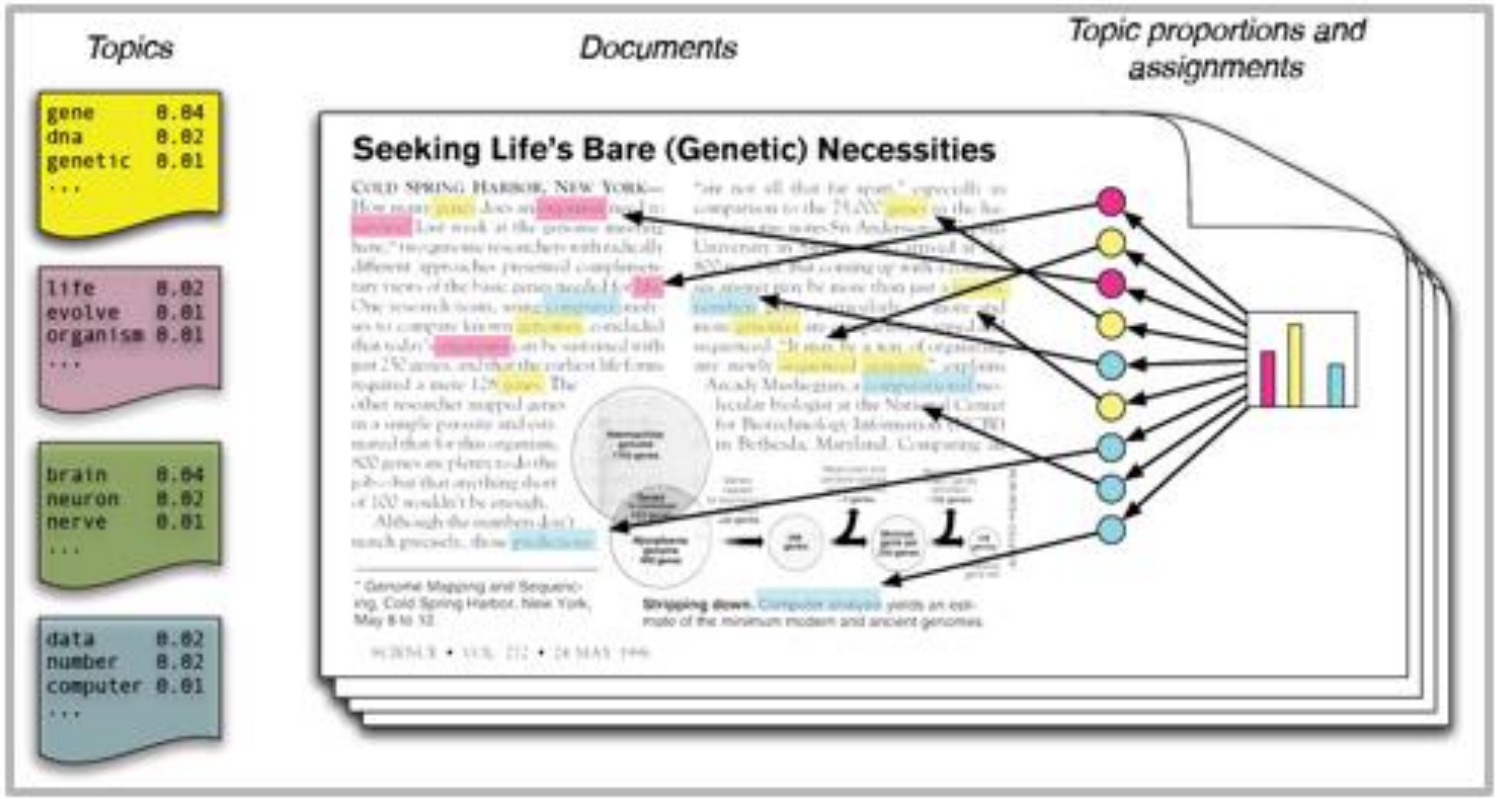
Semántico: utilizan diccionarios de términos con una orientación semántica que indica alguna polaridad u opinión

Felíz → Positivo
Mal → Negativo

Basado en aprendizaje: se construye un clasificador a partir de una colección de textos anotados utilizando técnicas de machine learning

Modelamiento de tópicos/temas

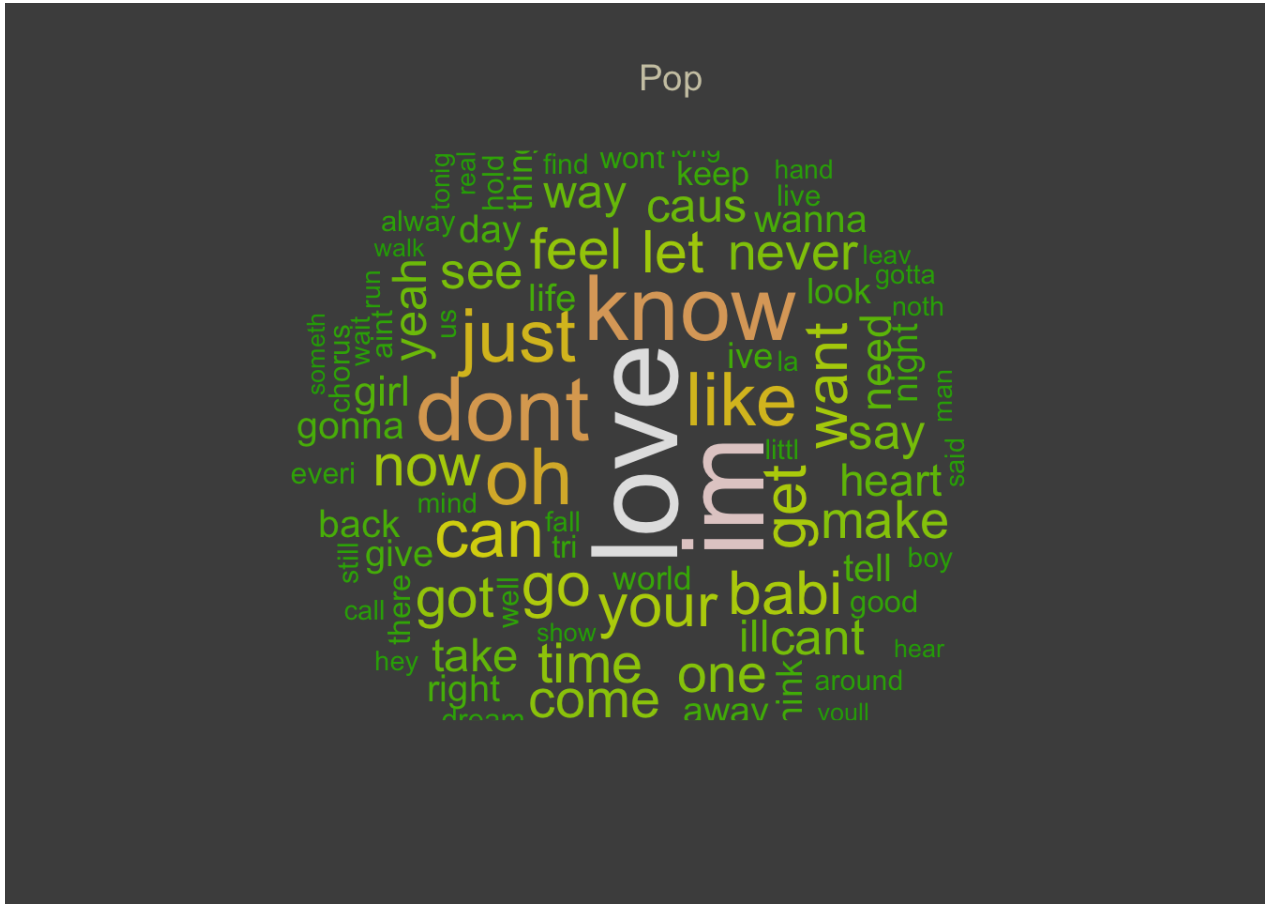
Identificación de los temas ocultos (latentes) que están presentes en una colección de documentos



https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

Modelamiento de tópicos/temas

El modelado de tópicos permite obtener los aspectos semánticos claves de los textos, mediante el descubrimiento de patrones de uso de palabras y cómo estos conectan los documentos que comparten patrones similares



https://towardsdatascience.com/text-analytics-topic-modelling-on-music-genres-song-lyrics-deb82c86caa2

Modelamiento de tópicos/temas

Los tópicos están representados por patrones recurrentes. Un tópico es un conjunto de palabras que tienden a aparecer juntas en los mismos contextos

- Conocer los trending topics en Twitter
- Conocer los tópicos de investigación en un área determinada (revisión de artículos)
- De reseñas qué podemos identificar como lo que gusta o no de un producto/servicio
- Tópicos principales en noticias, debates, entre otras

Conclusiones

- Existen múltiples aplicaciones de análisis de texto
- Los textos requieren de mucha limpieza y transformación para poder analizarlos
- Con exploración sencilla podemos extraer información valiosa
- Introducción a algunas técnicas de modelamiento de textos