
quantil

Introducción a Minería de Texto

Fabián Latorre
fabian.latorre@quantil.com.co

Contenido

- Qué es la minería de texto?
- Por qué es relevante? Por qué la estudiamos?
- Aplicaciones
- La complejidad del texto no estructurado
- Representación de texto como Bag of Words (Bolsa de Palabras)
- Modelo de espacio vectorial
  - Métodos de preprocesamiento
  - Evaluando la relevancia de ciertas palabras
  - Midiendo la similaridad entre los textos
- Aplicaciones en redes sociales
- Usando la API de twitter

Qué es minería de texto?

Es el uso de técnicas y métodos computacionales para extraer información de alta calidad sobre un texto.

Es un acercamiento computacional al descubrimiento de nueva información y conocimiento a partir de extracción automatizada de información de cantidades enormes de texto no estructurado.

Por qué es relevante?

- El texto no estructurado está presente en varias formas, y en cantidades enormes que crecen día a día:
  - Libros
  - reportes financieros
  - documentos administrativos
  - noticias y artículos periodísticos
  - blogs
  - wiki
  - mensajes en redes sociales
- Se estima que el 80% de todos los datos disponibles, son datos no estructurados.

Por qué es relevante?

- Para permitir el uso efectivo y eficiente de cantidades enormes de contenido textual, necesitamos métodos computacionales para:
  - extracción automatizada de información sobre texto no estructurado
  - análisis y resumen de la información extraída
- La investigación en minería de texto se centra en el desarrollo y mejora continua de estos métodos y sus aplicaciones.

Aplicaciones

- Clasificación de Documentos
- agrupamiento (clustering) y organización de documentos
- Resúmenes automáticos
- Visualización de Documentos
- Predicción de eventos (por ejemplo, predecir el comportamiento de la acción de alguna empresa, de acuerdo a noticias y comentarios en redes sociales)
- Sistemas de recomendación basados en contenido

La complejidad del texto no estructurado

- En general la interpretación / entendimiento de contenidos no estructurados (vídeos, imágenes, texto) usualmente es fácil para una persona, pero muy difícil para un programa de computador.
- En particular, la dificultad de entender automáticamente un texto están causadas por el hecho que el lenguaje humano / natural:
  - Está lleno de ambigüedades / sarcasmo
  - Usualmente depende de un contexto o de un conocimiento previo para definir un significado
  - Hay muchos términos probabilísticos (tonces, bn, quiubo, q tal, brexit)
  - Está influenciado por las relaciones interpersonales (el lenguaje usado depende de la persona)

Otros desafíos en minería de texto

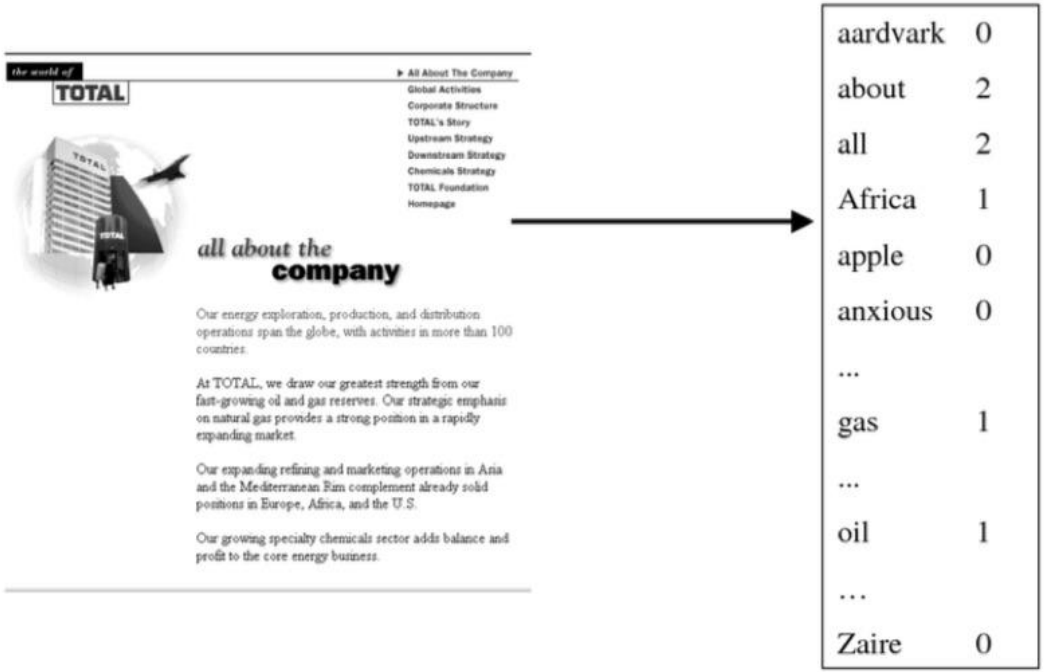
- El uso de métodos supervisados de aprendizaje es muy costoso
  - Se necesita una gran cantidad de documentos anotados (usualmente a través de trabajo humano) para poder entrenar un modelo
- Alta dimensionalidad en el espacio de atributos
  - Los atributos usualmente se definen como la presencia de términos o frases dentro de un documento
  - El vocabulario usado puede ser muy extenso

Primer paso: Representación Vectorial de un texto

Representación como Bag of Words (Bolsa de Palabras)

- Considera el texto simplemente como el conjunto de los términos que lo constituyen
- Basado en la siguientes suposiciones (muy debatibles)
  - Las palabras son mutuamente independientes (no hay relación entre sinónimos)
  - no importa el orden de las palabras
- A pesar de estas suposiciones y su simplicidad, da muy buenos resultados y puede ser usado como el modelo a mejorar

Representación de un texto como Bag of Words



Las distintas palabras usadas en el corpus forman un diccionario. Cada documento del corpus es descrito como el vector de frecuencias de los términos.

Vectorización de texto en general

- Generalización del modelo de Bag of Words
- Cada documento en el corpus es representado como un vector de múltiples dimensiones
  - Cada término representa una dimensión en el espacio vectorial
  - Un término puede ser una simple palabra o una frase compuesta

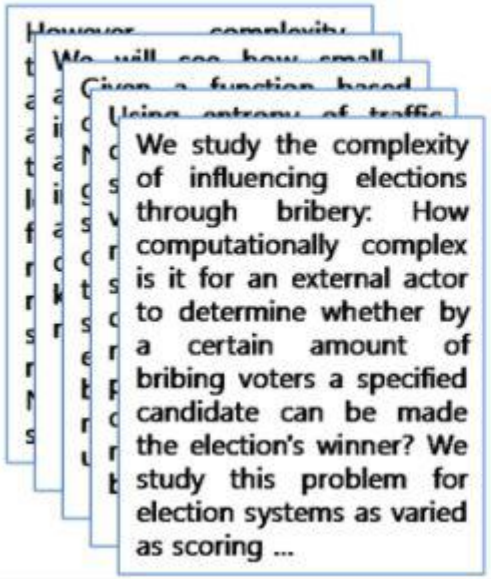
Representación vectorial de un documento

- Las entradas del vector representan pesos asociados con términos individuales. Esos pesos reflejan la relevancia de los términos correspondientes en el corpus
- Si un corpus tiene un diccionario de n términos, el documento d será representado con un vector (w_1, ..., w_n), donde w_i representa el peso del término i-ésimo.
- Las distancias en este espacio vectorial representan las relaciones entre los diferentes documentos del corpus
- Se asume que documentos que sean cercanos en esta distancia, también deben ser cercanos (similares) en su significado.

Matriz de Términos y Documentos

El corpus se representa como una matriz de términos y documentos (TDM), una matriz de dimensiones mxn con las siguientes características

- las filas (1, ..., m) representan los términos
- las columnas representan los documentos del corpus
- la entrada i,j contiene el peso del término i, en el documento j



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Preprocesamiento del texto

- Antes de crear la matriz, los documentos deben ser preprocesados
- Objetivo: Reducir el número de palabras a aquellas que uno espera, sean las más relevantes para el corpus dado
- El preprocesamiento usualmente incluye
  - Normalización del texto (por ejemplo, remover espacios innecesarios, pasar todo a minúsculas)
  - Remover términos con frecuencia muy alta o muy baja
  - Remover las llamadas ‘stop-words’
  - Reducir las palabras a su raíz mediante ‘stemming’ o lematización

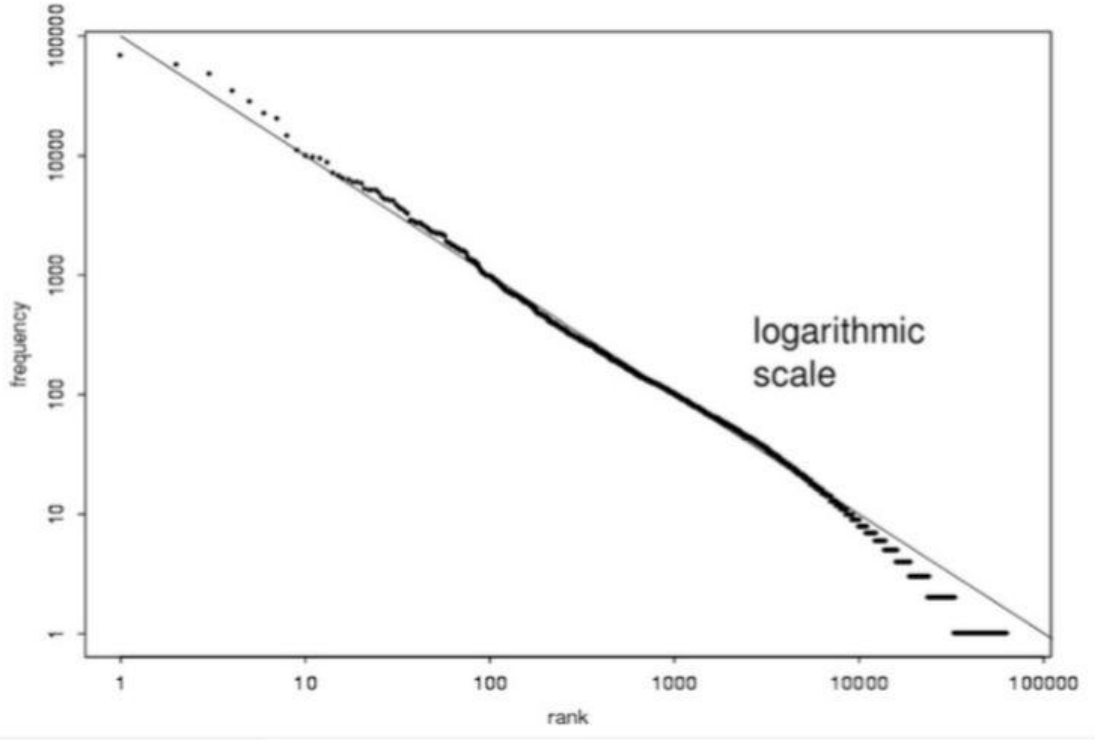
Normalización del texto

- Objetivo: Transformar varias formas del mismo término en una forma común normalizada
  - ejemplo: Manzana, manzana, MANZANA -> manzana
- Cómo se hace?
  - Siguiendo reglas simples
    - Remover todos los signos de puntuación
    - pasar todas las palabras a minúsculas
  - Usar un diccionario para reemplazar sinónimos con el término más común
    - carro, vehículo, automóvil, coche -> carro

Remoción de términos con alta y baja frecuencia

- Regla empírica
  - Unas pocas palabras con alta frecuencia
  - muchos términos con baja frecuencia
- La regla de Zipf:
  - La frecuencia de una palabra en un determinado corpus es inversamente proporcional a su rango en la tabla de frecuencias

Ilustración de la regla de Zipf



Implicaciones de la regla de Zipf

- Las palabras en la parte alta de la tabla de frecuencia corresponden a una porción significativa del corpus, pero son semánticamente casi inservibles:
  - el, la, ellos, eso, nosotros, tu, yo
- Por otra parte, las palabras en el fondo de la tabla de frecuencias contienen mucho significado, sin embargo tienen muy poca frecuencia
  - ejemplo: electroencefalografista
- El resto de palabras son las que representan mejor el corpus y son las que deberían determinar nuestro modelo de vectorización de texto

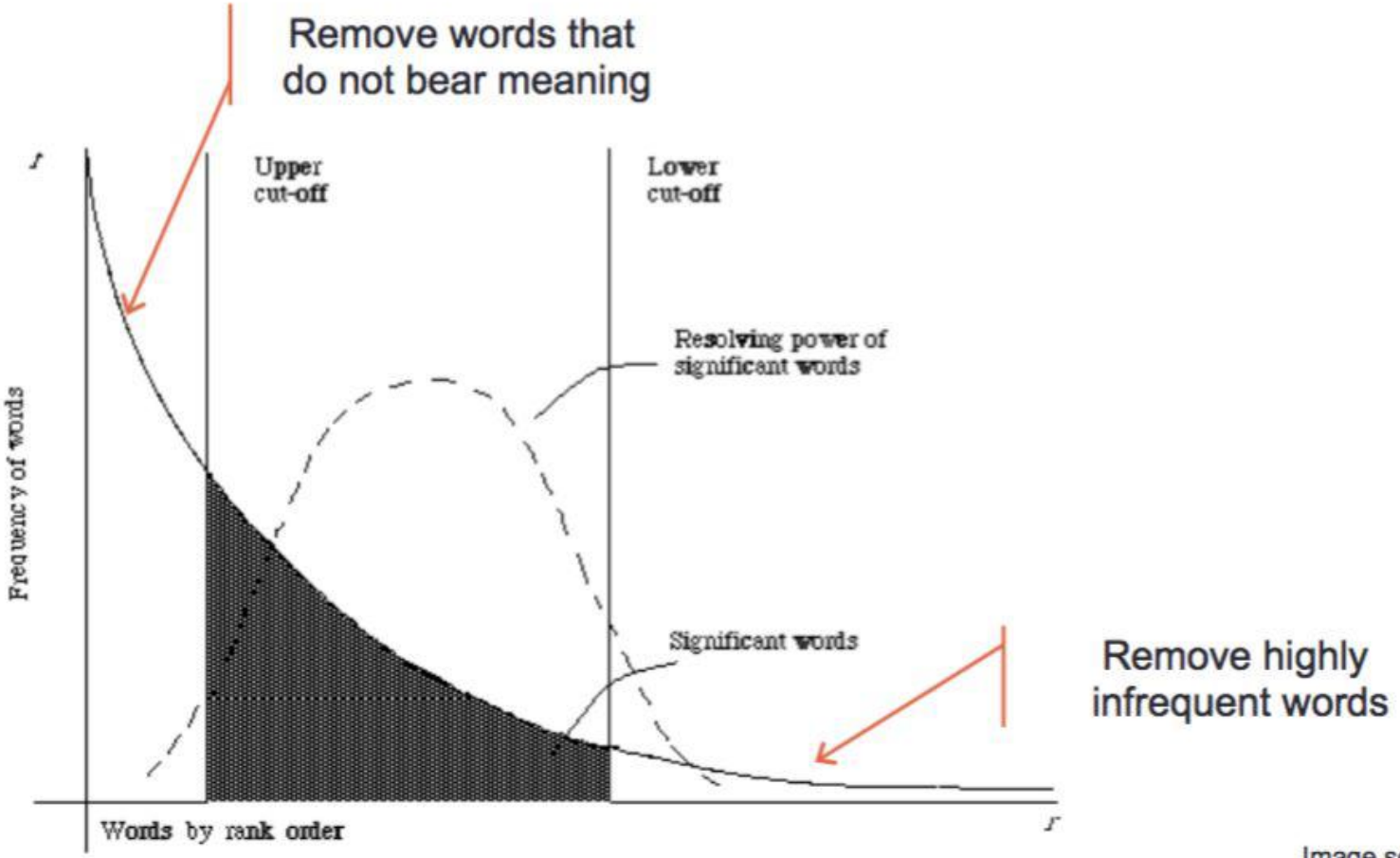


Image source:
http://www.dcs.gla.ac.uk/Keith/Chapter.2/Ch.2.html

Stopwords

- Una forma de eliminar palabras irrelevantes para el análisis
- stopwords son las palabras que por si mismas, no transmiten ningún significado
- se estima que pueden representar entre el 20% y el 30% de las palabras en cualquier corpus
- Existen muchas listas de palabras tipo stopwords
- Potenciales problemas
    - pérdida de significado ‘esta no es una buena opcion’ -> opcion
    - ‘ser o no ser’ -> NULL

Lematización y Stemming

- Dos procedimientos que buscan disminuir la variabilidad de las palabras reduciendolas a su raíz o forma básica
- Stemming es un proceso heurístico que remueve las terminaciones de las palabras sin considerar las propiedades linguisticas de la misma
    - ejemplo. argumentar, argumento, argumenta, argumentando -> argument
- Lematización se refiere al uso del vocabulario y análisis morfológico de las palabras, buscando retornar la forma base de la palabra, que se conoce como lemma
    - ej. argumentar, argumento, argumenta, argumentando -> argumentar

Calculando el peso de los términos

Hay diferentes formas de calcular los pesos de los términos.

- Pesos Binarios
- Frecuencia de Términos (TF)
- Frecuencia inversa de términos (IDF)
- TF-IDF

Pesos Binarios

Los pesos toman los valores 0 (cero) o 1 (uno), para reflejar la usencia o presencia de un término, respectivamente.

ejemplo

- Doc1: Text mining is to identify useful information.
- Doc2: Useful information is mined from text.
- Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Frecuencia de Términos

- La frecuencia de términos (TF) es el número de veces que aparece un termino t en el documento d
- supuesto
  - a mayor frecuencia, mayor relevancia del término dentro del documento

$TF(t) = c(t,d)$

c(t, d) es el número de veces que aparece el término t en el documento d

Frecuencia Inversa de términos (IFD)

- Idea: Asignar un mayor peso a términos poco usuales
- IDF se calcula al nivel del corpus y por lo tanto lo describe en su totalidad
- se calcula de la siguiente forma
  - $IDF(t) = 1 + \log(N(df(t)))$

N es el número de documentos en el corpus

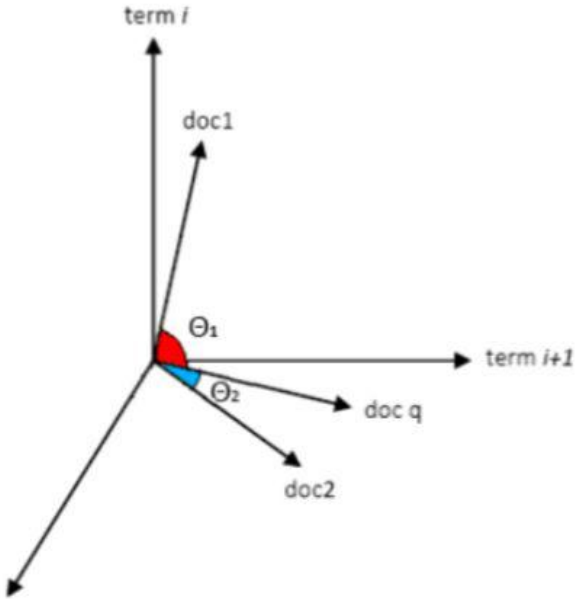
df(t) número de documentos donde aparece el término t

TF-IDF

- Idea: Darle valor a términos que no son muy comunes en el corpus (IDF relativamente alto) pero que tienen una frecuencia relativa alta dentro de un documento (TF relativamente alto)
- Es un esquema de vectorización muy usado
- Fórmula general para calcularlo
    - TF-IDF(t) = tf(t) * log(N/df(t))

Calculando la similaridad entre documentos

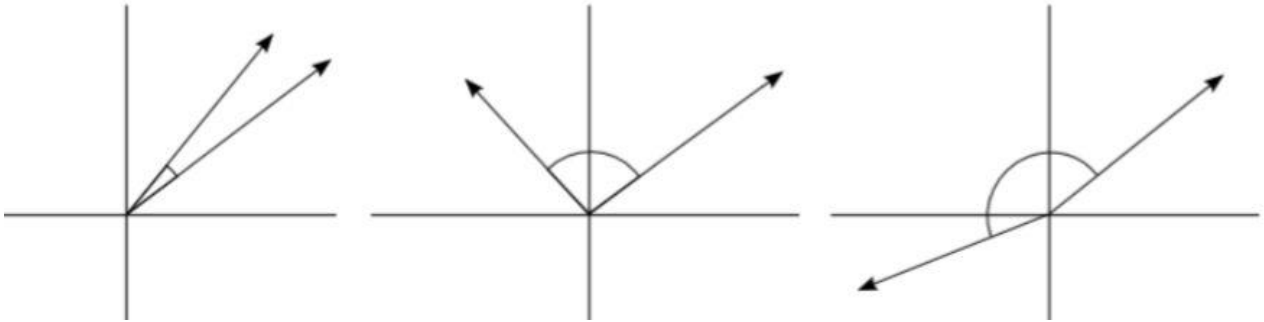
- Pregunta: Cuál métrica usar para similaridad de documentos?
- La más usada es la distancia coseno



Similaridad por distancia coseno

cos(d1, d2) = V1 • V2 / (||V1|| ||V2||)

V1 y V2 son los vectores que representan a los documentos d1 y d2



similares no relacionados opuestos

Ventajas/Desventajas

- Ventajas
  - Intuitivo
  - Fácil de implementar
  - Demostrado empíricamente que es efectivo
- Desventajas
  - Basado en que las palabras son mutuamente independientes
  - Ajustar los parámetros del modelo puede ser computacionalmente difícil

Modelos de Tópicos en Twitter

Aplicación a detección de disturbios