Introducción a la Minería de Textos

Tratamiento Inteligente de Datos

Master Universitario en Ingeniería Informática

UNIVERSIDAD DE GRANADA

Gabriel Navarro (*gnavarro@ugr.es*, *gnavarro@decsai.ugr.es*)

Karel Gutiérrez Batista (karel@decsai.ugr.es)

Objetivos

- Conocer la importancia de la minería de textos
- Entender el proceso y las fases de la minería de de textos
- Conocer las técnicas básicas de procesamiento de textos
- Conocer algunas aplicaciones de la minería de textos

Índice

- Concepto de minería de textos
- El problema de la minería de textos
- Preprocesamiento
  - Formas intermedia de representación
- Técnicas para la minería de textos
- Aplicaciones

Concepto de Minería de Textos

Se podría definir como

"… a knowledge-intensive process in which a user interacts with a collection of documents by using analytic tools in order to identify and explore interesting patterns"

C. C. Aggarwal and C. Zhai (editors), Mining Text Data, Springer, 2012.

Concepto de Minería de Textos

Es decir, siguiendo lo explicado en otros temas

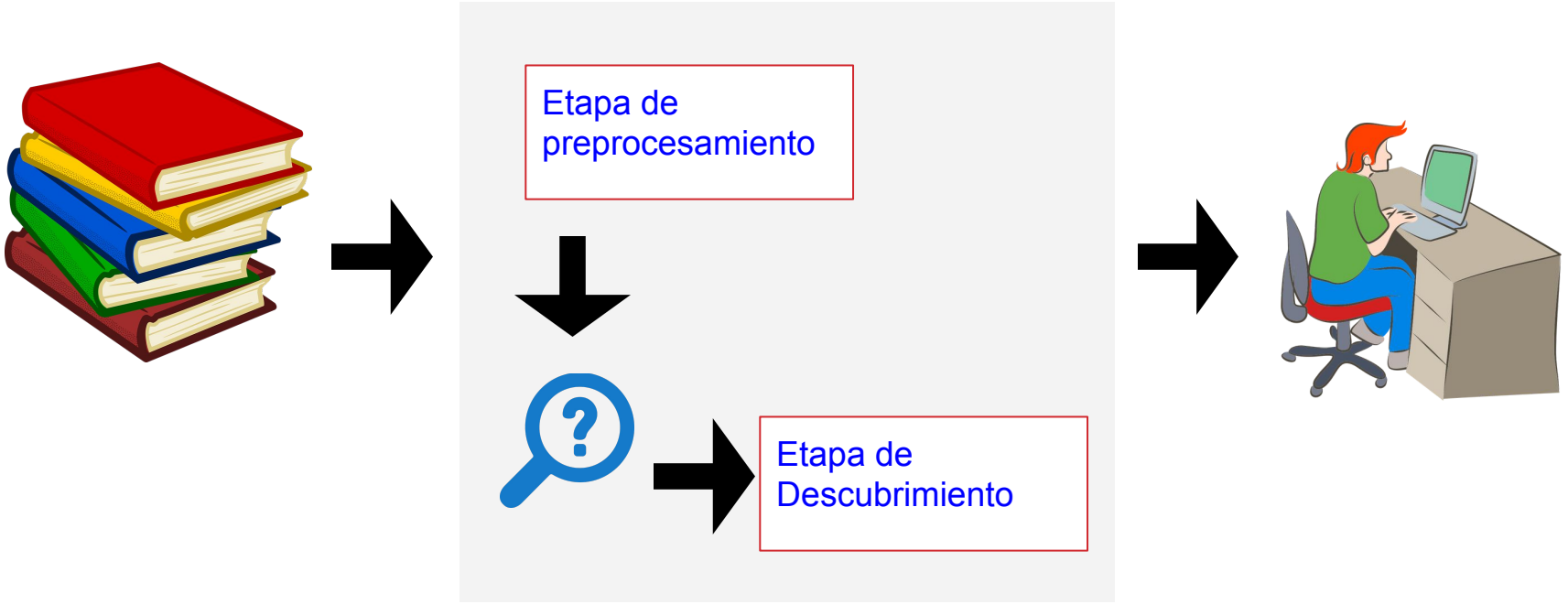
> *"Proceso de extracción de conocimiento o patrones, previamente desconocidos, no triviales e interesantes (potencialmente útiles) y comprensibles por los usuarios a partir de documentos de texto no estructurados."*

La Minería de Textos es una extensión de la Minería de Datos donde el descubrimiento se realiza a partir de bases de datos no estructuradas

El problema de la Minería de Textos

El problema de la Minería de Textos

Esquema sencillo del proceso de MT



El problema de la Minería de Textos

En general, no se pueden aplicar directamente las técnicas clásicas de Minería de Datos a la información textual.

Datos no estructurados/heterogéneos

- Diferentes formatos
- Diferentes idiomas
- Diferente semántica

Es necesario un tratamiento previo de los documentos para dar una estructura común (preprocesamiento)

El problema de la Minería de Textos

Esta ausencia de estructura constituye el mayor problema de la MT e implica la necesidad de preprocesar los textos, de pasarlos a una forma intermedia

- Bolsas de términos (bag of words)
- Estructuras matriciales (datasets)
- Grafos conceptuales o redes semánticas.
- Estructuras de tipo ”ontología”

El problema de la Minería de Textos

No se debe confundir **Minería de Textos (MT)** con **Recuperación de Información (RI)** a partir de bases de datos textuales. La RI busca ”documentos” de acuerdo con unos requerimientos.

En la MT buscamos:

- Conocimiento desconocido
- Comprensible por los usuarios
- No trivial
- Interesante

El problema de la Minería de Textos

Metodología para la Minería de Textos

1. Data overview (EDA)

2. Text preprocessing (http, lowercase, etc…)

3. Word embedding technique (Bag of word, TF-IDF, Hashing)

4. Using technique (Ridge Classifier,MultinomialNB, etc.)

5. Evaluating model (F1, precision vs recall, confusion matrix)

Motivación al preprocesamiento

Problema primario

Originalmente las colecciones documentales están en un formato (o varios formatos diferentes) que no puede ser tratado directamente por los sistemas de minería de textos

Solución

Convertir dicha colección a un formato manejable

Preprocesamiento

1. Preprocesamiento sintáctico
2. Preprocesamiento semántico

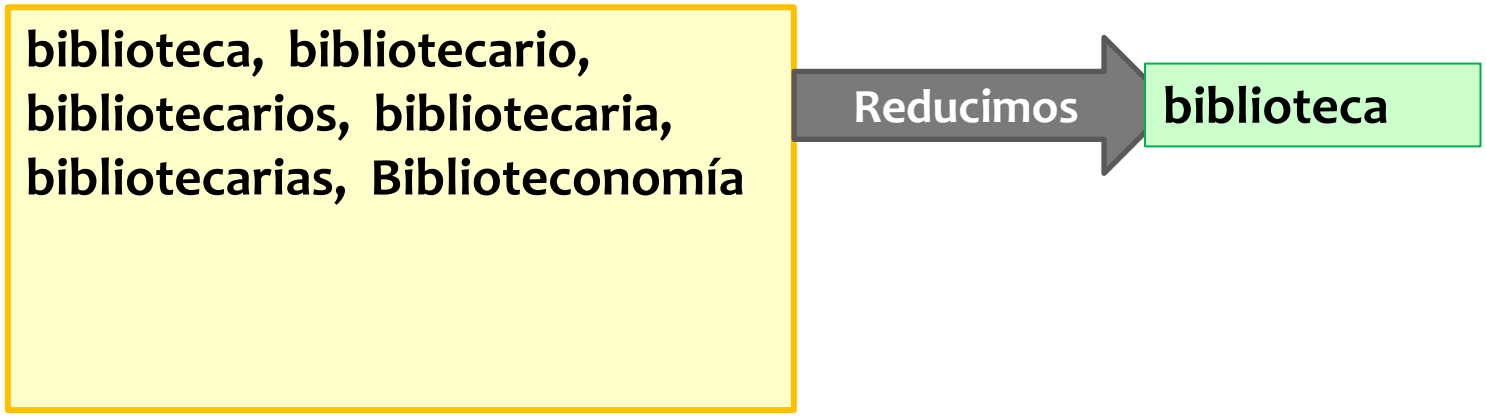
Preprocesamiento Sintáctico

1. Análisis léxico (tokenizing)
2. Eliminación de palabras vacías (stop words)
3. Segmentación (stemming) o lematización
4. Ponderación de términos (weighting)
5. Selección de los mejores términos
6. Reconocimiento de palabras múltiples. (n-gramas)
7. Análisis de la categoría gramatical (**POS-tagging)**

Segmentación y lematización

La segmentación (stemming) y la lematización son dos métodos usados para reducir el tamaño del vocabulario

- En lugar de indexar todas las palabras, se buscan sus "representantes" morfológicos, raíces o lemas
- La raíz es la parte de la palabra que queda al eliminar afijos (prefijos, infijos y sufijos)



Segmentación y lematización

Aunque se busca lo mismo, reducir palabras a una base común, son cosas diferentes

- Por segmentación se entiende a un proceso heurístico que corta el final de las palabras, y con frecuencia incluye la eliminación de los afijos derivativos
  - **coches** a **coche**, **automatic** a **automat**, **fuimos** a **fui**…
- Por lematización se entiende a hacer uso de un vocabulario y análisis morfológico de las palabras con el objetivo de eliminar las terminaciones flexivas y devolver la base de una palabra, que se conoce como el lema
  - **fuimos** a **ir**, **am** a **be**, …

Segmentación y lematización

Ventajas

- Reducción del vocabulario, eficiencia y ahorro de espacio

Desventajas

- Se pierde información sobre la palabra completa

Preprocesamiento Sintáctico

No siempre es útil llevar a cabo la lematización.

- En general para los casos en que se trabaje con frecuencia de términos es interesante realizarlo ya que resume varios términos en un sólo y aumenta frecuencia de este.

- Pero puede dar problemas si se hace un preprocesamiento semántico posterior

Preprocesamiento Sintáctico

Algunas categorías de POS en inglés:

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
POS	Possessive ending
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
WDT	Wh-determiner

Preprocesamiento Semántico

Idea básica

Una vez limpios y etiquetados los términos se utilizan relaciones de tipo semántico para reducirlos nuevamente.

Posibles relaciones entre los términos:

- Sinonimia: distinta forma, igual significado (clase, lección)

- Homonimia: misma forma distinto significado (banco institución financiera, sitio de sentarse)

Preprocesamiento Semántico

- Polisemia: misma forma ,distintos significado relacionado (Banco, Banco de sangre)
- Hiponimia Una palabra es una subclase de otra (perro, animal)
- Hiperonimia Una palabra es una superclase de otra (animal, perro)

Preprocesamiento Semántico

Desambiguación

Proceso mediante el cual se asigna a varios términos uno sólo que tiene el mismo significado que todos ellos.

- Existen herramientas que ayudan a trabajar con sinónimos e hipónimos, llegando a asignar cada conjunto de términos a una clase semántica (desambiguación).
- La más famosa es Wordnet; pero no existe un algoritmo que resuelva totalmente el problema y menos en varios idiomas..

Problemas

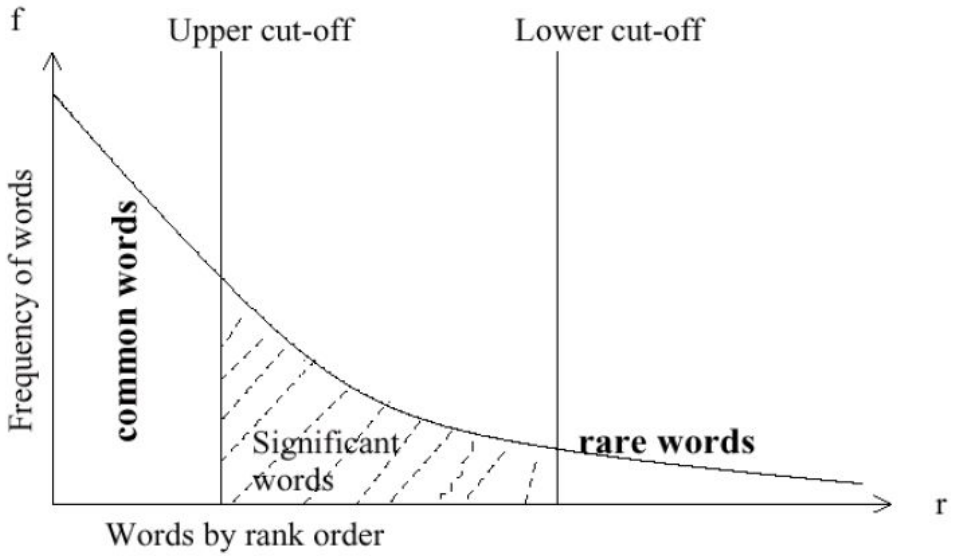
Con esto llegamos a que cada documento es un conjunto de términos de indexación (bag-of-words model)

Aquí surgen varios problemas:

- Considerar todos los términos sigue siendo muy costoso (selección de términos)
- Considerar todos los términos igual de importantes puede empeorar la eficacia de la MT (ponderación de términos)
- Representar cada documento simplemente como un conjunto restringe las operaciones sobre ellos (modelos de representación)

Selección de términos

Métodos basados en frecuencia de aparición



Según Luhn, las palabras importantes son aquellas que:

- son capaces de discriminar el contenido de documentos
- se sitúan en medio de dos umbrales

Ponderación de términos

- No todos los términos de un documento deberían ser igual de importantes…
- Podemos asociar a cada uno de los términos de indexación un peso que refleje la importancia del término en el documento

Resoluc 0.2, don 0.4, quijote 0.4, enfrasc 0.8, lectura 0.4, claro 0.7, dia 0.3 .....

Cada documento lo vemos como un conjunto de pares

$$d = \{(t_1, w_1), (t_2, w_2), \ldots, (t_n, w_n)\}$$

Ponderación de términos

Modelo Booleano

Considera la presencia (1) o ausencia (0) de un término en los documentos de la colección

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	0
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Ponderación de términos

Modelo Booleano extendido

Con la frecuencia de cada término en cada documento

$$tf_{t,d} = \text{num. ocurrencias de t en el documento d}$$

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Ponderación de términos

Problemas con esa forma de ponderar:

- Un término que aparece en pocos documentos en la colección tendrá un mayor poder discriminador que los que aparecen en casi todos
  - Solución: Frecuencia documental inversa
- Los términos aparecerán más en documentos más largos
  - Solución: Normalización de la frecuencia por la máxima frecuencia en el documento o por la longitud del mismo

Ponderación de términos

Inverse document frequency

- Dado un término t, un idf_t sencillo

$$idf_t = \frac{N}{N_t} \text{ donde } \begin{cases} N \text{ total de documentos} \\ N_t \text{ documentos con el término} \end{cases}$$

- Se daría un peso mayor a los términos que aparecen en menos documentos y un peso bajo a aquellos que aparecen en muchos.

Suavizamos tomando logaritmos $idf_t = \log \frac{N}{N_t} + 1$

Ponderación de términos

Medida tf-idf (term frecuency – inverse document frecuency)

Ponderamos la combinación de ambas frecuencias

$$w_{d,t} = tf_{d,t} \cdot idf_t = tf_{d,t} \cdot \left(\log \frac{N}{N_t} + 1 \right)$$

Objetivo: asignar pesos más altos a aquellos términos que

- sean frecuentes en documentos relevantes, pero...
- infrecuentes en la colección

Es el mejor esquema conocido para calcular pesos

Ponderación de términos

Normalización

Evitamos favorecer documentos largos

$$w_{d,t} = \frac{tf_{d,t} \cdot idf_t}{\sqrt{\sum_t (tf_{d,t})^2 \cdot (idf_t)^2}}$$

En cualquier caso…

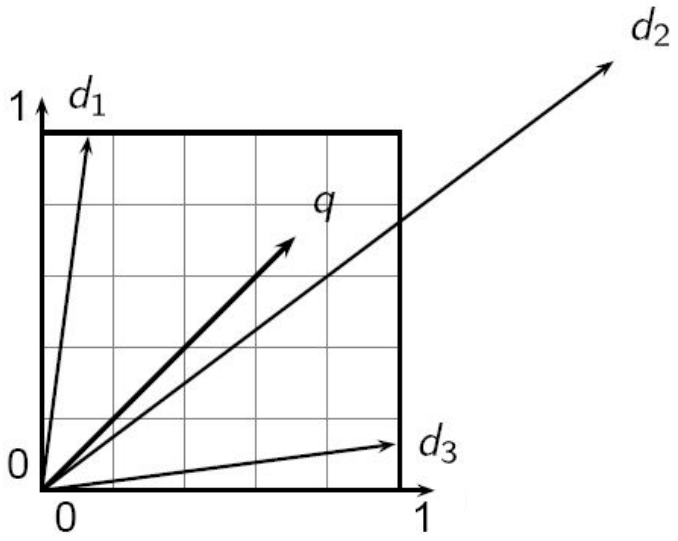
- podemos utilizar muchas modificaciones
- su utilidad depende de la evaluación y comparación

Modelo vectorial

Respecto a la similitud entre documentos, podemos utilizar la distancia Euclídea (como puntos del espacio)

[redacted]

Pero no es una buena idea, vectores muy parecidos pueden tener distancia grande

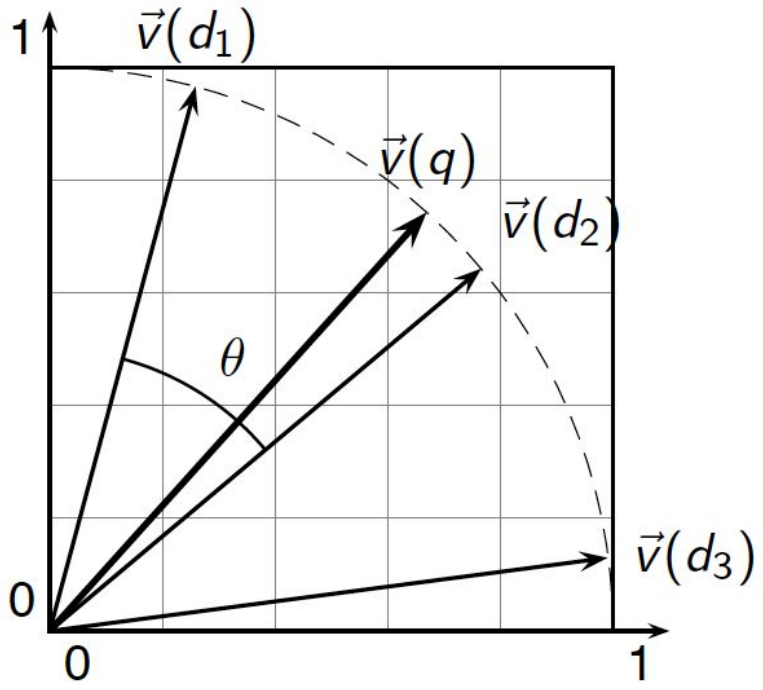


¿Qué pasa si hay un documento d y otro que es dos copias seguidas de d?

Modelo vectorial

Ángulo entre dos vectores

- A menor ángulo, más similitud
- Arregla la situación anterior
- Ahora, un documento y su copia doble tienen similitud máxima



Modelo vectorial

Entonces la **similitud** entre consulta y documento,

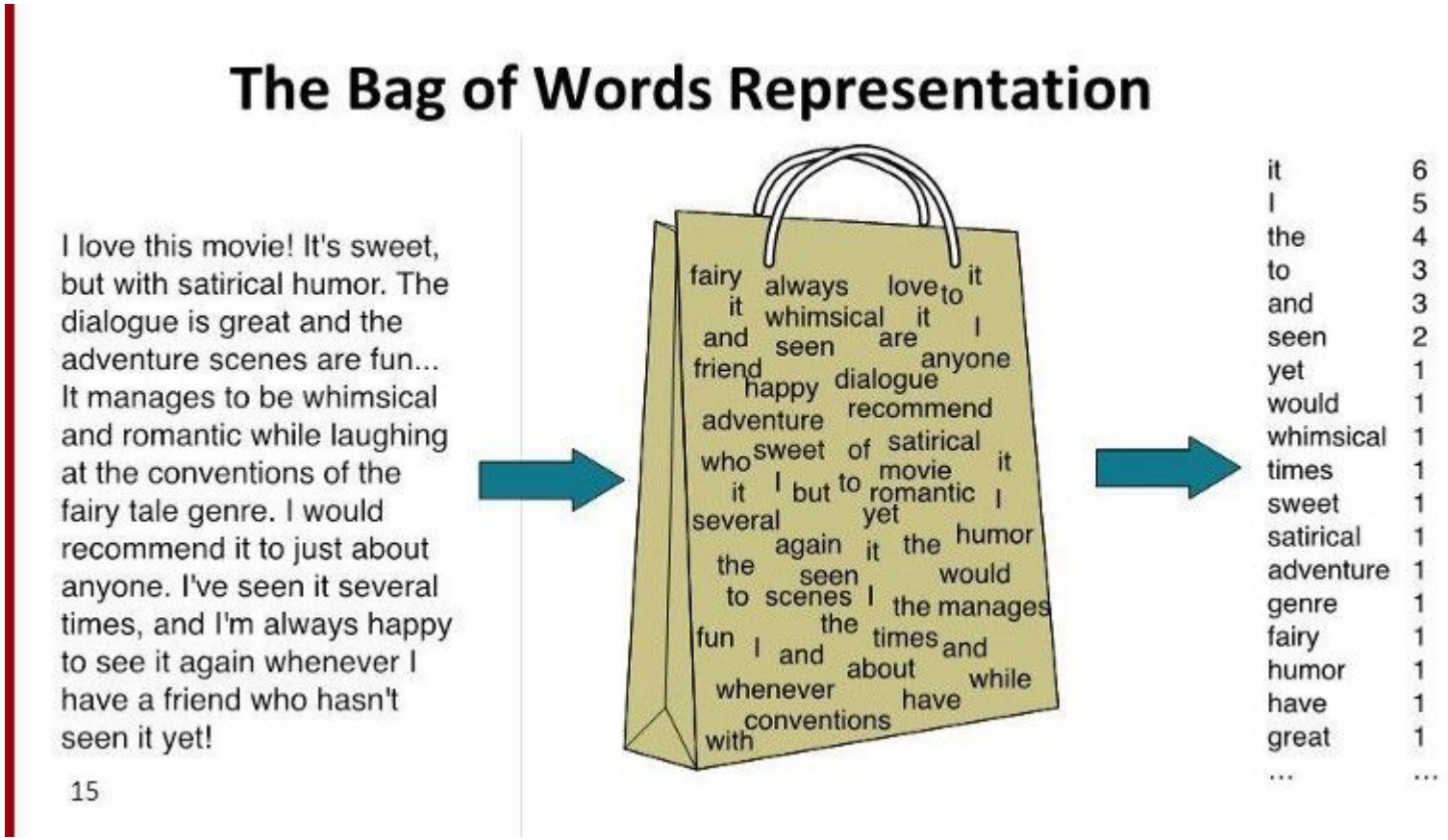
$$Sim(q,d) = \frac{q \cdot d}{|q||d|} = \frac{q}{|q|} \cdot \frac{d}{|d|}$$

Si los vectores están normalizados, es el producto escalar usual

$$Sim(q,d) = q \cdot d$$

Formas intermedia de representación

- Bag of words



Formas intermedia de representación

- TF-IDF

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

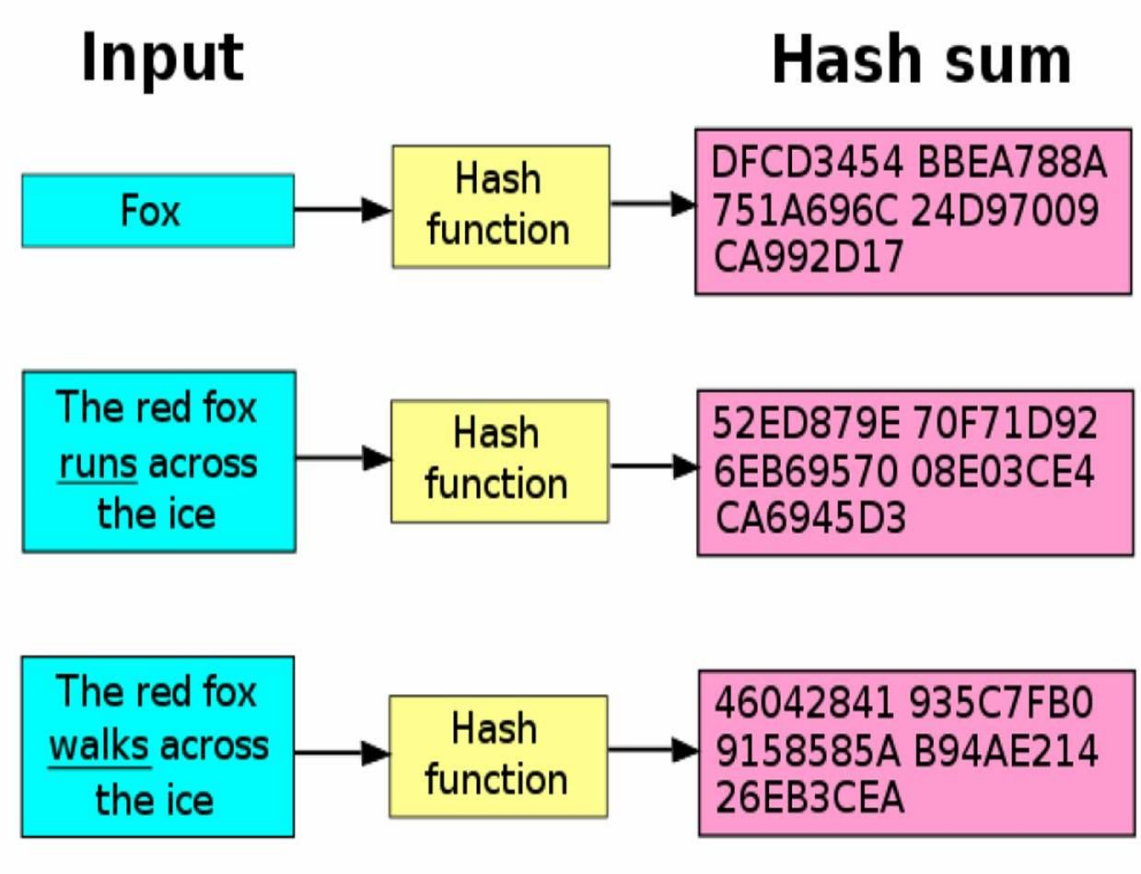
$\mathrm{tf}_{i,j}$ = total number of occurences of i in j

df_i = total number of documents (speeches) containing i

N = total number of documents (speeches)

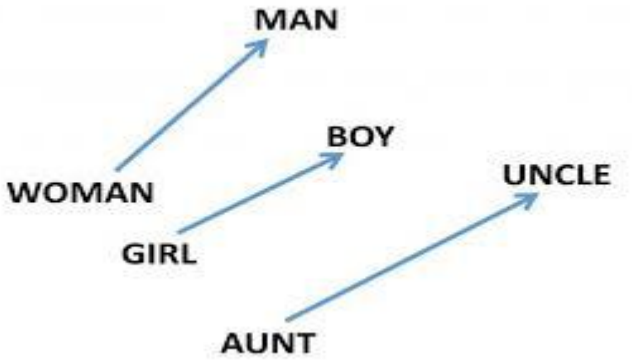
Formas intermedia de representación

- Hashing



Formas intermedia de representación

- Word embeddings
  - Representaciones distribuidas de una palabra en un espacio vectorial de baja dimensión.
  - Permite capturar característica sintácticas y semánticas del lenguaje (e.g. Male-Female relationship).



Técnicas para la Minería de Textos

No supervisadas

- Agrupamiento
- Reglas de asociación
- Modelado de tópicos (LDA, NMF, etc.)

Supervisadas

- SVM
- Naive Bayes
- Decision Tree
- Redes Neuronales (SNN)
- Deep Neural Networks (RNN, CNN y Language Models)

Aplicaciones

- Document summarization
- Machine translation
- Question answering
- Document recommendation
- Document classification (p.e. spam)
- Sentiment analysis (sentido de las críticas, escritos,..)
- Topic modeling
- …

Trabajos evaluables

(3 personas, 30 minutos) Detección de tópicos.

Bibliografía:

- Martin C, Corney D, and Goker A. Mining newsworthy topics from social media, in BCS SGAI Workshop on Social Media Analysis, Cambridge, UK, 2013. pp 35–46.
- Gao N, Gao L, He Y, Wang H, and Sun Q. Topic detection based on group average hierarchical clustering, in International Conference on Advanced Cloud and Big Data (CBD, 2013), IEEE, 2013. pp 88–92.
- https://medium.com/towards-artificial-intelligence/unlock-the-power-of-text-analytics-with-natural-language-processing-2e6d83b35f99

Trabajos evaluables

(3 personas, 30 minutos) Análisis de sentimientos.

Bibliografía:

- Kim, Soo-Min and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2004). 2004.
- Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
- Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press.doi:10.1017/CBO9781139084789

Bibliografía

- Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In KDD (Vol. 95, pp. 112-117).
- Introducción a la Minería de Datos. José Hernández Orallo, M.José Ramírez Quintana, Cèsar Ferri Ramírez. Pearson, 2004. Capítulo 21.
- Berry, MW. and Kogan, J. (2010). Text Mining. Applications and Theory. John Wiley & Sons
- C. Aggarwal, Data Mining: The textbook, Springer, 2015.
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. (2013)

Bibliografía

Algunas transparencias y gráficos tomados de:

- http://sci2s.ugr.es/docencia/in/
- http://elvex.ugr.es/idbis/dm/
- http://xeushack.com/on-hashes
- https://medium.com/shallow-thoughts-about-deep-learning/can-tfidf-be-applied-to-scene-interpretation-140be2879b1b
- https://www.programmersought.com/article/4304366575/