Minería de textos

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Introducción al preprocesado de textos y páginas web

Conceptos básicos

UCO ONLINE

“Si todos los datos del mundo se equiparasen al agua de la tierra, el texto sería los océanos”

F. Siegel, 2013

Minería de textos

- Los datos no estructurados (imágenes, texto, audio, vídeos, etc.) suponen la frontera a la ciencia de datos
- La minería de textos, más que cualquier otra técnica de minería de datos, cumple con la **metáfora del minero**:

  Separar los sucio de los elementos de valor (p.ej. metales)

- La minería de textos pretende separar palabras clave (*keywords*) o términos de valor de la masa de texto para identificar patrones significativos (descripción) o hacer predicciones
- Gran parte del esfuerzo de la minería de textos se ha dedicado a la descripción de documentos

Minería de textos

El texto es –por su naturaleza – difícil de analizar: debe convertirse en un activo no ambiguo, manejable y separado en elementos que puedan tratarse numéricamente

- ***Descriptive analytical methods***

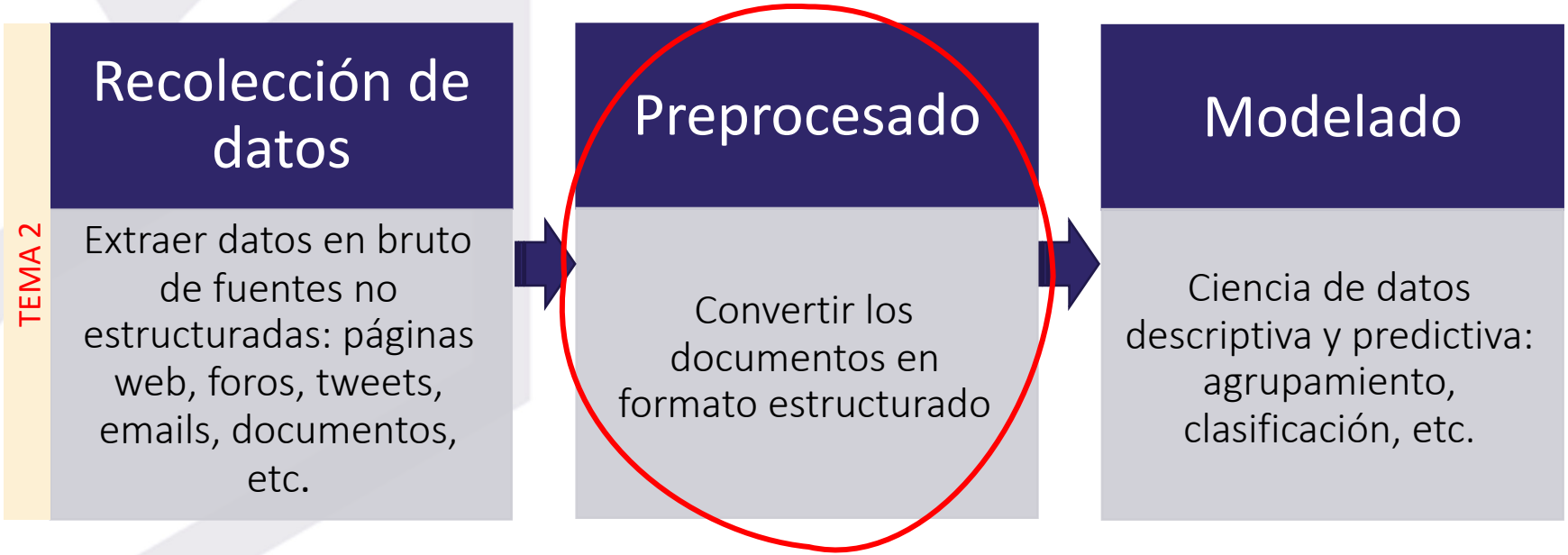
  Son métodos que muestran patrones de similitud, discutir qué hay cerca de un item de interés, cuentan ocurrencias y ofrecen una visión general de los patrones

- ***Predictive analytical methods***

  - Muestran si es posible observar alguna influencia, predicción, comportamiento o preferencia en la salida basada en las variables de entrada

Los modelos descriptivos son una buena base para construir modelos predictivos

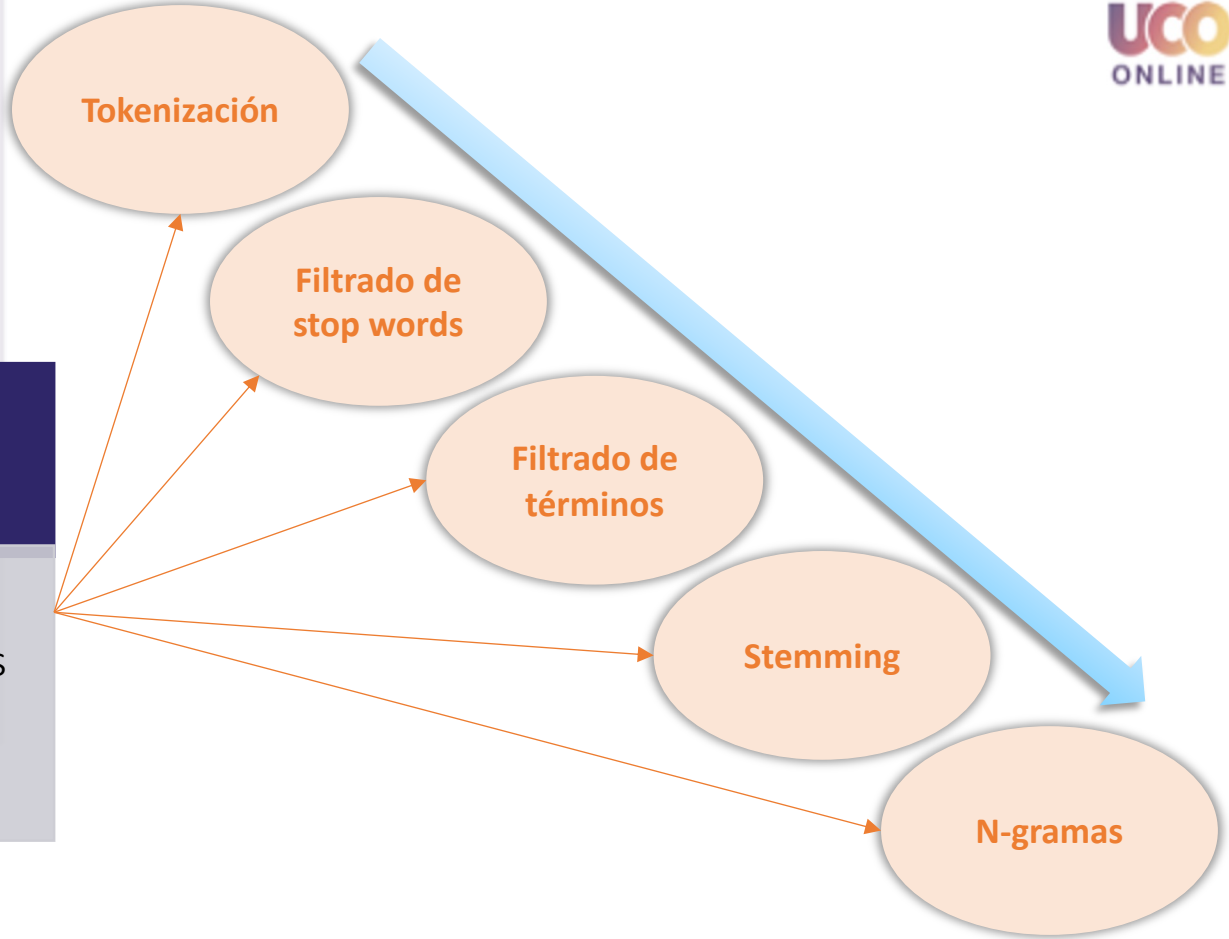
El proceso de la minería de textos



- El preprocesado de textos es un **paso fundamental** en la minería de textos
  - El objetivo es convertir el texto (no-estructura) en datos semi-estructurados
  - Los datos semi-estructurados ya pueden ser aplicados a métodos analíticos (clasificación, sumarización, agrupamiento, predicción, etc.)
- La minería de textos se focaliza en el análisis de **documentos**
  - La granularidad del documento la establece el científico de datos según sus necesidades
  - Un documento puede ser una frase, un tweet, un comentario, una página web o un artículo completo, todo dependiendo del contexto
- Un documento es una colección secuencial de **tokens**

Preprocesado

Convertir los documentos en formato estructurado



Consideraciones previas al preprocesado

Antes de tokenizar, lo habitual es preparar el texto: Eliminar caracteres especiales, comprobación de ortografía, cambio de mayúsculas o minúsculas, etc.

Introducción al preprocesado de textos y páginas web

Preprocesado de textos

UCO ONLINE

Tokenización

- Las búsquedas deben seguir los dos siguientes principios:
  - Dar un peso alto a aquellas palabras clave que son relativamente raras
  - Dar un peso alto a aquellas páginas web que contienen un gran número de instancias de palabras clave raras

¿Qué es una palabra clave rara?

Revisitamos la técnica TF-IDF

- **TF** (*Term Frequency*) es muy sencillo: el ratio entre el número de veces que una palabra clave aparece en un documento, $\mathbf{n_k}$ (k es *keyword*), y el número total de términos del documento, **n**

$$\mathrm{TF} = \frac{n_k}{n}$$

Tokenización

- **IDF** (*Inverse Document Frequency*) se define como:

$$\mathrm{IDF} = \log_2\left(\frac{N}{N_k}\right)$$

donde N es el total de documentos siendo minados y N_k es el número de documentos que contienen la palabra clave k

Por ejemplo, si se busca "*Libros de RapidMiner que describen la minería de textos*"

- "que" tendrá un valor muy alto de TF; además, aparecerá en todos los documentos (páginas web), por lo que el ratio N/N_k será cercano a 1 y, por tanto, IDF será cercano a cero.
- "RapidMiner" tendrá un valor muy bajo de TF; además, aparecerá en pocos documentos, por lo que el ratio N/N_k será mucho mayor que 1 y, por tanto, IDF será alto.

- **TF-IDF** se define como el producto:

$$\mathrm{TF-IDF} = \frac{n_k}{n} \times \log_2\left(\frac{N}{N_k}\right)$$

Tokenización

- **TF-IDF** se calcula para cada palabra del conjunto de documentos
  - Se presta más relevancia a los términos con valores altos de TF-IDF

- Se genera a continuación el **vector de documentos** o **matriz de documentos de términos** (TDM, *term document matrix*)
- En la TDM se estructuran los datos brutos creando una matriz sobre el conjunto de ejemplos (***example set***) donde las columnas son los tokens y las filas los documentos
  - Token → atributo
  - Documento → ejemplo

Tokenización

- Supongamos los siguientes documentos:

Document 1	This is a book on data mining
Document 2	This book describes data mining and text mining using RapidMiner

- El vector de documentos sería:

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

¿Qué ocurre si se añaden más documentos?

Ejemplo extraído de "Data science concepts and practice", 2nd Edition

Tokenización

Document 1	This is a book on data mining
Document 2	This book describes data mining and text mining using RapidMiner

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Cálculo de TF

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

TDM, *Term document matrix.*

Cálculo de TF-IDF (normalizado)

RapidMiner	This	a	and	book	data	describes	is	mining	on	text	using
0	0	0.577	0	0	0	0	0.577	0	0.577	0	0
0.447	0	0	0.447	0	0	0.447	0	0	0	0.447	0.447

Palabras vacías

- Es necesario reducir al máximo el número de términos en la matriz a aquellos que realmente aporten algún valor o significado
- La mayoría de los elementos gramaticales como artículos, conjunciones, preposiciones y pronombres deben ser filtrados
- El **filtrado de palabras vacías** (*stop word filtering*) se realiza justo tras la tokenización
- Requiere de un diccionario conforme al idioma que se esté tratando

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Palabras vacías

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Filtrado de palabras vacías

RapidMiner	book	data	describes	mining	text	using
0	1	1	0	1	0	0
1	1	1	1	2	1	1

Filtrado de términos

- El **filtrado de términos** (*term filtering*) consiste en la eliminación de palabras que son comunes en un determinado contexto
  - Por ejemplo, si se habla de la industria del automóvil puede ser interesante eliminar los términos “automóvil”, “coche”, “vehículo”, etc.
  - Son términos que en ese contexto probablemente no van a proporcionar una diferencia significativa pero que pueden reducir el espacio de características
- La **sustitución léxica** (*lexical substitution*) es el proceso de encontrar un término alternativo para una palabra en un contexto, que permita alinear los tokens antes de su análisis
  - Especialmente relevante en áreas con jergas específicas (p.ej. terminología clínica)
  - Por ejemplo, en según que contextos, “coche” y ”automóvil” podría sustituirse por “vehículo”
- Ambas técnicas pretenden reducir el tamaño de la matriz TDM

Stemming

Einstein es un científico muy reconocido

El nombre de Einstein es reconocible por los científicos

Cualquier científico reconoce el nombre de Einstein

Stemming

Einstein es un científico muy reconocido

El nombre de Einstein es reconocible por los científicos

Cualquier científico reconoce el nombre de Einstein

Todos estos términos comparten la misma raíz (root) → reconocer

Stemming

- Reducción de términos de un documento a sus raíces básicas → simplifica la conversión de texto no estructurado a datos estructurados
  - El proceso se reduce a ocurrencias de términos raíz, no de sus variantes
- Al proceso de reducción al mínimo esencial de un término (raíz) se denomina ***stemming***
- El **Método de Porter** (1980) es el más conocido para minería de textos en inglés
  - Se basa en la simplificación (eliminación o reemplazo) de sufijos de las palabras
  - Ejecuta reglas sobre el texto como: (1) reemplazar “ies” → “y”; (2) eliminar “s”
  - Es un método muy eficiente pero que comete errores que pueden ser costosos en contextos que requieren precisión
- Hay otros métodos disponibles → su elección depende de la experiencia en el dominio

Stemming

- La **lematización** (*lemmatization*) es un tipo especial de stemming que normaliza las palabras al mismo tiempo que intenta averiguar su parte del discurso
- En este caso, la reducción de las palabras a la raíz puede diferir de cómo se utilizan
  - Por ejemplo, el sustantivo “venda” no sería reducido, mientras que el verbo “venda” sí sería reducido a su raíz vender

---

- En entornos web hay que prestar atención también a la **revisión de la ortografía** del texto de usuarios, eliminación de palabras sin sentido e iconografía
- La **extracción de entidades con nombre** (***named entity extraction***) consiste en identificar un conjunto o grupo de palabras para que el ordenador entienda que tienen un único significado
  - A menudo se requiere de un diccionario o léxico especial que se aplica

Stemming – Errores comunes

- ***Over-Stemming***. Términos con diferentes significados son transformados a una misma raíz

  Por ejemplo: universidad – universo.

- ***Under-Stemming***. Términos con similar significado no son reducidos a una misma raíz

  Por ejemplo: máquina – maquinaría.

El *over-stemming* reduce la precisión y el *under-stemming* reduce el *recall*

Un **modelo predictivo** usable raramente tendría más de 70-80 variables

Incluso un **modelo descriptivo** dejaría de funcionar adecuadamente con más de 100 variables

N-gramas

- En el lenguaje, hay familias de palabras que habitualmente aparecen juntas
- El agrupamiento de estos términos, llamados **n-gramas**, y su análisis estadístico pueden aportar nuevos descubrimientos de los datos
- **Usos habituales**: consultas en buscadores web, traducción automática, identificación de patrones de habla, detección de entidades, comprobación de faltas ortográficas, extracción de información, etc.
- Aunque Google es capaz de encontrar un n- alto, normalmente 2- y 3-gramas son suficientemente útiles en la mayoría de aplicaciones
- Los algoritmos para formar y almacenar n-gramas son computacionalmente costosos y con resultados inmanejables, por lo que elegir “n” dependerá del número de documentos y tamaño del *corpus*

N-gramas

RapidMiner	book	book_data	book_descr...	data	data_mining	describes	describes_data	mining	mining_text	mining_usi...	text_0	text_mining	using	using_RapidMiner
0	0.447	0.447	0	0.447	0.447	0	0	0.447	0	0	0	0	0	0
0.243	0.243	0	0.243	0.243	0.243	0.243	0.243	0.485	0.243	0.243	0.243	0.243	0.243	0.243

Los n-gramas significativos muestran los valores TF-IDF más altos

N-gramas

- Se han desarrollado una serie de métodos para conseguir que las palabras tengan una forma más manejable – todos funcionan sobre textos ya preprocesados
- **Ventana deslizante** (*sliding window*) –
  - Consiste en una caja de longitud fija (p.ej. 6 palabras) que se desliza por el texto
  - Las ventanas deslizantes cuentan la frecuencia con la que las palabras aparecen juntas en la caja a medida que ésta se desplaza por el documento, por lo que son especialmente valiosas para encontrar patrones de similitud y asociación entre las palabras
  - La ventana deslizante de n-gramas permite obtener todas las representaciones pictóricas de los patrones (nubes de palabras).

- **Codificación automática** (*automated coding*) –
  - Consiste en dividir el texto en agrupaciones significativas, codificando los documentos según el significado de las palabras y frases
  - El programa es capaz de desarrollar un intrincado sistema de códigos y subcódigos, en el que las ideas más amplias contienen otras más específicas
  - Lo más probable es que encuentre algunas cosas que necesiten ser ajustadas

- **Análisis de factores** –
  - Agrupa las palabras en función de su uso en los documentos → Las palabras que entran en un mismo factor tienden a reflejar un tema común
  - El humano es responsable de nombrar cada factor basándose en la idea que parecen representar las distintas palabras reunidas en ese factor. Ver si se pueden encontrar nombres sencillos para los factores es una prueba excelente para decidir si la solución tiene sentido
  - Lo ideal es que las variables más fuertes en los factores (que tienen las cargas más fuertes) se refieran todas a un único concepto subyacente
  - El análisis factorial puede utilizarse para modelos predictivos

¡Gracias!

UCO ONLINE