#AprendeDesdeCasa
#AprendeConLosPioneros

Online | DMC Perú

CURSOS ANALYTICS

Machine Learning Advanced

- Text Analytics -

Docente: Manuel Montoya

Agenda

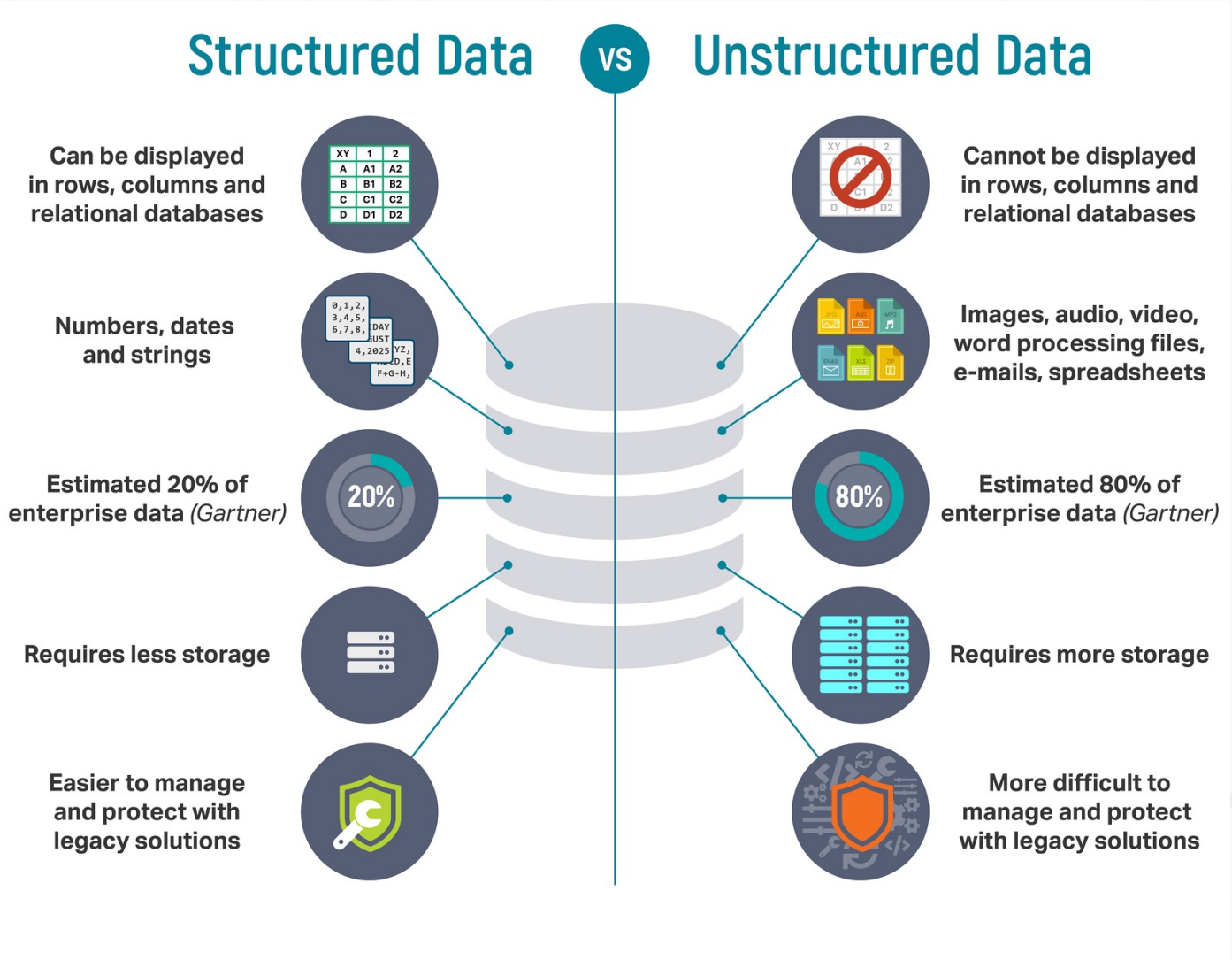
1. Data no estructurada
2. Text Mining
3. Representación y preprocesamiento de textos

Data no estructurada

Data estructurada y no estructurada

La data estructurada usualmente se encuentra en forma matricial: archivos planos, excel, bases de datos relacionales con columnas que representan las variables y filas que representan cada registro.

La data no estructurada es **todo lo demás:** imágenes, videos, información de redes sociales, mensajes de texto, correos, libros, llamadas a un call center, coordenadas, información en sitios web, etc..



https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

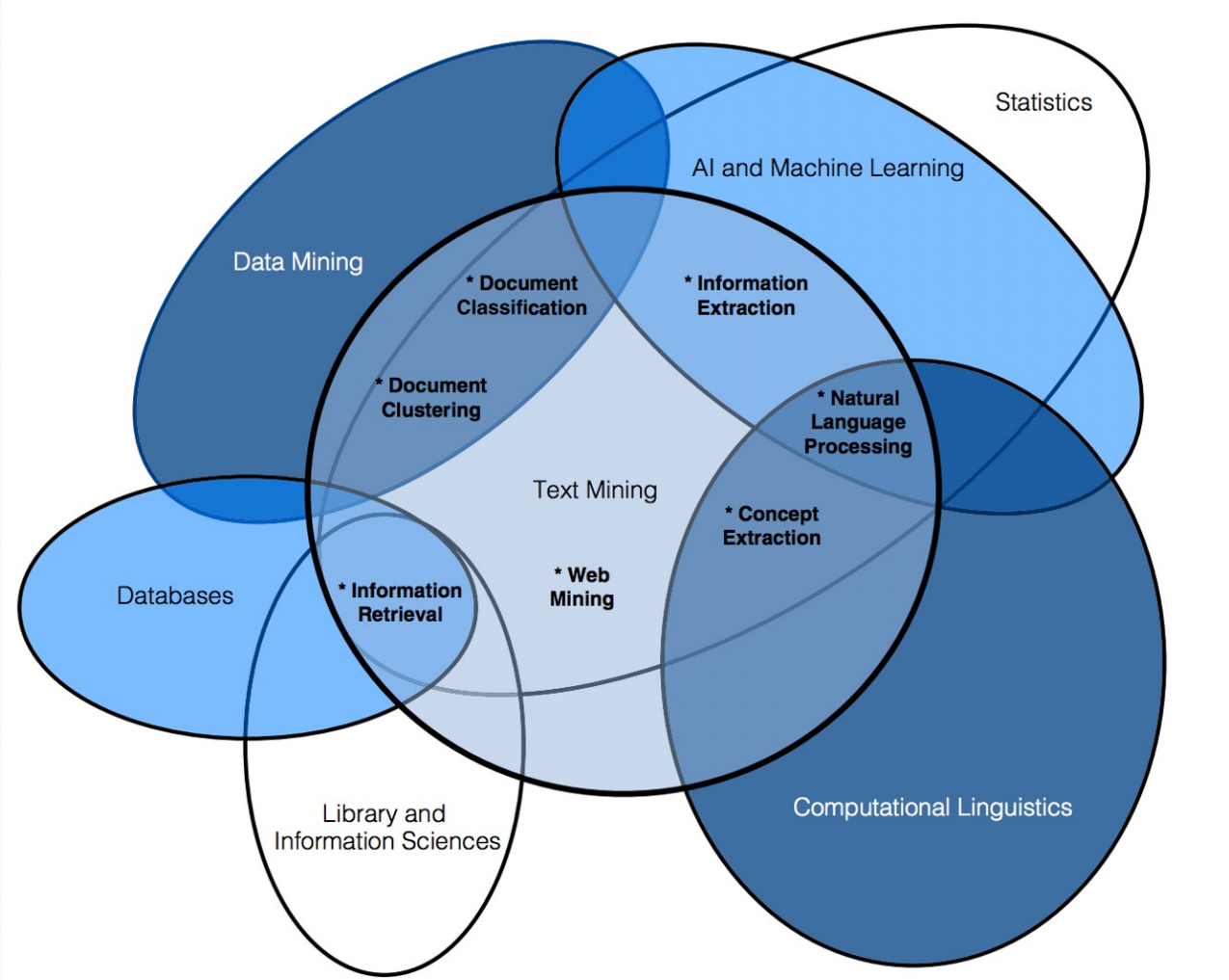
Text Mining

¿Qué es Text Mining?

Existen muchos ejemplos de documentos basados en textos: correos, páginas web, encuestas, currículums, papers de investigación, noticias, transcripciones de llamadas a call centers.

¿Tenemos suficiente tiempo (o paciencia) para leerlos todos?

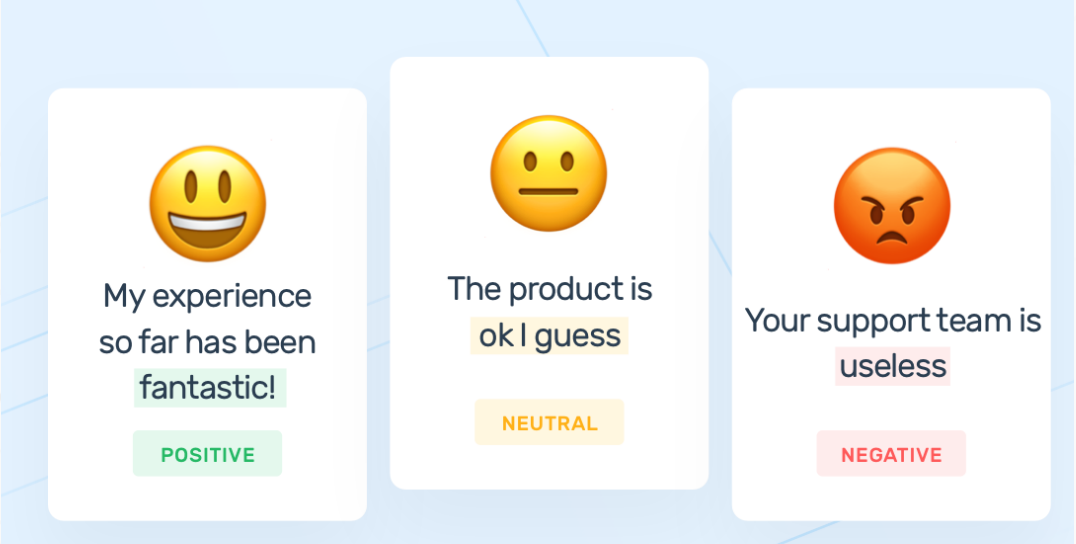
Queremos encontrar una forma de obtener información (resumida) de estos textos sin tener que leerlos o examinarlos todos.


Statistics
AI and Machine Learning
Data Mining
* Document Classification
* Information Extraction
* Document Clustering
* Natural Language Processing
Text Mining
* Concept Extraction
Databases
* Information Retrieval
* Web Mining
Library and Information Sciences
Computational Linguistics

Clasificación y Clustering de documentos

Clustering:
Agrupar automáticamente los documentos relacionados según su contenido

Clasificación:
Asignar un conjunto de etiquetas a los documentos y entrenar un modelo predictivo

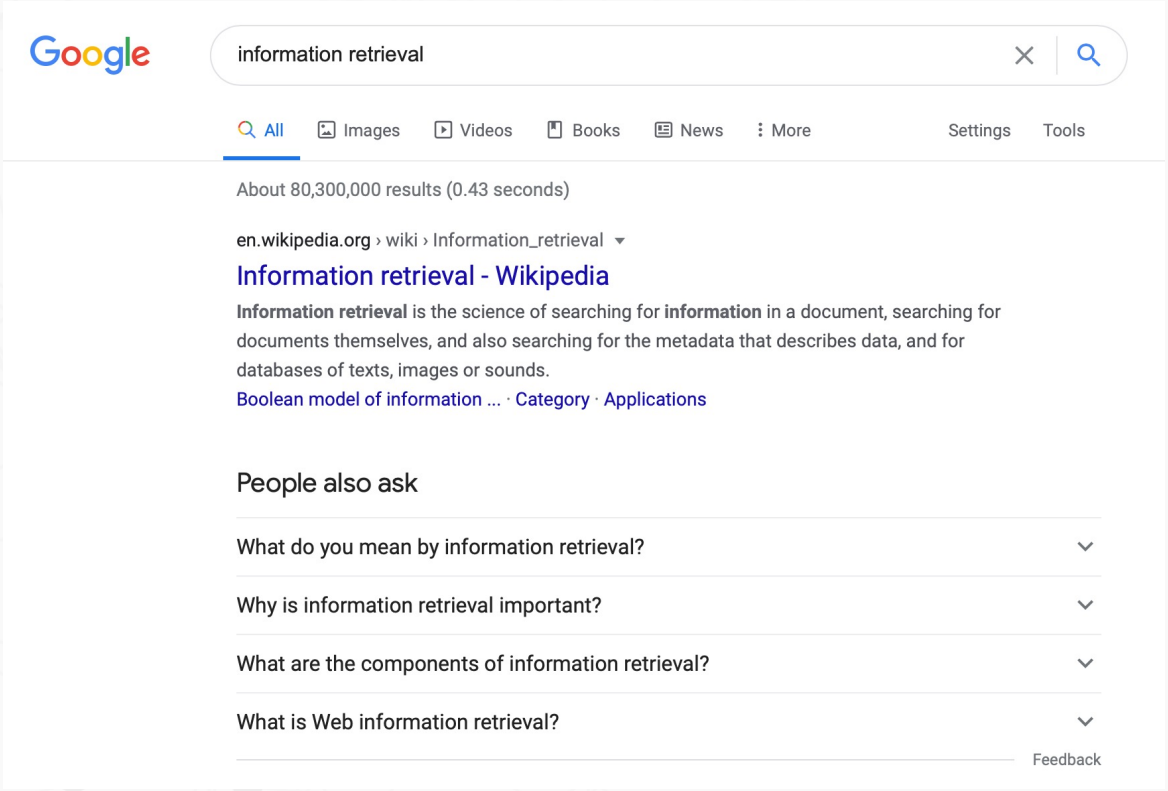


Information Retrieval

Problemática: Existe una gran cantidad de documentos y quiero encontrar los más similares a una query en específico

Ejemplos: búsqueda web, catálogos en una biblioteca, estudios similares

Data mining ‘interactivo’. No hay un query especificado



Otros análisis

Wordclouds: Resumir el contenido de un texto. Recibe un texto como entrada, encuentra las palabras más relevantes y las muestra de forma gráfica

Web mining: Utiliza información del uso y contenido de una web para obtener insights. Ejemplos: búsquedas web, recomendar el mejor producto siguiente, publicidad.

Opinion mining: Utiliza información de opiniones (subjetivas) sobre productos o entidades para poder entender el sentimiento general hacia este.

★★★★★ **Would Reccommend**

Calificado en Estados Unidos el 6 de abril de 2018

I have had this phone for about one year now and it has never failed me. It was been able to withstand many drops and has not broken. The battery lasts a full day at moderate use. Great phone!

A 2 personas les resultó útil

Representación y preprocesamiento de textos

Term – document Matrix

- Término: típicamente una palabra, pero puede ser una frase como ‘muy bueno’
- Documento: la entidad que tiene el texto a extraer

La cantidad de términos puede ser muy grande (¿cuántas palabras?)
La representación puede ser binaria o de frecuencia

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

```
D1 = (1, 1, 1, 0, 0, 0)
```

Cada documento ahora es un vector de términos (frecuencia o binario)

Stopwords

Muchas de las palabras más comunes en cualquier idioma no conllevan mayor significado

En inglés: the, of, and, to ...
En español: de, a, algún, por ...

También podemos construir un conjunto de palabras asociadas a una aplicación específica

¿Por qué necesitamos eliminar las stopwords del análisis?

- Reduce significativamente la cantidad de términos utilizados (~ 20% - 30%)
- Mejor eficiencia: Las stopwords no son útiles para búsquedas o text mining
- Las stopwords aparecen muy frecuentemente

Stemming

Es una técnica para encontrar la raíz (stem) de una palabra

Usuarios
Usuario
Usado
Usando

Stem: **Usar**

Engineer
Engineered
Engineer

Stem: **Engineer**

El Stemming es útil para:

- Mejorar la efectividad de la búsqueda y text mining
- Combinar las palabras con una misma raíz reduce considerablemente el tamaño del vocabulario utilizado (~ 40% - 50%)

Ver: Algoritmo de Porter para stemming, Lematización

IDF - Ponderación en el espacio TD

No todas las palabras o frases son igual de importantes.
- David es menos importante que Beckam

Si un término aparece frecuentemente en muchos documentos tiene menor poder de discriminación

Una forma de corregir esto es la frecuencia de documento inversa (IDF)

La importancia final del término es el producto de la matriz de frecuencias TF por la frecuencia inversa IDF.

Un término es importante si tiene alta TF y/o alta IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

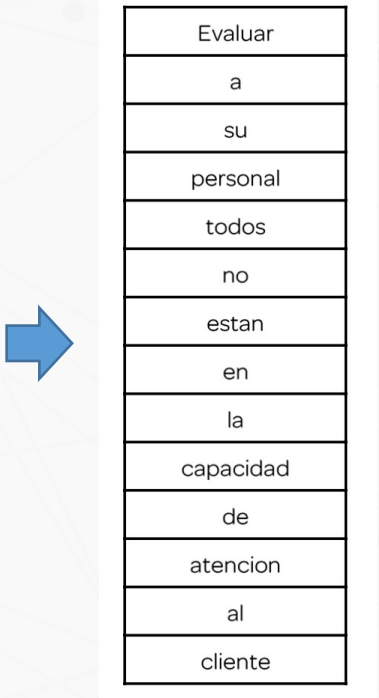
df_x = number of documents containing x

N = total number of documents

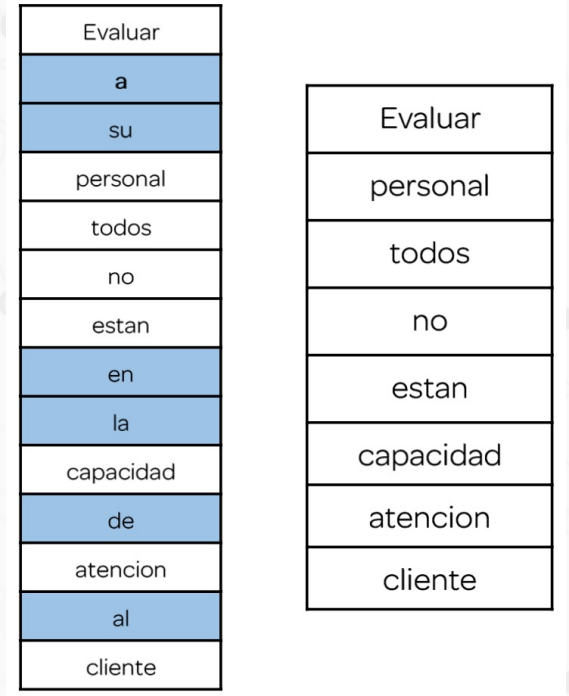
Resumen

Comentario

"Evaluar a su personal, todos no están en la capacidad de atención al cliente".



Tokenización



Stopwords y Stemming

Doc	Evalua	Person	Tod	no	Est
1	1	1	1	1	1
2	0	1	0	1	0
3	0	1	1	0	1

Matriz TF o TF-IDF

¡Gracias!

Online | DMC Perú