1. Analítica y análisis de textos, búsqueda y recuperación de información

Texto y multimedia → tipo de dato no estructurado → estructura impredecible.

``Datos no estructurados" son datos que no tienen una estructura clara, semánticamente abierta y fácil de usar. Por ello no se pueden almacenar en una base de datos tradicional o RDBMS. Casi ningún dato es realmente ``no estructurado´´.

Para poder analizar un texto tenemos que transformarlo en vectores, matrices o cosas parecidas (lenguaje que entienda el ordenador).

Podemos hacer algunas distinciones →

- La analítica de textos → proceso de análisis de texto no estructurado, se coge info importante, proporciona información cuantitativa.
- Minería o análisis de texto → se obtiene información cualitativa mediante texto no estructurado, proporciona información cualificada.

Búsqueda y recuperación de información; es encontrar material (datos no estructurados) para satisfacer unas necesidades.

El proceso del lenguaje natural (NLP)

Es una rama de la IA que ayuda a los ordenadores a entender, interpretar y manipular el lenguaje humano.

2. Tipos de tareas y aplicaciones

Podemos realizar varias acciones sobre un conjunto de documentos o textos.

- Indexación de documentos.
- Resumen de documentos o un grupo de ellos.
- Etiquetado.
- Recuperar documentos o textos de una base de datos.
- Clasificación de documentos en varios grupos.
- Búsqueda de los grupos o documentos de agrupación.
- Comprender el sentimiento de un texto y contexto.
- Procesamiento avanzado del lenguaje natural (NLP).

  Traducir a otro idioma, comprender y responder preguntas.

- Gestión de riesgos.
- Comentarios de los clientes, escucha social y aplicaciones centradas en las personas.
- Publicidad mediante medios digitales.
- Enriquecimiento de contenido.
- Filtrado de spam.
- Prevención de delitos cibernéticos.

③ Clasificación de textos, análisis de sentimiento y recuperación de información: Clasificador de Bayes

Tenemos un conjunto de clases → buscamos a qué clase pertenece el objeto.

· Temas → clases generales
Clasificador de texto → tarea de clasificar los textos.

* Usaremos el algoritmo de Bayes para clasificar texto.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \rightarrow \text{Método de aprendizaje probabilístico.}$$

Probabilidad de que un documento (d) esté en la clase (c);

$$P(c|d) = P(c) \times \prod P(t_k|c) \cdot (1)$$

$P(c|d)$ → d o cma en c $P(c)$ = probabilidad previa.

- Probabilidad previa

$$P(c) = \frac{Nc}{N\,docs} = \frac{N^o \text{ documentos clase } c}{N^o \text{ documentos en total}}$$

- Probabilidad condicional

$$P(c|d) = \frac{(Tct) + 1}{(T'ct) + (B)}$$

(Tct) → N° de veces que aparece t en los documentos de la clase C

(T'ct) → N° total de palabras (longitud) de cada clase.

(B) → N° de términos diferentes en el vocabulario.

Otras maneras de calcularlo.

1- Bernoulli
2- Máquina vectorial de soporte
3- Redes neuronales.

4. Recuperación de información: Modelo de espacio vectorial por puntuación.

Representación de un conjunto de documentos como vectores → modelo de espacio vectorial.

Proceso para puntuar y variar un conjunto de documentos: Pasos

- Caracterización de la base de datos de documentos
a) construcción de la matriz de frecuencias del conjunto de documentos.
- Ponderación de las palabras relevantes:
- Frecuencia inversa del documento.
- Término frecuencia del documento x matriz de frecuencia inversa del documento.
- Caracterización de la consulta: vector frecuencia.
- Clasificamos los documentos de la similitud más alta a la más baja.

Caracterización de la base de datos de documentos.

Frecuencia del documento → nº de veces que cada palabra se repite en estos documentos. Cuanto más frecuente más relevante. Como hay muchos elementos repetidos, necesitamos eliminar aquellos que no aporten. Para ello:

Frecuencia inversa del documento

$$idf_t = \log \frac{N}{df_t}$$

N → Número de documentos en la base de datos.

df_t → Frecuencia del término en todos los documentos

idf de un término raro → alto
idf de un término frecuente → bajo

Frecuencia inversa + Frecuencia del término

$tf \cdot idf$

Caracterización de la consulta → vector de frecuencias de término.

$tf \cdot idf$

matriz term frecuencia

Similitud d_1 y d_2 → calcular similitud del coseno

$$sim(d_1, d_2) = \frac{V(d_1) \cdot V(d_2)}{\|V(d_1)\| \cdot \|V(d_2)\|}$$

$V(d_1) \cdot V(d_2)$ → producto de puntos de los vectores d_1 y d_2

$\|V(d_1)\| \cdot \|V(d_2)\|$ → producto de las longitudes euclidianas de los vectores del documento. Normaliza la longitud del vector

a mayor ángulo → menor coseno = menos similitud entre el documento y la consulta