3.- Lenguaje Natural (Contenido de otro curso)

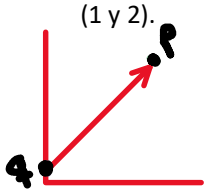
jueves, 22 de febrero de 2024 10:20 p. m.

Similitud y disimilitud de datos

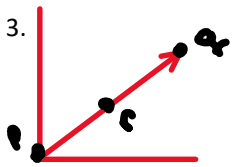
La similittud es una forma de medir que tan relacionadas están las observaciones de datos (samples/instances). Por otro lado, la disimilitud se refiere a que tan distintas son.

- La similitud generalmente expresada como un numero entre cero y uno. Por convención es uno pero puede ser cualquier número positivo (+infinito). Donde cero significa que no hay similitud o parecido y uno representa total similitud.
- La disimilitud también se expresa numéricamente en un rango entre cero y uno. Donde cero significa que no hay parecido entre las observaciones y el uno implica total similitud.

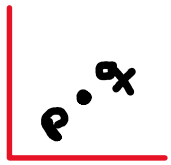
(1 y 2).



3.



4.



Asimismo, la proximidad puede usarse para referirse a similitud o disimilitud.
Por su parte una distancia (proximidad) dada tiene el fin de ser una métrica si y solo si satisface lo siguiente:
1.- No negatividad: d(p,q) >= 0, para cualquier observación de dos muestras p y q.
2.- Simetría d(p,q) = d(q,p) para toda "p" y "q".
3.- Inequidad triangular: d(p,q) <= d(p,r) + d(r,q) para toda "p", "r" y "q".
4.- d(p,q) = 0 si p = q.

Distancias:
La técnica (métrica) utilizada para medir la distancia o proximidad dependerá de cada caso específico sobre el que se está trabajando.

Vectorización: Es la transformación del lenguaje natural a un vector con valores numéricos que pueda entender la maquina (A esta etapa se le llama la transformación).

Limpieza: Mover ruido, remover palabras que se repiten frecuentemente (stop words) o sinónimos de palabras que generan la misma idea.

A los párrafos de texto se les conoce como documentos, al conjunto de documentos se le conoce como cuerpo (corpus).
Lo que se busca es pasar texto a forma numérica para que entienda la máquina, convirtiéndolo en un vector de N dimensiones.

Bolsa de palabras
Es un modelo donde tiene una "bolsa" donde están todas las palabras del documento o los documentos.

Vocabulario: Conjunto de palabras que existen en el documento y se colocan dentro de la "bolsa".

Técnica #1: One hot encodign.
Le importa si está o no está la palabra
Se debe de establecer la forma del vocabulario.
Se realiza una limpieza del stopword
Se realiza una matriz de representación de cada palabra que aparece en los documentos y se determina su relación entre el vocabulario y el documento con 0 o 1.

A los vectores que resultan de esta técnica se le conoce que son densos y esparcidos (dense & sparse), esto porque tienen muchas dimensiones, y los valores difieren entre documentos.

Técnica #2: Term Count (conteo de términos)
Le importa cuantas veces aparece la palabra.
Cuenta el número de veces que aparece la palabra en el documento con base al vocabulario del cuerpo.

Técnica #3: Probabilidad
Le importa la probabilidad de la palabra.
Mide la probabilidad de que aparezca la palabra del vocabulario en un documento, con respecto al total de palabras que hay en el cuerpo

Ejemplo de vectores de las diferentes técnicas.
One hote | 1 | 0 | 1 | 1 | 0 |
Row cou. | 1 | 0 | 5 | 3 | 0 |
Probabil. | 7/60 | 0 | 5/60 | 3/60 | 0 |

TF (Term Frecuencia)

TF = número de veces que el termino aparece en el documento/número de términos del documento

	Proyecto	
Doc1	1/15	doc1 = 15 ; Proyecto = 1
Doc2	1/13	doc2 = 12 ; Proyecto = 1

IDF (Inverse Document Frequence)
Se saca la representación del valor de cada palabra

IDF = log (# Total del documento / # Documentos que contienen el termino) + 1

IDFproyecto = log(2/2)
IDFinicial = log(2/1)

TF-IDF = TF * IDF

Tarea: Investigar conceptualmente que hace TF-IDF, con respecto a que le interesa saber o conocer, poner el foco en IDF

TF-IDF
Le interesa medir qué términos son más relevantes para el asunto, analizando la frecuencia con la que aparecen en una página, en comparación con su frecuencia en un conjunto más grande de páginas.
Aplicado en motores de búsqueda, sistemas bibliotecarios y la mineración de textos.

Importancia:
"La importancia del término (valor TF-IDF) aumenta de acuerdo con el número de veces que la palabra aparece en el documento (TF). Pero se compensa por el número de repeticiones en la colección de documentos (IDF), lo que sirve para ajustar el hecho de que algunas palabras aparecen con más frecuencia en general."