**Taller Preparación Parcial.**

**Curso de NLP**

**Prof. Rubén Manrique**

**Fecha de Entrega. 22 de Febrero**

**BY: JHON BUESAQUILLO - JUAN PABLO HERNÁNDEZ - JHOAN SAAVEDRA**

Considere el siguiente corpus documental:

D1: Yo tengo que estudiar para pasar el parcial.

D2: Si no practico Python no voy a avanzar en el curso.

D3: En clase estudiamos que era un modelo en bolsa de palabras.

D4: El modelo de bolsa de palabras requiere procesar el texto.

D5: Remover stopwords y normalizar son algunas de las tareas de procesar el texto.

D6: Para ponderar la importancias de las palabras en el modelo de bolsa de palabras se usa tfidf.

D7: Para calcular similitudes entre documentos debemos vectorizar el modelo de bolsa de palabras.

D8: Usando la distancia coseno podemos calcular las similitudes entre documentos.

D9: En python implementamos diferentes medidas de similitud y estudiamos diferentes métricas de evaluación.

D10: Entre las métricas más empleadas de evaluación se encuentra la medida F1.

**Archivo adjunto: “TallerPreparcial”**

**El punto 1 al 5 se realiza en Word (manual).**

1. Construya el modelo de bolsa de palabras.
   1. Considere las siguiente lista de stop-wprds: <https://github.com/Alir3z4/stop-words/blob/master/spanish.txt>
   2. Suponga el siguiente stemmer que define las siguientes equivalencias.
      1. estudiar = estudiamos =estudio
      2. similitudes = similitud
      3. Usando = usa
      4. Modelo = modelos = modelamiento
   3. Defina su vocabulario.
   4. Utilice una ponderación tf-idf (puede hacerla en una tabla excel).

**RT//** Para este punto se quitaron los stops words y puntuaciones. Los documentos quedaron de la siguiente manera:

D1: estudiar pasar parcial

D2: practico python voy avanzar curso.

D3: clase estudiar modelo bolsa palabras.

D4: modelo bolsa palabras requiere procesar texto.

D5: remover stopwords normalizar tareas procesar texto.

D6: ponderar importancias palabras modelo bolsa palabras usa tfidf.

D7: calcular similitud documentos debemos vectorizar modelo bolsa palabras.

D8: usa distancia coseno podemos calcular similitud documentos.

D9: python implementamos diferentes medidas similitud estudiar diferentes métricas evaluación

D10: métricas empleadas evaluación encuentra medida f1

El vocabulario,el tf- idf están en el excel “TallerPreparcial”

1. Vectorize su modelo de bolsa de palabras. En un excel indique el vector correspondiente a cada documento.

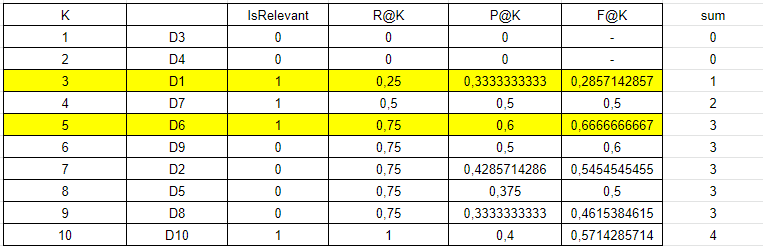
RT/Están en el excel “TallerPreparcial”

1. Calcule la similitud de los documentos con el query: “modelo de bolsa de estudio”.

RT// Quitando los stops words y puntaciones la query queda

query: modelo bolsa estudiar

1. Si los documentos relevantes para el query anterior son D6, D7, D10 y D1, calcule P@5, R@5, F1@5, P@3, R@3 y F1@3.



El procedimiento está en el excel

1. Suponga que existe una escala de relevancia D6:4, D7:3, D10:3, D1:5 (otros documentos relevancia de 0). Calcule el NDCG para los primeros 10 documentos.

Rt// Procedimiento está en excel



1. Verifique los resultados anteriores usando Python en Notebook (el notebook es un entregable).

Rt// Archivo “TallerParcial”

1. Suponga que los documentos D3, D6 y D7 son documentos etiquetados como spam, mientras que D1, D2, D4, D5, D8, D9 y D10 como NOspam. Construya un clasificador Naive Bayes y determine la clase de los siguientes documentos:
   1. “El modelo de estudio basado en Python funciona más que las palabras.”
   2. “En el parcial hay que calcular similitudes entre modelos”.