INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence
Primavera 2016

Profesores: Juan D. Velásquez
Gaspar Pizarro V.
Prof. auxiliar: Tomás Valdivia H.

Solución Control 1

Pregunta 2

1. Dados los siguientes documentos: *4 puntos*

   D1: *"Melania acusa a los medios de atacar injustamente a Trump por video"*

   D2: "Melania Trump: Donald fue incitado a los comentarios sexistas"

   D3: "Melania Trump: Mi esposo fue 'orillado a decir cosas sucias' por Billy Bush"

   D4: "Tras video polémico, Melania Trump defiende a su esposo"

   Determine cuál es el más parecido a D1. Para esto considere:

   I. Utilice como características uni-gramas (*Bag of Words*) de las palabras para calcular la matriz Tf-Idf.

   II. Compare mediante la similitud del coseno para determinar el documento más parecido.

   $$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \times |d_2|}$$

   III. Justifique cada decisión tomada.

   Supuestos válidos:

   - *Tokenización o eliminación de puntuación*
   - *Normalización de mayúsculas*
   - *Eliminación de stopwords*
   - *Lematización o stemming*

   La matriz tf-idf debe mostrar los supuestos usados (si dice que va a utilizar stemming o lematización que las palabras estén cortadas o en raiz, si elimina stopwords no deben haber stopwords, etc.). 0.5 puntos. Matriz tf-idf: Ver pauta en `P2.xlsx`*. 3 puntos. En caso de no haber consecuencia entre supuestos y matriz, asígnese máximo 2.5 puntos. Documento más similar a D1: D4. 0.5 puntos por cálculo de similitud de coseno.*

2. Sobre text mining:

   a) ¿Cuál es la ventaja de usar Idf en el cálculo de la matriz de términos? *1 punto*

   Permite disminuir la relevancia de términos que son usados en muchos documentos, creando una suerte de "stopwords de dominio específico", a la vez que aumentar la relevancia de términos que son ocupados en pocos documentos y pueden diferenciarlos aun más, incluso aunque en estos documentos sean ocupados pocas veces.

Profesores: Juan D. Velásquez
Gaspar Pizarro V.
Prof. auxiliar: Tomás Valdivia H.

b) Discuta brevemente las ventajas y desventajas de usar n-gramas con n pequeño y con n grande. *1 punto*

N pequeño. Ventajas: Más compacto, menor tiempo de procesamiento. Desventajas: No captura nombres o palabras compuestas. N grande. Ventajas: Permite capturar nombres o palabras compuestas. Desventajas: Aumenta el espacio vectorial, más tiempo de cómputo.