Examen RI Julio 2018 Soluciones

Recuperación da Información (Universidade da Coruña)

Examen **Recuperación de Información** Julio 2018

Apellidos: _____Nombre: _____

Examen sin libros, apuntes, ni dispositivos electrónicos.
Tiempo: 2h. 15m.

1. (2 puntos) Cada apartado de esta pregunta se puntúa con 0.2 puntos, pero **cada respuesta incorrecta invalida una correcta**. La pregunta no tendrá puntuación negativa. Responda cada pregunta con Verdadero/Falso
   i) Suponga que una colección de documentos se representa por una matriz términos x documentos (o lo que es lo mismo, palabras_únicas x documentos) que recoge en cada posición de la matriz el tf(t, d). La matriz ocupa un espacio de almacenamiento en bytes de número_términos x número_documentos x bytes_por_posición. Considere que se eliminan los tokens (una palabra que se repite dos veces en la colección son 2 tokens) que representan el 30% de tokens de la colección pero que se eliminan A) empezando por los tokens del término de orden 1 según la Ley de Zipf y siguiendo por los tokens del término de orden 2 y así consecutivamente hasta completar el 30% de tokens totales de la coleccíon B) empezando por el orden opuesto según la ley Zipf, es decir, empezando por el término que está en el extremo derecho de la cola de la Ley de Zipf y siguiendo por los términos hacia la izquierda hasta completar el 30% de tokens. Haciendo B) se reducirá más espacio en la matriz términos por documentos que haciendo A)
   RESPUESTA CORRECTA:____True_____
   ii) Se usa la media harmónica en la métrica F1 en lugar de la aritmética para poder promediarla para todas las queries
   RESPUESTA CORRECTA:_____False ___
   iii) La medida de similaridad de coseno es ampliamente usada en el modelo de espacio vectorial de IR por su menor complejidad computacional con respecto a la distancia euclídea.
   RESPUESTA CORRECTA:___False_____
   iv) Puede haber una query para la que P@2 (Precision@2) puede tener el valor 0.2.
   RESPUESTA CORRECTA:___Falso_____
   v) Dado un benchmark de evaluación de IR de gran tamaño (gran número de documentos, queries y juicios de relevancia), si el sistema A alcanza un MAP=0.33 y el sistema B un MAP=0.81, esto implica que la mejora del sistema B con respecto al sistema A es estadísticamente significativa.
   RESPUESTA CORRECTA:__ Falso___
   vi) Un search engine vertical (diseñado para búsqueda temática) el crawling debe hacerse con una estrategia de primero en profundidad
   RESPUESTA CORRECTA:____Falso_____
   vii) Con el modelo de IR de Language Models el smoothing no debe aplicarse a los documentos que contienen todos los términos de la query
   RESPUESTA CORRECTA:_____Falso ___
   viii) El algoritmo de procesamiento de consultas Document-at-a-Time puede adaptarse para computar el score con el modelo de Query Likelihood con suavización de Dirichlet
   RESPUESTA CORRECTA:_____Verdadero____
   ix) El algoritmo de procesamiento de consultas Document-at-a-Time puede adaptarse para computar el score con el modelo de Query Likelihood con suavización de Jelinek-Mercer
   RESPUESTA CORRECTA:___Verdadero_____
   x) Con el método de pooling, en la iniciativa TREC, con un esfuerzo asesor razonable, se consiguen etiquetar todos los documentos relevantes de las colecciones usadas en TREC.
   RESPUESTA CORRECTA:__ Falso___

2. (1 punto) Considere un grafo web con 4 nodos y los siguientes enlaces: 1-->2, 1-->4, 2-->1, 2-->2, 2-->3, 2->4, 3-->4, 4-->3, 4-->2.

   a) Compute la matriz de transición de probabilidad con teleporting del 80%
   b) Considere el modelo del random surfer de Page Rank. Si se conoce con certeza que en el estado inicial el random surfer está en el nodo 1, ¿cuál es la distribución de probabilidad para el estado siguiente?
   Debe indicar el resultado final y los cómputos.

Matriz de tranisición de probablidad

$$\begin{pmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

b) MTP con teleporting 80%

$$0.20\,x \begin{pmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix} \textbf{ + (0.80/4) x } \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0.1 & 0 & 0.1 \\ 0.05 & 0.05 & 0.05 & 0.05 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0.1 & 0.1 & 0 \end{pmatrix} + \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix} = \begin{pmatrix} 0.2 & 0.3 & 0.2 & 0.3 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix}$$

c)

$$\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \text{ x } \begin{pmatrix} 0.2 & 0.3 & 0.2 & 0.3 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix} = \begin{pmatrix} 0.2 & 0.3 & 0.2 & 0.3 \end{pmatrix}$$

3 . (1 punto) Considere un documento D en una colección. Suponga que se añade otro documento que es una copia de D a la colección. Considera que no se produce el caso de que existan palabras que aparecen en todos los documentos de la colección.
a) Para todas las palabras de la colección el idf logarítmico, idf(w), aumenta
b) Para todas las palabras de la colección el idf logarítmico, idf(w), disminuye
c) Para las palabras de D el idf disminuye y para el resto aumenta
d) Para las palabras de D el idf aumenta y para el resto disminuye
e) Ninguna de las anteriores es cierta

Respuesta correcta: _____C_____
Para puntuar la pregunta debe ser correcta la respuesta y el razonamiento aportado.

Pista: Escriba la fórmula de idf log y razone lo que pasa

N, es el número de documentos de la colección. df(w) es la frecuencia de la palabra w en la colección

idf(w) = log (N+1 / (df(w)+1)), si w está en D
= log (N+1/ df(w)), si w no está en D

Si w está en D, idf disminuye porque N es mayor o igual que df(w) (no ser que una palabra aparezca en todos los documentos, N=df(w), en cuyo caso quedaría igual.
Si w no está en D, idf aumenta porque el numerador aumenta y el denominador sigue igual

Por tanto la correcta es la c).

4. (1 punto) Considere una colección de documentos y un documento D. Considere que el documento se cambia duplicando su contenido. Considere modelos de lenguaje con suavización de Jelinek-Mercer. Considerando el antes y después del cambio en D,
a) El modelo de lenguaje del documento D queda igual
b) El modelo de lenguaje del documento D cambia
c) El modelo del lenguaje del documento D cambia sólo para los términos de la consulta
d) El modelo del lenguaje del documento D cambia sólo para los términos que no están en la consulta

Respuesta correcta: __B____
Para puntuar la pregunta debe ser correcta la respuesta y el razonamiento aportado.

Pista: haga las cuentas con un ejemplo pequeño

Imagine dos documentos: d1 con 2 ocurrencias de w y 10 en total; d2 con 3 ocurrencias de w y 10 en total. Al duplicar el contenido de d2 el MLE de las palabras de d2 queda igual, pero ya se ve que para w la probabilidad en la colección cambia. Antes la probabilidad en la colección para w era 5/20 = ¼ y ahora pasa a ser 8/30 = 4/15 > ¼.
Por tanto la respuesta correcta es la b). Las respuestas c y d no tienen sentido ya que un modelo de lenguaje es una distribución de probabilidad que se define para todas las palabras del vocabulario.