Examen: Inteligencia Artificial
19 de diciembre de 2012

Considere:

- Puede utilizar una calculadora para desarrollar los ejercicios.
- El uso de un aparato electrónico móvil está prohibido.
- Si es sorprendido copiando, se le asginará la nota 1.
- Se prohibe el uso de cualquier tipo de apunte durante el examen.

1. Recuperación de Información (25 puntos)

Para desarrollar las preguntas de esta sección considere los siguientes cuatro fragmentos de texto y sus respectivos bag-of-words (BoW):

doc0::"*El flamante técnico de la selección chilena, Jorge Sampaoli, cumplió con su segundo objetivo en su gira por Europa y este jueves concretó su reunión con el técnico de FC Barcelona, Tito Vilanova, para hablar respecto a Alexis Sánchez.* "

{Alexis=1, Barcelona=1, Europa=1, FC=1, Jorge=1, Sampaoli=1, Sánchez=1, Tito=1, Vilanova=1, chilena=1, concretó=1, cumplió=1, flamante=1, gira=1, hablar=1, jueves=1, objetivo=1, respecto=1, reunión=1, segundo=1, selección=1, técnico=3}

doc1::"*El técnico de FC Barcelona, Tito Vilanova, aseguró este martes que se reunirá con el flamante entrenador de la selección chilena, Jorge Sampaoli, en la gira que el argentino está realizando por Europa.*"

{Barcelona=1, Europa=1, FC=1, Jorge=1, Sampaoli=1, Tito=1, Vilanova=1, argentino=1, aseguró=1, chilena=1, entrenador=1, flamante=1, gira=1, martes=1, realizando=1, reunirá=1, selección=1, técnico=1}

doc2::"*Sin cifras oficiales confirmadas y en medio de opiniones divididas por las ofertas que estuvieron disponibles, las empresas asociadas con la Cámara de Comercio de Santiago para el nuevo Cyber Monday chileno ya sacan cuentas optimistas*. "

{Comercio=1, Cyber=1, Cámara=1, Monday=1, Santiago=1, Sin=1, asociadas=1, chileno=1, cifras=1, confirmadas=1, cuentas=1, disponibles=1, divididas=1, empresas=1, estuvieron=1, medio=1, nuevo=1, ofertas=1, oficiales=1, opiniones=1, optimistas=1, sacan=1}

doc3::“*Un positivo balance realizó la Cámara de Comercio de Santiago sobre la segunda versión del Cyber Monday realizado en Chile a través de un grupo de empresas adheridas a su Comité de Comercio Electrónico. ”*

{Chile=1, Comercio=3, Comité=1, Cyber=1, Cámara=1, Electrónico=1, Monday=1, Santiago=1, adheridas=1, balance=1, empresas=1, grupo=1, positivo=1, realizado=1, realizó=1, segunda=1, través=1, versión=1}

Preguntas:

1. Muestre el Indíce Invertido no-posicional para una colección que considera sólo los documentos doc0 y doc1. (5 puntos)

Alexis=1:1
Barcelona=2:1,2
Europa=2:1,2
FC=2:1,2
Jorge=2:1,2
Sampaoli=2:1,2
Sánchez=1:1
Tito=2:1,2
Vilanova=2:1,2
argentino=1:2
aseguró=1:2
chilena=2:1,2
concretó=1:1
cumplió=1:1
entrenador=1:2
flamante=2:1,2
gira=2:1,2
hablar=1:1
jueves=1:1
martes=1:2
objetivo=1:1
realizando=1:2
respecto=1:1
reunirá=1:2
reunión=1:1
segundo=1:1
selección=2:1,2
técnico=2:1,2

2. ¿Cuál sería el número recomendado de skip pointers? (2 puntos)

sqrt(1)=1 y sqrt(2)=1.4

3. ¿Qué documentos recuperaría la consulta: "*martes*" AND "*jueves*" ? (3 puntos)

Ninguno

4. ¿Cómo se construiría el índice Permuterm y k-gram de "*hablar*"? ¿Qué problema observa?: (5 puntos)

hablar={hablar$,ablar$h,blarha,larhab,ar$habl,r$habla,r$habla}
hablar={ha,hab,abl,bla,lar,ar}

El índice crece substancialmente. El PermuTerm es más grande que el k-gram.

5. ¿Cómo se resolvería una consulta ha*ar con un índice k-gram (k=3)? ¿Qué problema avisora en una colección más grande?: (5 puntos)

$ha AND ar$

El problema que se observa es que puede hacer matching con palabras que no están relacionadas con el query $haXXXXar$, XXXX=cualquier cosa.

6. Calcule la ley de Heap para la colecciónconlos documentosdoc0 y doc1. Utilice los valores típicos para los parámetros. (5 puntos)

M=44*pow(39,0.5)=274.78

2. Semántica Distribucional (25 puntos)

1. De un ejemplo de las siguientes relaciones, e indique si son simétricas: (5 puntos)

a) Hiperonimia: fruta -> pera, manzana, sandía...
b) Cohiponimia: pera <-> manzana
c) Meronimia: hoja -> árbol
d) Parónimos: afecto <-> efecto

2. ¿Cuál es la diferencia entre la meronimia y la hiponimia? (2 puntos)

Una es conceptual (hiponimia) y la otra tiene relación al mundo físico (meronimia).

3. ¿Cuál es la idea detrás de LSA? (8 puntos)

Representar un modelo/matriz de vectores documentos vs. términos en un espacio latente formado por las componentes principales. De esta forma, se permite reducir el ruido y hacer inferencias semánticas.

4. Compare LSA y LRA. (10 puntos)

La principal diferencia es que LSA se enfoca a las palabras, i.e., sinónimos, y LRA a las relaciones, i.e., analogías. Ambas utilizan SVD para operar, pero la representación de LSA es documentos versus términos, y LRA pares de palabras versus patrones.

3. Aprendizaje No-Supervisado (25 puntos)

1. Compare las estrategias de clustering jerarquico aglomerativo y devise. (5 puntos)

El primero es bottom-up, y el segundo es top-down. Ambos construyen un árbol, pero el primero uniendo datos y el otro separando. Ambos utilizan una métrica de distancia, y pueden ocupar las mismas políticas para decidir unir/dividir datos.

2. Calcule las distancias de Manhattan, eudlideana y coseno para TODOS los documentos vistos en la primera parte. (10 puntos)

Manhattan (1,0)=2.0
Manhattan (2,0)=0.0
Manhattan (2,1)=0.0
Manhattan (3,0)=0.0
Manhattan (3,1)=0.0
Manhattan (3,2)=2.0

Euclideana (1,0)=2.0
Euclideana (2,0)=0.0
Euclideana (2,1)=0.0
Euclideana (3,0)=0.0
Euclideana (3,1)=0.0
Euclideana (3,2)=2.0

Coseno (1,0)=0.6024640760767093
Coseno (2,0)=0.0
Coseno (2,1)=0.0
Coseno (3,0)=0.0
Coseno (3,1)=0.0
Coseno (3,2)=0.3344968040028363

3. ¿Qué problema existe para las métricas de distancia si un feature va de 0 a 1 y el otro de 0 a 1000? Ejemplifique. (10 puntos)

El problema es que la distancia está totalmente dominada por el feature que tiene ordenes de magnitud mayores, por ejemplo dos puntos (0.1,20) y (0.9,720), la distancia es 700.000457. Lo que hay que hacer es normalizar, lo que dejaría el 20 en 0.02 y el 720 en 0.8, y la distancia en 1.063.

4. Aprendizaje Supervisado (25 puntos)

1. Calcule la entropía si una de las dos clases tiene el 80% de los datos. (2 puntos)

0.2*log2 0.2 - 0.8*log2 0.8 = 0.722

2. ¿Por qué son necesarios los kernels cuando se construyen máquinas de soporte vectorial? (3 puntos)

Porque no todos los problemas no son linealmente separables, por ende, no se puede aplicar eficientemente SVM. Los kernels ayudan a resolver este problema.

3. ¿Cómo se tratan los atribnutos contiguos? ¿Qué problema existe? (5 puntos)

Se discretizan de acuerdo a intervalos definidos por los valores observados. El problema que existe es que hay que ordenar los valores y eso hace más lento el proceso de aprendizaje.

4. ¿Qué se hace con los valores “perdidos”? ¿Qué problema existe? (5 puntos)

Se utiliza una etiqueta “UNKNOWN” o bien el valor más frecuente. En caso de atributos contiguos, se utiliza la media.

5. Dada la siguiente matriz de confusión, calcule: Recall, Precision, Accuracy, y F(1)-Score, con respecto a la clase positiva. Comente. (10 puntos)

	Clasificados como Positivo	Clasificados como Negativo
Etiquetados Positivos	123	789
Etiquetados Negativos	987	321

Recall = 123/(123+987)=0.110810811
Precision = 123/(123+789)=0.134868421
Accuracy = (321+123)/(789+987+321+123)= 444/2220=0.2
F(1)-Score = (0.110810811*0.134868421)/(0.110810811+0.134868421) = 0.014944879/0.245679232=0.060830861

Estos resultados representan un clasificador de dos clases con un mal desempeño, ya que la Accuracy es muy baja. Además, vemos que entregaría muchos resultados malos en un motor de recuperación de información ya que tiene un bajo recall de la clase positiva. Es decir, encontraría poco de los documentos buscados. La lista de resultados estaría poblada de resultados irrelevantes a la consulta.