Universidade do Minho

Mestrado em Engenharia Biomédica

Ano Lectivo de 2011/2012

Bibliotecas Digitais

Frequência- Duração 3 horas

1. Considere o seguinte grafo Web: a página A aponta para C e D; a página B aponta para A e C; a página C aponta para B; a página D aponta para E; a página E aponta para A. **(6 valores)**
    a. Calcule a matriz de adjacência para este grafo
    b. Calcule a matriz de probabilidade para este grafo. A probabilidade de *teleporte* é a=0,5
    c. Suponha que um passeio aleatório comece na página C. Apresente um vector de probabilidade para esta página.
    d. Calcule uma aproximação para o *PageRank* das páginas deste grafo usando 3 iterações.

2. Considere a seguinte uma colecção de 10000 documentos que contém d1,d2,d3 com as seguintes características. **(4 valores)**

Termo	d1	d2	d3	df
actor	12	35	55	123
movie	15	24	48	240
trailer	52	13	12	85

    a. Calcule a representação vectorial para esses documentos usando pesos *tf-idf* e normalização euclideana.
    b. Calcule a similaridade do coseno entre os documentos
    c. Se tivesse duas classes como agruparia os documentos com base no 1NN?

3. Assuma que queremos classificar documentos de Ciências da Computação nas seguintes três categorias: *Systems, Theory, AI*. Considere que estamos a fazer uma classificação *Naive Bayes* simples em que apenas verificamos se as palavras significativas ocorrem ou não no documento. Foram estimadas as seguintes probabilidades analisando um corpus de treino pré-classificado.

c	Systems	Theory	AI
P(c)	0.35	0.40	0.25
P(theorem|c)	0.05	0.80	0.10
P(search|c)	0.30	0.40	0.60
P(heuristic|c)	0.05	0.01	0.50
P(disk|c)	0.30	0.02	0.01
P(data|c)	0.50	0.01	0.20

Assumindo que a probabilidade de cada palavra considerada como evidência na classificação é independente dada a categoria do texto, classifique os pequenos textos a seguir apresentados. Assuma que as categorias são disjuntas e completas para esta aplicação. As palavras que não estão na tabela devem ser ignoradas. **(4 valores)**

D1: Data on heuristic search for theorem proving
D2: Search for data stored on disk

4. Considere as seguintes respostas a uma determinada interrogação para a qual existem 8 documentos relevantes na colecção. **(6 valores)**

S1: RRRNRNNRNR
S2: RNRNRNRNRR

a. Calcule a Precisão e o *Recall* da resposta
b. Calcule a Precisão interpolada aos 55% de *Recall* para cada sistema
c. Calcule o *MAP (Mean Average Precision)* e o *R-Precision* para cada sistema
d. Qual dos dois sistemas é mais eficaz? Porquê?
e. Considere para a interrogação acima dois juízes analisaram 40 documentos da colecção com os seguintes resultados . R significa relevante e NR não relevante. O nível de concordância dos juízes é aceitável? Justifique.

Nº Docs	Juiz 1	Juiz 2
6	R	R
4	R	NR
3	NR	R
27	NR	NR