Information Retrieval and Data Mining

Prof. Marco Tagliasacchi
Prof. Matteo Matteucci

July, 8 2015

Very Important Notes

- Answers to questions 1, 2, and 3 should be delivered on a different sheet with respect to 4 and 5
- If you need a calculator this should not be to any extent programmable or network connected

1. **Question (8 pts)**: Consider a set of $N = 4$ objects, a, b, c, d, and $M = 5$ annotators providing preference judgments about the ordering of the objects

u_1	u_2	u_3	u_4	u_5
a	a	b	a	c
c	b	a	b	b
b	c	d	d	a
d	d	c	c	d

Find the aggregated ranking using the following methods

(a) Borda's count

(b) Condorcet's winner

(c) Kemeny's rank aggregation (using Kendall tau distance). For the sake of the exercise, consider only candidate orderings which rank a as top-1 and d as top-4.

(d) Median rank aggregation.

2. **Question (6 pts)**: Describe the following graph centrality indices: betweennes, closeness and harmonic centrality. Illustrate by means of a small-scale example how to compute these indices.

3. **Questions (5 pts - each statement can be either TRUE or FALSE)**

(a) Let U denote the following term-topic matrix computed with LSI, where U_{ij} denotes the relevance of term i in topic j. Assume $\sigma_1 = \sigma_2 = 1$.

$$U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{1}$$

Let $q = [0, 1, 0, 0]^T$, $a = [1, 1, 0, 0]^T$, $b = [0, 1, 0, 1]^T$.

- $\boxed{\text{T}}\boxed{\text{F}}$ In the original term-document space, document a is closer to the query than document b according to cosine similarity.
- $\boxed{\text{T}}\boxed{\text{F}}$ The term-topic representation of document a is $[1, 0]^T$.
- $\boxed{\text{T}}\boxed{\text{F}}$ In the topic space, document a is closer to the query than document b according to cosine similarity.
- $\boxed{\text{T}}\boxed{\text{F}}$ Adding new documents to the dataset, the matrix U needs to be updated.

(b) Consider a multi-dimensional indexing scheme based on kd-trees.

- $\boxed{\text{T}}\boxed{\text{F}}$ A kd-tree is a binary tree.
- $\boxed{\text{T}}\boxed{\text{F}}$ The number of leaf nodes in a kd-tree is equal to the number of vectors to index.
- $\boxed{\text{T}}\boxed{\text{F}}$ Given a set of vectors to index, the kd-tree has a unique construction.

(c) Consider the construction of an inverted index.

- $\boxed{\text{T}}\boxed{\text{F}}$ The adoption of stemming increases precision and decreases recall.
- $\boxed{\text{T}}\boxed{\text{F}}$ Accessing the dictionary costs $O(M)$, where M is the number of terms.
- $\boxed{\text{T}}\boxed{\text{F}}$ The number of terms to be indexed grows approximately as $\sqrt{N}$, where N is the number of documents.

4. **Question (8 pts)**: Consider the following dataset composed by 4 binary attributes $\{A, B, C, D\}$ and a binary output Y

A	B	C	D	Y
0	1	1	0	1
1	1	0	1	0
1	0	1	0	1
1	0	0	1	1

(a) Build the decision tree to predict the output Y from selected attributes by using the *Information Gain* criterion[1]

(b) Describe the difference between the *Gini Index* and the *Information Gain Ratio* with respect to *Information Gain*

(c) Describe the procedure to split on real valued attributes using *Information Gain*

5. **Question (5 pts)** Answer the following questions:

(a) What is the goal of frequent pattern mining (a.k.a. association rules mining)?

(b) How *Support* and *Confidence* are defined?

(c) Describe the a-priori algorithm for frequent itemset generation

[1] You might need to know that $\lim_{x \to 0} x \ln(x) = 0$, $\ln(1) = 0$, $\ln(2) = 0.69315$, $\ln(3) = 1.0986$, $\ln(4) = 1.3863$, $\ln(5) = 1.6094$, $\ln(6) = 1.7918$.