TÉCNICO LISBOA

PRI 2021/22

Exam 1 - Version **A**

PART I [15.3v]

Consider the following collection D with 5 documents, 3 terms, term frequency (TF) entries, and query-independent scores by an IR system

	blue	*whale*	*balloon*	*IR score* (R*elevance*)
d1	3	0	3	0.4 (R)
d2	0	1	2	0.5 (R)
d3	2	3	1	0.3 (NR)
d4	1	**?**	0	0.2 (NR)
d5	1	2	0	0.1 (NR)

An referenced expert ordered documents by relevance, d1 > d2 > d3 > d5 > d4, where d1, d2 and d3 were seen as relevant, and d4 and d5 irrelevant.

1. [1v] Estimate **?** range of possible values for the two following scenarios:

a) [0.5v] ({d3,d4,d5},{*blue*,*whale*,*ballon*}) is an order-preserving concept with no noise.

b) [0.5v] Collection frequency of *whale* is 10.

IMPT: For all the remaining questions consider **?=1**

2. [2.2v] Assuming an integer takes 4 bytes and a pointer takes 3 bytes, identify the size of the postings in the *inverted index* of D (ignore the dictionary part) knowing:

a) [0.7v] the inverted index is non-positional and the term frequency per document is recorded

b) [0.7v] the inverted index is positional (term frequencies not recorded)

c) [0.8v] the inverted index is positional and the presence of gap-based encodings knowing that the current collection is a *dynamic collection* with no more than 30 documents and no more than 300 tokens per document

3. [3.5v] Let us now assess the performance of the given IR system

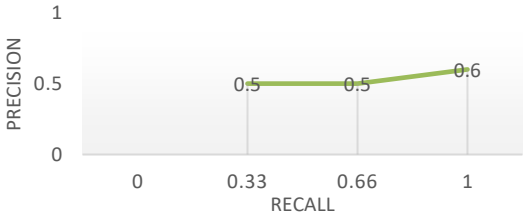
a) [0.6v] Draw the confusion matrix

b) [0.6v] Identify the MAP of the system.

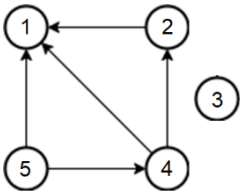
c) [0.7v] Compute Kendall tau to assess whether rankings are concordant with reference ones?

d) [0.8v] Do the IR system and expert ratings agree by chance? Use the *kappa* statistic to answer this question.

e) [0.8v] Considering a second IR system with the following precision-recall curve. Identify which of the 5 documents are seen as relevant. Justify.



4. [1.2v] Consider that the given five documents have the following link structure.



Compute the PageRank score for each page in the graph, assuming a uniform teleportation step and a weight α=1/2. Let disconnected documents to have fixed/unchanged uniform probability, 1/N. Fix *sinks*. Consider *one* iteration in the algorithm for computing the scores, starting with an initial uniform vector.

5. [0.6v] Compute the quality of the formal concept B=({d1,d3,d5},{*blue*, *balloon*}) under binarization threshold of 1.5.

6. [1v] Consider the expert feedback as magnets in SMART Rocchio with $\alpha = 1$, $\beta = 1, \gamma = 0$. Identify the modification to the query $q = \{whale\}$.

7. [1.6v] Given documents d6="*blue whale*" and d7="*ballon*":

a) [1v] Estimate d6 and d7 scores by the target IR system using a kNN regressor with k=3, simple Jaccard distance, and median estimator.

b) [0.6v] Knowing the system scores both as 0.5, compute the RMSE.

8. [2.2v] Consider q = "*blue blue*" and classic inverse document frequency, IDF= $\log_{10} \frac{N}{n_i}$:

a) [1.2v] Rank d3 using BM25 (k=1.2, b=0.75).

b) [1v] Rank d3 using TF-IDF under the ***ntn.lnn*** scheme
(note: if you do not recall ntn.lnn use another schema to get 50% of grading)

9. [2v] Assuming agglomerative clustering with minimum/single linkage and distance given by the sum of TF differences:

c) [1.2v] Showing the pairwise distance matrix, plot the dendrogram with the documents in D.

d) [0.8v] Given ({d1,d2,d3},{d4,d5}) solution, compute the silhouette of the smaller cluster.

PART II [4.7v]

1. [0.3v] Given a large collection C, sort by frequency in ascending order: i) *words* in C, ii) *terms* in C, iii) *tokens* in C, and iv) *topics* in C.

2. [0.3v] Identify an IR task that benefits from pseudo relevance feedback.

3. [0.5v] Given a collection, consider three terms appearing in $2n$, $3n$ and $10n$ documents, respectively. Considering an inverted index, how many operations are necessary to answer a Boolean query with these three terms?

4. [3.6v] True or False (+0.18v correct, –0.09v wrong)

a. Documents in the vector space model are generally dense vectors
b. A topic is a (multinomial) distribution over the terms in vocabulary
c. In latent Dirichlet allocation (topic modelling), the β parameter controls the topic-word density
d. Concept lattices can be used to infer ontologies for content categorization
e. In contrast with the minimum link, hierarchical clustering with a maximum link tends to break large clusters, generally yielding clusters with a more balanced number of documents
f. The *robots.txt* file specifies a list of pages to be indexed by a crawler
g. Locality sensitive hashing (min hash) is used to efficiently detect near-duplicate documents
h. Front queues enforce politeness
i. The Soundex algorithm can be used for spelling correction
j. The word “jaguar” can be subjected to asymmetric expansion in the normalization phase
k. One of the aims of lemmatization is to reduce inflectional morphology
l. Co-occurrence statistics are relevant for producing entries in the thesaurus
m. In query expansion, a false negative is a pair of unrelated words in the thesaurus
n. k-gram similarity with k=3 between *porta* and *cortas* is 0.6
o. $'m'$ and $'n'$ letters are more distant in classic edit distances than weighted edit distances
p. Proximity queries can be answered using biword indexes
q. Skip pointers are more elicited when intersecting dissimilar posting lists than when intersecting with similar posting lists
r. Variable length encoding on postings is a lossless form of compression
s. In collaborative filtering (CF), cosine is a good measure to assess similarity between items
t. CF is challenged by the high density (lack of missings) in rating matrices

END ☺