TÉCNICO LISBOA

PRI 2022/23

Exam 1

Duration: 2h

Part I (141pts)

Consider a static collection D with the following 3 documents after stop word removal, term normalization, and term selection:

d1: "grass stone ocean stone"
d2: "ocean grass ocean ocean"
d3: "unicorn stone grass ocean"

1. [14pts] Draw the positional inverted index.
Use gap-based encoding on the positional part of the inverted index.
Annotate terms with both document frequency and collection frequency.

2. [14pts] Consider the drawn positional index. Assume pointers are stored using 3 bytes and efficient typing for integers. What is the size of the posting lists (ignore the dictionary part)?

3. An expert ordered the documents by relevance d1 (relevant) < d2 (relevant) < d3 (irrelevant).

a. A user posted a query q that returned {d2,d3} documents.
   i. [10pts] Compute the confusion matrix and identify the balanced F-measure.
   ii. [6pts] Knowing that coverage is 0.5, is d1 known to the user? Show your calculus

b. An IR system assigned the following relevance score (i.e. the higher, the better) per document: s(d1)=0.4, s(d2)=0.2, s(d3)=0.6.
   i. [12pts] Calculate *interpolated* MAP, i.e. values taken from the *interpolated* PR curve.
   ii. [8pts] Assess agreement against reference truth using Kendall.

4. Given $q = \{stone, ocean\}$

a. [8pts] Identify the selected documents using 2NN with Jaccard-based similarity. Show all calculus

b. [11pts] Rank document d1 using TF-IDF under the nnn.ntc scheme
(note: if you do not recall nnn.ntc use another schema to get 50% of grading)

c. [11pts] Rank d1 using BM25.

5. [8pts] Considering the pairwise document distances d(d1,d2)=0.2, d(d1,d3)=0.5, d(d2,d3)=0.4. Identify the distance at which the three documents are aggregated in a single cluster using agglomerative clustering with: i) min (single) linkage, and ii) max (complete) linkage.

6. Consider a 100% quality in concept analysis, and concept $B = (\{d1, d2, d3\}, \{grass, whale, ocean\})$.

a. [8pts] Is B a coherent concept? If yes, under which coherence assumptions? If not, justify.

b. [7pt] Can B be formal concept? If yes, identify the range of valid binarization thresholds.

7. Consider that the query $q = \{whale\}$ returned d1 and d2, after which the user interactively selected d1 as relevant and d2 as non-relevant. SMART Rocchio (under $\alpha = 1$, $\beta = 1$)

a. [11pts] After modification, $q' = \{grass = 0.5, whale = 3, ocean = -0.5\}$. Is positive feedback more relevant than negative feedback? Justify with calculus.

b. [4pts] The modified queries can hamper efficiency and interpretability of IR systems. Briefly indicate the underlying unique reason for this.

8. [9pts] Consider the document links {d1→d2, d2→d3, d2→d1}, compute the normalized hub and authority scores for each document given by the HITS algorithm after one iteration.

Part II (59pts)

Important note: the open questions in this part are *objective*. Provide *clear* and *compact* answers.

1. [10pts] Consider the following four data structures: i) trees, ii) variable-length arrays, iii) linked lists, and v) hash tables. Which of them are generally used to represent: a) *dictionaries*, and b) *posting lists*? For each assignment, identify one advantage.

2. [7pts] If classifiers provide a deterministic output (e.g. relevant or non-relevant), how can they be used to rank documents?

3. [3pts] Identify one source of indirect relevance feedback: _____

4. [8pts] Consider an IR system that upon being queried with "aircraft" is unable to return documents with "airplane".

In the absence of feedback, identify two strategies to handle this problem.

5. Consider a clustering solution with two clusters, {d7} and {d1,d2,d3,d4,d5,d6,d8,d9,d10}.

a) [3pts] Select which agglomerative algorithm more probably return this division:

□ minimum/single linkage □ average linkage

b) [6pts] Knowing that {d1, d2} belong to a specific class and the remaining documents to another. Identify the purity of the clustering solution.

6. [22pts] Annotate each statement as **True** or **False** (+1.7pt correct, –0.3pt wrong)

1) Ranking searches are susceptible to the "feast and famine" problem: queries often result in either too few or too many results
2) In a classic TF-IDF stance, "John is quicker than Mary" and "Mary is quicker than John" are seen as identical statements.
3) IDF has no effect on ranking documents against single term queries
4) In latent Dirichlet allocation (topic modelling), topic-term density can be controlled
5) A thesaurus commonly stores near-synonym terms using their grammar relationships
6) Pearson coefficient is generally preferred over Spearman to assess non-linear correlations
7) The distance between *orta* and *hortas* is 2 using Levenshtein edit distance, considering both classic and Damerau variant
8) The noun phrase associated with "London to Washington" is a bi-word
9) The use of positional indexes to answer lengthy phrase queries is susceptible to false positives
10) Page Rank outputs a single score for each document in the base set
11) Page Rank is sensitive to the term content of a document
12) DNS resolution has a significant impact on crawling speed
13) A back queue in crawlers is generally implemented as FIFO (first in, first out)

END