THE UNIVERSITY OF
MELBOURNE

Student Number: ________________

Faculty: Computing and Information Systems

Subject Number: COMP90042

Subject Name: Web Search and Text Analysis

Writing Time: 2 hours

Reading Time: 15 minutes

Open Book Status: Closed book

Number of Pages: This paper has 5 pages including this cover page.

Authorized Materials: None

Instructions to Students: Every question and every part of each question should be attempted. A suggested amount of time for each question is shown next to the question.

Instructions to Invigilators: Invigilate vigilantly!

Paper to be held by Baillieu Library: Yes

Extra Materials Required: None

Question 1:

a). Consider the following Python function skeleton:

```
def tfidf_norm(dfdt, df, ndocs):
    """
Calculate unit-length-normalized TFIDF value for document.

Arguments:
  - dfdt: a dictionary of { term : fdt }, giving the
      fdt for each term in the documet
  - df: a dictionary of { term : ndocs }, giving the
      number of documents that each term occurs in
  - ndocs: the total number of documents in the collection
      (an integer)
Return value:
  Returns a { term : w_dt } dictionary, giving the
     unit-length normalized TFIDF weight of each term
     in the document.
    """
    pass
```

Implement this function. Use the TF and IDF formulae specified for Project 1. [15 minutes]

b). Let $\vec{q}_o$ be the original query vector, and $D_r = \{\vec{d}_1, \ldots, \vec{d}_i, \ldots, \vec{d}_r\}$ the document vectors of the top r documents returned for $\vec{q}_o$. What is the formula for the expanded pseudo–query-document, $\vec{q}_e$, in Rocchio pseudo-relevance feedback? (You may use $+$ to represent vector addition; that is, $\vec{a} + \vec{b} = \{a_1 + b_1, \ldots, a_i + b_i, \ldots, a_n + b_n\}$.) [10 minutes]

[25 minutes in total]

Question 2:

Consider the following two formulae:

$$\mathbf{X}_{t\times d} = \mathbf{T}_{t\times t}\mathbf{\Sigma}_{t\times d}\left(\mathbf{D}_{d\times d}\right)^T \quad (1)$$

$$\widehat{\mathbf{X}}_{t\times d} = \widehat{\mathbf{T}}_{t\times k}\widehat{\mathbf{\Sigma}}_{k\times k}\left(\widehat{\mathbf{D}}_{d\times k}\right)^T \quad (2)$$

Equation 1 indicates the matrix operation known as singular value decomposition; and Equation 2 is a reduced-rank representation of the SVD.

a). What is the name of the text analysis technique that is built upon Equation 2? [5 minutes] *b).* Under the text analysis technique mentioned in Question 2.a, what does the dimension k in Equation 2 represent? [5 minutes]

c). Under the text analysis technique mentioned in Question 2.a, what information does the i'th row of $\widehat{\mathbf{T}}$ give us? [5 minutes]

[15 minutes in total]

Question 3:

We run a binary classifier against a test set of 1000 examples and get the following confusion matrix:

	True class	
Predicted class	1	0
1	6	14
0	24	956

where "1" is the positive class, "0" is the negative class.

a). What recall has the classifier achieved on the test set? [5 minutes]

b). What precision has the classifier achieved on the test set? [5 minutes]

c). What accuracy has the classifier achieved on the test set? [5 minutes]

d). The accuracy metric gives a very different picture of classifier performance from precision and recall. Why? Which metric score (taken at face value) is a better indicator of how well the classifier has done? [10 minutes]

[25 minutes in total]

Question 4:

In probabilistic IR, we aim to rank documents by decreasing probability of relevance to the query, $P(R|d,q)$. For general use, we only care about the ranking, not the precise probability of relevance. Therefore, any monotonic transformation of $P(R|d,q)$ will do.

a). One such transformation is the log odds ratio:

$$\log O(R|d,q) \propto \log P(d|R,q) - \log P(d|\bar{R},q) \tag{3}$$

Show the working to get the above expression from:

$$O(R|d,q) = \frac{P(R|d,q)}{P(\bar{R}|d,q)} \tag{4}$$

[10 minutes]

b). To calculate Equation 3, we need a model for $P(d|R,q)$. One such model is the binary independence model (BIM). There are two main assumptions of the BIM; what are they? [5 minutes]

[15 minutes in total]

Question 5:

a). In the language model approach to information retrieval, we try to estimate $P(d|q)$, using the monotonic inversion:

$$P(d|q) \propto P(q|d) \tag{5}$$

Regarding the n-word query q as a list of word occurrences $\{q_1, \ldots, q_i, \ldots, q_n\}$, write an expression for $P(q|d)$ under the unigram language model. [5 minutes]

b). Jelinek-Mercer smoothing defines the quantity:

$$P_{\text{JM}}(w|d) = (1-\lambda)P_{\text{mle}}(w|d) + \lambda P(w|C) \tag{6}$$

How is $P_{\text{mle}}(w|d)$ usually estimated?

[5 minutes]

c). In Equation 6, how is $P(w|C)$ usually estimated? [5 minutes]

d). Smoothing performs two main functions in language models for IR. What are they? [5 minutes]

[20 minutes in total]

Question 6:

a). In Naive Bayes classifiers, we estimate the probability that the n-word document $d = \{w_1, \ldots, w_i, \ldots, w_n\}$ belongs to class c as:

$$P(c|d) \propto P(c) \prod_i P(w_i|c) \tag{7}$$

How is $P(c)$ estimated? [5 minutes]

b). The Laplace-smoothed estimate for $P(w_i|c)$ in Equation 7 can be written:

$$P(w_i|c) = \frac{a+1}{b+|V|} \tag{8}$$

where $|V|$ is the size of the vocabulary, and a and b are standing in for more meaningful variable names. What is a and what is b? [5 minutes]

c). Naive Bayes tends to give too-extreme estimates of the probability of relevance. For instance, if we examined a set of documents each of which a Naive Bayes classifier estimated were 99% likely to be relevant, we might find that only 80% (say) were. Why does Naive Bayes tend to exaggerate proabilities like this? [10 minutes]

[20 minutes in total]

[120 MINUTES FOR EXAM]

end of exam