FINAL EXAM 2018

CLASSIFICATION (25%)

(a) [7%] You have a two-class text classifier γ at your disposal. Discuss how you can make use of γ to extend the two-class classification problem to $J > 2$ classes when:
   (i) The classes are not mutually exclusive, i.e., it is an *any-of* problem where a document can belong to several classes simultaneously, or to a single class, or to none of the classes.
   (ii) The classes are mutually exclusive, i.e., it is a *one-of* problem where a document must belong to exactly one of the classes.

(b) [10%] Classifiers can be organized according to some of their theoretical properties. For each of the dimensions below, describe what discerns or characterizes the two, provide examples of text classifiers, and list some pros and cons.
   (i) Linear versus non-linear classifiers.
   (ii) Parametric versus non-parametric classifiers.
   (iii) Generative versus discriminative classifiers.

(c) [8%] Classify the following statements as either true or false. Briefly justify your answers.
   (i) Training of a linear SVM to solve a two-class classification problem will fail if the two classes are not linearly separable.
   (ii) Having trained a linear SVM, we need to have the full training set available to classify a new data point.
   (iii) If a two-class classification problem is not linearly separable in an n-dimensional space then it is not linearly separable if we map the data points to an m-dimensional space either, with $m > n$.
   (iv) For a linear SVM trained to classify documents as being either relevant or not relevant to a given query, it is possible to use this for ranking.

RELEVANCE EVALUATION (20%)

(a) [8%] For a given query q, the breakeven point is defined as the position k in a ranked list of documents where precision equals recall. Can there exist multiple breakeven points? Either give an example having multiple breakeven points or prove that the breakeven point is unique.

(b) [6%] The F_1 score is a common metric that combines precision and recall into a single measure. How is the F_1 score defined? What can you say about the F_1 score at the breakeven point?

(c) [6%] Explain what mean average precision is, and how it is computed.

QUERY EXECUTION (15%)

(a) [6%] Let $p_i = (j)$ denote a posting indicating document number j. Furthermore, let q denote the conjunctive query *s AND t* where *s* and *t* are dictionary terms associated with the following posting lists:

s: $[p_1 = (2), p_2 = (6), p_3 = (12), p_4 = (19)]$
t: $[p_5 = (1), p_6 = (6), p_7 = (12), p_8 = (21)]$

(i) Assuming that q is evaluated using term-at-a-time scoring, indicate the order in which the postings are processed.
(ii) Assuming that q is evaluated using document-at-a-time scoring, indicate the order in which the postings are processed.

(b) [9%] Discuss some of the relative merits of term-at-a-time scoring and document-at-a-time scoring. Are there situations where only one of the approaches can be used?

PAGERANK (20%)

(a) [4%] Describe the intuition behind the "random surfer model" and how this applies to arriving at a static quality score $g(d)$ for each document.

(b) [10%] Consider the four-document web graph having nodes *V* and edges *E* as follows:

V = {1, 2, 3, 4}
E = {(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 1), (4, 1), (4, 3)}

For this web graph, compute the transition probability matrix *P* for a random surfer assuming a teleportation probability of 0.5. Show how you arrive at the answer.

(c) [6%] Given *P* as computed above, explain how you can use this to compute the PageRank score for each document.

MISCELLANEOUS (20%)

(a) [6%] A vector space representation of text is a useful way to represent both document and queries.
(i) Define, mathematically, what TFIDF scoring is. Describe, intuitively, what TFIDF scoring tries to achieve.
(ii) Define, mathematically, what cosine similarity is. Describe, intuitively, what cosine similarity expresses.

(b) [8%] Approximate string matching techniques are useful when query terms are wrongly or only partially expressed.
(i) Explain how a permuterm index can be used to evaluate the wildcard query *fi*mo*er*.

(ii) Describe a way to efficiently find all strings in a large dictionary that have an edit distance less than or equal to *k* from the malformed query *fiskmongeer*. You can assume that *k* is a small number, e.g., *k = 2*.

(c) [6%] Compression of both the dictionary and the posting lists can have a beneficial impact on search performance.

(i) You have devised a new compression algorithm for integers that results in a smaller data footprint than any other algorithm you know. Yet you decide to not use it as part of your search engine. Provide at least two reasons why you might arrive at such a conclusion.

(ii) Describe the idea behind the Simple-9 compression algorithm.