Information Retrieval and Text Analytics

W. Kraaij and C. Veenman

Final Exam

8 June 2017
B02
14.00-17.00

Student name:

Student number:

This exam consists of 5 pages and 10 questions for a total of 500 points. The grade will be computed as follows: ((points/50)*0.9 + 1)

Instructions:

- Carefully read the instructions and all exercises at the start of the exam.
- Write your name and student number on this hand-out and the answer sheets.
- Verify that your copy of this hand-out is complete and legible.

- Write your answers in a readable form.
- Always provide an explanation for your answer.

- This is a closed-book, closed-notes, individual exam.
- You are not allowed to use your laptop, smartphone or any photographing or telecommunication device. Only a simple calculator is allowed.
- You can work on this exam only within the allocated time-slot.
- Do not unstaple or tear off pages of this hand-out.
- Do not write your answers on this hand-out.
- You must return all pages of this hand-out to the proctor at the end of the exam, regardless of whether or not you have written anything on it.

1. Boolean retrieval and posting lists

a. Consider a collection of N documents about Dutch history. The number of documents containing *Leiden* and *Amsterdam* is *l* and *a* respectively. What is the time complexity of evaluating the query `'Leiden OR NOT Amsterdam'` ? (20p)

b. Rewrite the following query in disjunctive normal form: `(Leiden OR Amsterdam) AND NOT Rembrandt` (15p)

c. For the evaluation of a two word AND query, we need to merge two posting lists {1,5,6,7,13, 15,16,20,21} and {15}. Compute the number of comparisons for the following two cases: (i) standard posting lists; (ii) posting lists with skip pointers using the standard heuristic for skip length. (15p)

2. Stemming is an example of a term normalization function *Fn* generating equivalence classes.

a. *Fn* is a surjective non-injective function, what is the impact of using a normalization function on the size of the dictionary of an IR system? (15p)

b. What is the motivation of term normalization for IR systems? (15p)

c. Which properties should be met by the equivalence classes in order to achieve optimal retrieval effectiveness ? (20p)

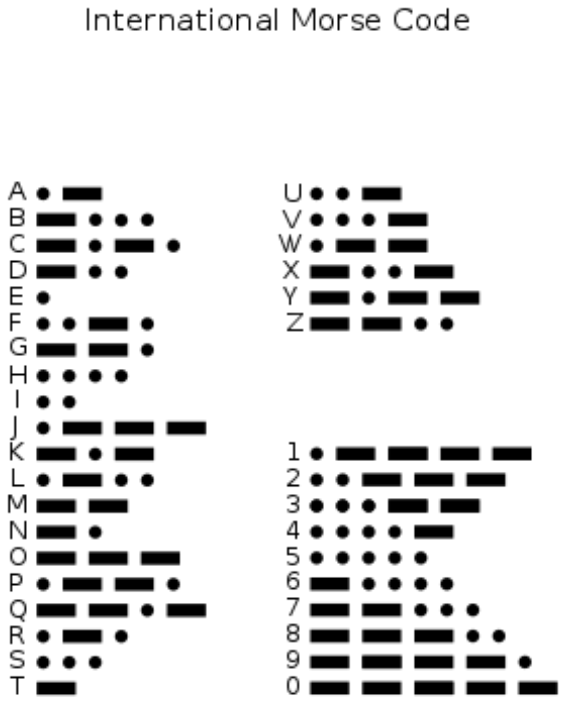
3. Query processing

a. Discuss the pros and cons of using i) hash tables or ii) binary search trees for implementing a dictionary. (15p)

b. Compute the Jaccard coeffcient for the character bigrams of the terms ‘quote’ and ‘qoute’. (10p)

c. Compute the Levenshtein (edit) distance between ‘ford’ and ‘frog’, explain the procedure with a matrix showing the transitions. Use the following cost parameter setting: the 'insert' and 'delete' operation have a cost of 1, the substitute operation has a cost of 2. (25p)

4. Consider the Morse code alphabet in the figure below



a. Explain why the Morse code table has different code lengths for different characters. (10p)

b. Now consider that we replace all but the 5 most frequent letters by the symbol '\$'. The new alphabet consists thus of 6 symbols. Relative symbol frequencies are tabulated in the table below. What is the theoretically minimum average number of bits per symbol for the language modeled in the table? Please motivate. (20p)

E	13%
T	9%
A	8%
O	7%
I	6%
\$	57%

c. Construct a Huffman coding scheme for this reduced alphabet. What is the average number of bits per symbol used for a typical fragment of the language modeled in the above table, using your code Huffman code scheme? (20p)

5. Let *D* be a set of documents and *T* a set of terms.
   a. The tf-idf score of a term *t* within a document *d* is given by:
   $$tf\text{-}idf_{t,d} = (1 + \log(tf_{t,d})) \times \log(\frac{N}{df_t})$$
   Explain the name of this formula and explain the components, structure and intuition behind the formula. (25p)
   b. The cosine-measure for document *d* and query *q* is given by:
   $$sim(q,d) = \sum_{t \in T} q(t) \cdot d(t)$$
   Explain the name of this formula and explain the components, structure and intuition behind the formula. (25p)

6. We assume a test collection consisting of 20 documents, two queries q_1 and q_2 and a set of relevance judgements. The following table shows the relevance judgements for the top 15 results for each query using system *S*. The '*' symbol indicates a document being relevant. There are 8 relevant documents in the result set for query q_1 and 10 for q_2.

q_1	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}
	*	*	*	*	*									*	*

q_2	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}
		*	*	*		*	*				*			*	

   a. Compute the interpolated precision values at 11 recall values {0,0.1,0.2,...,1.0} for q_1 using the definition of the IIR book. (25p)
   b. Compute the mean average precision of system *S* for this test collection consisting of two queries. (25p)

7. Consider the following table of unigram conditional probabilities for a document *D* and a document collection *C*:

model M_D		model M_C	
w	$P(w\|M_D)$	w	$P(w\|M_C)$
the	0.2	the	0.15
a	0.15	a	0.13
of	0.1	of	0.12
to	0.08	to	0.085
Dutch	0.0001	Dutch	0.0002
voters	0.0006	voters	0.0007
give	0.005	give	0.0051
clear	0.003	clear	0.002
signal	0.0003	signal	0.0002
change	0	change	0.00001
...	...	....	...

Consider also two sentences: s_1: "Dutch voters give a clear signal." and s_2: "Dutch voters signal change."

a. Decompose the probability $P(s_1)$= P("Dutch voters give a clear signal.") using the chain rule. Hint: the outcome is a product of N (conditional) probabilities, where N is the sentence length.(15p)
b. Compute the generative probability (query likelihood) of s_1 and s_2 given the unigram model for D: $P(s_2|M_D)$ (see Table above).(20p)
c. Compute the generative probability of s_1 and s_2 where M_D is interpolated with background model M_C using interpolation parameter $\lambda = 0.5$ (15p)

8. Text classification
   a. Explain the name of the naive Bayes classifier. (15p)

Consider the following formula:

$$C_{map} = argmax_{c_i} \frac{P(c_i)P(x_1, \dots, x_n|c_i)}{P(x_1, \dots, x_n)}$$

   b. How can we simplify it into a Naïve Bayes classifier? (20p)
   c. Text collections have large numbers of features. Give two reasons why the large number of features can be a problem. (15p)

9. Consider the following formula:

$$P(p \mid q) = \frac{1}{N} \cdot d + \frac{\delta_{q \to p}}{O(q)} \cdot (1 - d)$$

   a. Explain the components and the structure of this formula, and the intuition behind this formula. (25p)
   b. Why is it essential that the Markov chain describing the random walk process is ergodic? (25p)

10. Entities

   a. Give examples of three different entity types and their types. (10p)
   b. Give three fundamentally different ways to recognize entities in texts. (15p)

Evaluation of named entity recognition systems

   c. What are criteria for correctly identified named entities? (10p)
   d. Give two measures that are typically used for evaluation of entity recognition systems, their formula and meaning. (15p)