IR&TM Probeklausur

Zugelassene Hilfsmittel: Taschenrechner; sonst keine.

Bearbeitungszeit: 60 Minuten

It will be impossible to complete all assignments. Don't worry, this is not required for getting the top grade. The final exam will have a few more exercises, but this does not influence the number of points needed to pass.

You may answer the questions in English or German.

1 Questions

Each question is 2 points.

1.0.1 Question

Define the number of types/tokens in a sentence.

1.0.2 Question

What is lemmatization? Give an example.

1.0.3 Question

Define Levenshtein edit distance.

1.0.4 Question

What is the feast or famine problem?

1.0.5 Question

Write down the formula for cosine similarity between query q and document d.

1.0.6 Question

What are the components of an information retrieval benchmark?

1.0.7 Question

Define the kappa measure.

1.0.8 Question

What does marginal relevance measure?

1.0.9 Question

What is the main independence assumption of Naive Bayes?

1.0.10 Question

Why is feature selection used? Give the two main reasons?

1.0.11 Question

For linearly separable problem, how many different linear decision boundaries are there that separate the two classes of the training set perfectly?

1.0.12 Question

How does an SVM classify a test set point in the margin?

1.0.13 Question

Does K-means find the global optimum? Why (not)?

1.0.14 Question

What are the advantages of a search engine ad compared to other types of ads (radio, television, newspaper)?

1.0.15 Question

What does politeness mean for a crawler?

2 Problems

2.0.16 Problem – 10 points

(i) Compute the Levenshtein matrix and the distance l for the distance between the strings "apfel" (input) and "poems" (output). Assume that all operations have the same cost. (ii) How many different sequences of l Levenshtein operations are there? List them.

2.0.17 Problem – 12 points

Compute the similarity between the query "smart phones" and the document "smart phones and video phones at smart prices" by filling out the empty columns in the table below. Use the similarity function ddd.qqq = lnc.ltn (as given in the table). Assume $N = 10{,}000{,}000$. Treat *and* and *at* as stop words. What is the final similarity score? When computing length

normalized weights you can round the length of a vector to the nearest integer.

	query					document			
word	tf-raw	tf-wght	df	idf	weight	tf-raw	tf-wght	n'lized	product
smart			5,000						
video			50,000						
phones			25,000						
prices			30,000						

Here are some log values you may need:

x	1	2	3	4	5	6	7	8	9
$\log_{10} x$	0	0.3	0.5	0.6	0.7	0.8	0.8	0.9	1.0

2.0.18 Problem – 15 points

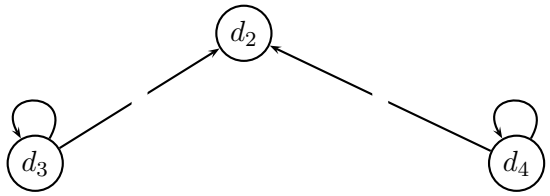

Figure 1: A web graph

Compute PageRank for the web graph in Figure 1 for each of the three pages.

Assume that at each step of the PageRank random walk, we teleport to a random page with probability 0.6, with a uniform distribution over which particular page we teleport to.

Using symmetries to simplify and solving with linear equations might be easier than using iterative methods.

2.0.19 Problem – 10 points

Rank the documents in collection $\{d_1, d_2\}$ for query q using the language model approach to IR introduced in class. Use the mixture coefficient $\lambda = 0.4$. Ignore punctuation marks.

- d_1: In Financial Crisis, No Prosecution of Top Figures
- d_2: Wall Street and the Financial Crisis: Anatomy of a Financial Collapse
- Query q: Financial Crisis