Taaltheorie en Taalverwerking 2017

Homework 6

1. Consider the following query and the following collection of documents: 3 points

   Query: *What is the Dam in Amsterdam, Netherlands?*

   Doc 1: *The Dam is the main town square in the Dutch capital.*

   Doc 2: *Damhotel is a popular and cheap hotel in downtown Amsterdam. It is located between the Dam and Centraal, the main station in Amsterdam (10 min. walk from the Dam).*

   Doc 3: *Jam in the Dam is a unique 3-day music festival bringing US bands and fans to the fabulous Melkweg music hall in Amsterdam, the Netherlands.*

   (a) Build a term-by-document matrix considering as terms all the proper nouns in the collection (*Dam, Dutch, Damhotel, Amsterdam, Centraal, Jam, US, Melkweg, Netherlands*) and as the value of each feature the un-weighted term frequency.

   (b) Calculate the similarity between the query vector and the vector for each of the documents. Recall that the query vector is obtained in the same way as the document vectors.

   – first calculate the dot product similarity without normalising the vectors;
   – then calculate the cosine similarity.

   Do this methods give substantially different results in this case? Justify your answer.

2. Let us assume that (i) the three documents in the previous exercise belong to a collection that contains a total of 15 documents, 5 of which (including Doc 1) are about Dam square in Amsterdam, and (ii) an information retrieval system returned the three documents above (without ranking them) in response to the specified query. 3 points

   (a) Which documents were true positives and which ones were false positives? How many documents were false negatives and how many were true negatives?

   (b) What is the precision, the recall, and the F-score of the results?

   (c) Assuming the term *Dam* appears in 10 of the 15 documents in the collection and the term *Melkweg* appears in only 2 of them, what is the value for each of these terms in Doc3 after applying `tf-idf` weighting?

   Justify your answer in all cases.

3. Using the statistics from the Berkeley Restaurant Project given in Jurafsky & Martin, Chapter 4, compute the probability of the following sentences. Show and justify how you computed your answers. 2 points

   (a) *I want Chinese food* with unsmoothed n-grams

   (b) *I want Chinese Chinese food* with unsmoothed n-grams

   (c) *I want Chinese food* with Laplace-smoothed bigrams

   (d) *I want Chinese Chinese food* with Laplace-smoothed bigrams