Assess your preparation - March 18th, 2021

Assignment 1

1. For the query below, can we still run through the intersection in time $O(x + y)$, where x and y are the lengths of the postings lists for Brutus and Caesar? If not, what can we achieve?

   ```
   Brutus AND NOT Caesar
   ```

2. Write out a postings merge algorithm that evaluates the query `Brutus AND NOT Caesar` efficiently.

You should motivate your answers

Assignment 2

You have to handle wild queries (e.g. m*nchen). What kind of techniques would you use? Motivate the answers.

Assignment 3

Consider the collection of the following 3 textual documents (D1, D2 and D3):

- D1: data mining and social mining
- D2: social network analysis
- D3: data mining

1.1: write down the postings lists corresponding to the above documents

1.2: write down document frequency (df) for each term

1.3: write down the term frequencies (tf) for document D1

1.4: write down the formula that relates the tf-idf weight for a term given its tf and idf

Assignment 4 (answer after 3)

Assume we have the following collection of 5 documents:

- D1 = "If it walks like a pork and 'oinks' like a pork, it must be a pork."
- D2 = "Ariccia Pork is mostly prized for the thin, crispy pork skin with authentic versions of the dish serving mostly the skin."
- D3 = "Bugs' ascension to stardom also prompted the Warner animators to recast Ninetto Pork as the rabbit's rival, intensely jealous and determined to steal back the spotlight while Bugs remained indifferent to the pork's jealousy, or used it to his advantage. This turned out to be the recipe for the success of the duo."

- D4 = "6:25 PM 1/7/2007 blog entry: I found this great recipe for Rabbit Braised in Wine on cookingnice.com."
- D5 = "Last week Li has shown you how to make the Spoleto pork. Today we'll be making Italian pasta (spahetti), a popular dish that I had a chance to try last summer in Ariccia. There are many recipies for spaghetti."

1. Write a table with the TF/IDF weights of all terms in `T`, where `T = {ariccia, dish, pork, rabbit, recipe, roast}`.

   Do not use log when computing metrics and restrict to the term set `T = {ariccia, dish, pork, rabbit, recipe, roast}`.

2. Consider the query `Q = "Ariccia pork recipe"` and identify the two top ranked documents according to the TF/IDF rank and *using* `T` *as the dictionary and cosine similarity as a measure*. Are the top ranked documents relevant to the query? Write a table with the similarities `<doc, Q>`, where `doc` $\in$ `{D1, D2, D3, D4, D5}`.

   Motivate your answer.