Intelligent Data Analysis 2020 - Exercise sheet 1

Martin Russell 16th January 2020

Question 1

According to Zip's Law, with C = 0.1 and $\alpha = 1$, how many times would the most frequent word occur in a document that contains 180,000 words?

Question 2

Two documents D_1 and D_2 have the following forms:

- D_1 : Delays on Southern Rail trains peaked over the Christmas period
- D₂: Industrial action caused train cancellations and train delays in the south of England over Christmas

After stop-word removal and stemming these become:

- d_1 : delay south rail train peak christmas period
- d_2 : industry action cause train cancel train delay south england christmas

The IDFs of the words that occur in these documents are given in the Table below:

Term (t)	IDF(t)	Term (t)	IDF(t)	Term (t)	IDF(t)
action	0.4	delay	0.8	period	0.5
cancel	0.6	england	2.1	rail	2.2
cause	0.3	Industry	1.6	south	0.8
christmas	1.5	peak	0.6	train	1.9

- 1. Calculate the TF-IDF similarity $sim(d_1, d_2)$ between d_1 and d_2 .
- 2. Assuming that the vocabulary in the table above is the complete vocabulary and is ordered according to the table, write down the document vectors $vec(d_1)$ and $vec(d_2)$.
- 3. Suppose that the term "delay" is repeated N times in d_1 . Write down a formula for the angle θ_N between $vec(d_1)$ and $vec(d_2)$ as a function of N.
- 4. What is the limiting value of θ_N as $N \to \infty$? (Note: it is possible to answer this question without the formula from the previous part.

Question 3

Suppose that d_1 and d_2 are documents. Show that

$$0 \le Sim(d_1, d_2) \le 1 \tag{1}$$

where $Sim(d_1, d_2)$ is the TF-IDF similarity between documents d_1 and d_2 .

Question 4

Consider the following set of documents:

- The cat sat on the mat
- The dog chased the cat
- The cat sat on the dog
- The dog chased another dog
- The cat sat on the dog's mat

After text pre-processing these become:

- d_1 : cat sat mat
- d_2 : dog chase cat
- d₃: cat sat dog
- d₄: dog chase dog
- d_5 : cat sat dog mat

With the vocabulary ordered alphabetically as follows {cat, chase, dog, mat, sat}

- Calculate the Inverse Document Frequency of each word in the vocabulary
- Calculate the document vectors $vec(d_1), \cdots vec(d_5)$
- Calculate the TF-IDF similarity CSim(d₂, d₄)

Question 5

Let d_1 and d_2 be documents. Show that the cosine similarity between $vec(d_1)$ and $vec(d_2)$ is the same as the TF-IDF similarity between d_1 and d_2 . In other words show that

$$CSim(d_1, d_2) = Sim(d_1, d_2)$$
(2)