Intelligent Data Analysis

Solutions to Exercise sheet 1

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

   Solution:

   According to Zipf's Law the rank-frequency distribution F is given by:
   $F(r) = \frac{C}{r^{\alpha}}$
   Where $\alpha \approx 1$ and $C \approx 0.1$.
   For the most frequent word $r = 1$ and $F(1) = 0.1$. Therefore the number of occurrences of the most frequent word is predicted to be $180{,}000 \times F(1) = 180{,}000 \times 0.1 = 18{,}000$.

2. Two documents D_1 and D_2 have the following forms:

   D_1: *Delays on Southern Rail trains peaked over the Christmas period*

   D_2: *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

   a. Calculate the TF-IDF similarity $sim(d_1, d_2)$ between d_1 and d_2.

   Solution:

   I think it is easiest to do this type of calculation in a table. In the table below, the 5th, 8th and 9th columns contain the numbers that you need to calculate the similarity: the square root of the sum of column 5 (8) is the length of document 1 (2) and the sum of column 9 is the numerator in the similarity calculation.

t	$IDF(t)$	f_{t,d_1}	w_{t,d_1}	${w_{t,d_1}}^2$	f_{t,d_2}	w_{t,d_2}	${w_{t,d_2}}^2$	$w_{t,d_1} \times w_{t,d_2}$
action	0.4	0	0	0	1	0.4	0.16	0
cancel	0.6	0	0	0	1	0.6	0.36	0

cause	0.3	0	0	0	1	0.3	0.09	0
christmas	1.5	1	1.5	2.25	1	1.5	2.25	2.25
delay	0.8	1	0.8	0.64	1	0.8	0.64	0.64
England	2.1	0	0	0	1	2.1	4.41	0
industry	1.6	0	0	0	1	1.6	2.56	0
peak	0.6	1	0.6	0.36	0	0	0	0
period	0.5	1	0.5	0.25	0	0	0	0
Rail	2.2	1	2.2	4.84	0	0	0	0
south	0.8	1	0.8	0.64	1	0.8	0.64	0.64
train	1.9	1	1.9	3.61	2	3.8	14.44	7.22

$$\|d_1\| = \sqrt{2.25 + 0.64 + 0.36 + 0.25 + 4.84 + 0.64 + 3.61} = 3.55,$$

$$\|d_2\| = 5.05$$

The numerator in the formula for TF-IDF similarity is just the sum of entries in the final column. Therefore,

$$sim(d_1, d_2) = \frac{10.75}{3.55 \times 5.05} = 0.6$$

b. 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 $vec(d_2)$.

Solution: The elements of a document vector are just the TF-IDF weights for each term for that document. Hence the document vectors for documents d_1 and d_2 can be obtained directly from columns 4 and 7 of the table:

$$vec(d_1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1.5 \\ 0.8 \\ 0 \\ 0 \\ 0.6 \\ 0.5 \\ 2.2 \\ 0.8 \\ 1.9 \end{bmatrix}, vec(d_2) = \begin{bmatrix} 0.4 \\ 0.6 \\ 0.3 \\ 1.5 \\ 0.8 \\ 2.1 \\ 1.6 \\ 0 \\ 0 \\ 0 \\ 0.8 \\ 3.8 \end{bmatrix}$$

c. 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.

Solution: If "delay" is repeated N times in d_1 this will not have any effect on $IDF(delay)$ but it will change w_{delay,d_1}. In this case:

$$w_{delay,d_1} = N \times 0.8, \|d_1\| = \sqrt{11.95 + (0.8N)^2}$$

$$\|d_2\| = 5.05 \text{ (unchanged)}$$

$$\sum_{t \in d_1 \cap d_2} w_{t,d_1} \times w_{t,d_2} = 10.11 + 0.64N$$

Therefore, since cosine similarity and TF-IDF similarity are the same,

$$cos(\theta_N) = \frac{10.11 + 0.64N}{5.05 \times \sqrt{11.95 + 0.64N^2}}$$

Therefore

$$\theta_N = cos^{-1}\left(\frac{10.11 + 0.64N}{5.50 \times \sqrt{11.95 + 0.64N^2}}\right)$$

d. What is the limiting value of θ_N as $N \rightarrow \infty$? (Note: it is possible to answer this question without the formula from part (iv)).

Solution: There are at least two ways to do this. Denote the angle by θ. Then

From the formula from the previous part:

$$cos(\theta) = lim_{N\rightarrow\infty} \frac{10.11 + 0.64N}{5.05 \times \sqrt{11.95 + 0.64N^2}} =$$

$$lim_{N\rightarrow\infty} \frac{0.64N}{5.05 \times \sqrt{0.64N^2}} = lim_{N\rightarrow\infty} \frac{0.8}{5.05} = 0.158$$

So $\theta = cos^{-1}(0.158) = 1.412 \text{ radians} = 80.9^o$.

Alternatively, notice that as $N \rightarrow \infty$ the term "delay" will dominate and the direction of the vector representation of d_1 will tend towards

$$e = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \text{ Hence } cos(\theta) = \frac{e \cdot vec(d_2)}{\|d_2\|} = \frac{0.8}{5.05} = 0.158.$$

3. Suppose that d_1 and d_2 are documents. Show that

$$0 \le sim(d_1, d_2) \le 1$$

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

Solution:

This follows immediately from the result from week 2 that $sim(d_1, d_2)$ is equal to the cosine of the angle between $vec(d_1)$ and $vec(d_2)$.

Alternatively, you can obtain the result from first principles by considering the cases where $d_1 = d_2$ and where d_1 and d_2 have no words in common

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
   - d_3: cat sat dog
   - d_4: dog chase dog
   - d_5: cat sat dog mat

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

   a. Calculate the Inverse Document Frequency of each word in the vocabulary

   Solution:

Word w	d_1	d_2	d_3	d_4	d_5	ND_w	IDF(w)
cat	1	1	1	0	1	4	0.22
chase	0	1	0	1	0	2	0.92
dog	0	1	1	2	1	4	0.22
mat	1	0	0	0	1	2	0.92
sat	1	0	1	0	1	3	0.51

   b. Calculate the document vectors vec(d_1), …, vec(d_5)

   Solution:

   First calculate the TF-IDF weights using the values in the table above:

Word w	d_1	d_2	d_3	d_4	d_5
cat	0.22	0.22	0.22	0	0.22
chase	0	0.92	0	0.92	0
dog	0	0.22	0.22	0.45	0.22

mat	0.92	0	0	0	0.92
sat	0.51	0	0.51	0	0.51

$$vec(d_1) = \begin{bmatrix} 0.22 \\ 0 \\ 0 \\ 0.92 \\ 0.51 \end{bmatrix}, vec(d_2) = \begin{bmatrix} 0.22 \\ 0.92 \\ 0.22 \\ 0 \\ 0 \end{bmatrix}, vec(d_3) = \begin{bmatrix} 0.22 \\ 0 \\ 0.22 \\ 0 \\ 0.51 \end{bmatrix},$$

$$vec(d_4) = \begin{bmatrix} 0 \\ 0.92 \\ 0.45 \\ 0 \\ 0 \end{bmatrix}, vec(d_5) = \begin{bmatrix} 0.22 \\ 0 \\ 0.22 \\ 0.92 \\ 0.51 \end{bmatrix},$$

c. Calculate the similarity sim(d_2,d_4)

Solution:

$$\|d_2\| = 0.97, \|d_4\| = 1.02$$

$$sim(d_2, d_4) = \frac{\sum_{w \in d_2 \cap d_4} w_{w,d_2} w_{w,d_4}}{\|d_2\| \|d_4\|} = \frac{0.94}{1.07 \times 1.02} = 0.95$$

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

$$CSim(d_1, d_2) = sim(d_1, d_2)$$

Solution:

This was covered in the lecture in week 2 and is in the notes on Canvas.