Handin 7 - 10 ECTS

We have used Generative Artificial Intelligence tools in doing this assignment, for the following legitimate use cases only: to get background information or understand the topic / problem, to improve writing of own text, to find gaps in our knowledge. The solution of the assignment is entirely our own.

Exercise (a)**.**

You are given a collection of three short documents:

1. *The sky is blue.*
2. *The sun is bright.*
3. *The sun in the sky is bright.*

Please calculate the TF-IDF score for the word "sky" for each of the documents. Make sure to include the intermediate steps of your calculation in your solution.

The TF-IDF score is a combinantion of the term-frequency and the inverse document frequency. It is calculated in the following way,

TF-IDF

Given a term i and document j, the term frequency TF_{ij} is given by,

$$TF_{ij} = \frac{f_{ij}}{\sum_i f_{ij}}.$$

Where f_{ij} is the frequency of the term i in the document j, which is normalized by the sum of all the frequencies of any other terms. The inverse document frequency is given by,

$$IDF_i = \log \frac{N+1}{n_i+1}.$$

Where N is the total number of documents, and n_i is the number of documents that mention the term i. The TD-IDF score is found by multiplying the two,

$$w_{ij} = TF_{ij} \cdot IDF_i.$$

We start by calculating the IDF score for the term *sky*, as this only has to be calculated once for all the documents,

$$IDF_{sky} = \log \frac{N+1}{n_{sky}+1} = \log \frac{3}{2}.$$

To calculate the TF-scores, we essentially have to count the number of terms in each of the documents and use that as our normalization factor. But before doing so, we should filter out non-significant. I have chosen the following words as insignificant,

$$\text{insignificant\_words} = \{\text{is, The, in, the}\}.$$

The $TF_{sky,j}$ scores for the three are,

$$TF_{sky,1} = \frac{1}{2} \quad , \quad TF_{sky,2} = \frac{0}{2} = 0 \quad , \quad TF_{sky,3} = \frac{1}{3}.$$

Combining TF and IDF we get,

$$w_{sky,1} = \frac{1}{2}\log\frac{3}{2} \quad , \quad w_{sky,2} = 0 \quad , \quad w_{sky,3} = \frac{1}{3}\log\frac{3}{2} \quad , \quad .$$

So document 1 is the most relevant, document 2 is completely irrelevant and document 3 is somewhat relevant.

Exercise (b)**.**

Assume we have 2 documents:

1. The vector for the first document is $(1,1,1,1,1,1,0,0)$
2. The vector for the second document is $(0,1,0,1,1,1,1,1)$ Please calculate the document similarity using cosine similarity. Make sure to include the intermediate steps of your calculation in your solution.

The cosine similarity is calculated in the following way,

Cosine similarity

The *cosine similarity* Q_{ab} of two vectors a and b is given by,

$$Q_{ab} = \frac{a \cdot b}{||a|| ||b||}.$$

Lets start by calculating the dot product where $a = (1,1,1,1,1,1,0,0)^T$ and $b = (0,1,0,1,1,1,1,1)^T$

$$a \cdot b = 4.$$

And the norm of each,

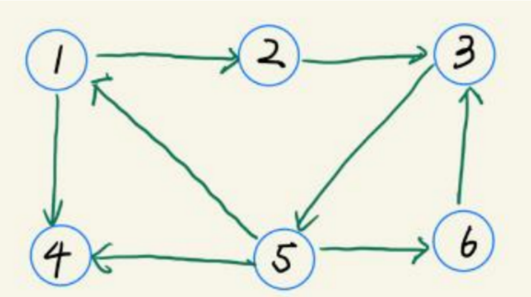
$$||a|| = ||b|| = \sqrt{6}.$$

So Q_{ab} is found to be,

$$Q_{ab} = \frac{4}{\sqrt{6}^2} = \frac{2}{3}.$$

Exercise (c)**.**

Consider the following graph representation of six web pages 1,...,6 and the hyperlinks between them. Initially, all these web pages have the same PageRank score which is 1/6. Please calculate the PageRank of all the nodes for the next iteration. Make sure to include the intermediate steps of your calculation in your solution



We use the pagerank formula,

PageRank

Given a directional graph, the rank P of the node A is given by,

$$P(A) = \sum_i \frac{P(X_i)}{O(X_i)}.$$

Where X_i are the nodes pointing to A, and $O(X_i)$ is the number of vertices leaving X_i

During a single iteration, we calculate the scores of all the nodes using the initial state. The page rank of each node is then updated at the end of the iteration (this is at least what was done in the example in the slides). For a bigger graph, it would make sense to update the pageranks at some intermediate steps. Since all our nodes are in the initial state we have,

$$P(X_i) = \frac{1}{6} \qquad \text{for all } i.$$

The pagerank of each node can then be calculated as,

$$\begin{aligned}
P(X_1) &= \frac{P(X_5)}{O(X_5)} = \frac{1}{6} \cdot \frac{1}{3} = 0.55 \\
P(X_2) &= \frac{P(X_1)}{O(X_1)} = \frac{1}{6} \cdot \frac{1}{2} = 0.083 \\
P(X_3) &= \frac{P(X_2)}{O(X_2)} + \frac{P(X_6)}{O(X_6)} = \frac{1}{6} + \frac{1}{6} = 0.33 \\
P(X_4) &= \frac{P(X_4)}{O(X_4)} + \frac{P(X_5)}{O(X_5)} = \frac{1}{6} \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{1}{3} = 0.13 \\
P(X_5) &= \frac{P(X_3)}{O(X_3)} = \frac{1}{6} = 0.166 \\
P(X_6) &= \frac{P(X_5)}{O(X_5)} = \frac{1}{6} \cdot \frac{1}{3} = 0.055.
\end{aligned}$$

Which gives us the following ranking from highest to lowest: $(X_3, X_5, X_4, X_2, X_1, X_6)$

Exercise (d)**.**

Consider a scenario where we are asked to develop a diagnostic tool for a serious disease by finding relevant medical literature. Failing to identify information about a sick patient could be life-threatening, but providing nonrelevant information would only be a minor irritation.

| **.** In this case, which is more important: precision or recall? Please argue for your choice.

In this case, recall is the most important. A high recall with low precision, is likely to return some cases that aren't relevant, but should find most relevant cases. On the other having high precision with low recall, will return a sample dense in relevant cases, but is likely to forget some relevant cases. As missing relevant cases is deadly, the first error is preferable in this case.

. Given the following summary of the retrieval results, please calculate the recall, precision, and F1 score. Make sure to include the intermediate steps of your calculation in your solution.

- 95 relevant medical documents were retrieved.
- 30 irrelevant documents were retrieved.
- 5 relevant medical documents were missed.
- 22 irrelevant documents were not retrieved.

The different scores are defined in the following way,

Recall, Precision and F1 scores

Let R be the recall score and P the prescision score. Let D be the number of documents, D_p the relevant documents, D_f the irrelevant documents and D_r and D_m the retrieved and missed/not retrieved documents. The scores are then found by,

$$
\begin{aligned}
R &= \frac{D_{rp}}{D_p} \\
P &= \frac{D_{rp}}{D_r} \\
F_1 &= \frac{2RP}{P+R}.
\end{aligned}
$$

In our case we have,

$$
\begin{aligned}
D_{rp} &= 95 \\
D_p &= D_{rp} + D_{mp} = 5 + 95 = 100 \\
D_r &= D_{rp} + D_{rf} = 95 + 30 = 125.
\end{aligned}
$$

The R and P scores are then,

$$
\begin{aligned}
R &= \frac{95}{100} \\
P &= \frac{95}{125}.
\end{aligned}
$$

And finally the F_1 score is,

$$
F_1 = \frac{2 \cdot \left(\frac{95}{100}\right) \cdot \left(\frac{95}{125}\right)}{\left(\frac{95}{100} + \frac{95}{125}\right)} \approx 0.84.
$$