IRTM Gxcercise sheet 2

Task 1

$$J(q,d_i) = \frac{|q \cap d_i|}{|q \cup d_i|}$$

$J(q,d_1) = \frac{2}{8} = \frac{1}{4} = 0{,}25$

$J(q,d_2) = \frac{1}{11}$

$J(q,d_3) = \frac{2}{2} = 1$

The document d_3 is most relevant to the query q as its Jaccard index is the highest of all of the documents. This makes sense since the document d_3 includes all of the queries ~~of~~ words ~~a~~ and has no additional which means the query q and the document d_3 match perfectly.

Task 2

tf-score(q,d_1) = 2

tf-score(q,d_2) = 1

tf-score(q,d_3) = 2

tf-score$(q,d_i) = \sum_{t \in q \cap d_i} (1 + \log tf_{t,d_i})$

tf-idf-score$(q,d_1) = \log\frac{3}{2} + \log\frac{3}{3} \approx 0{,}176$

tf-idf-score$(q,d_2) = 1 \cdot \log\frac{3}{3} = 0$

tf-idf-score$(q,d_3) = \log\frac{3}{2} + \log\frac{3}{3} \approx 0{,}176$

tf-table	d_1	d_2	d_3
algorithm ~~the~~	1	0	1
intersection	1	1	1

Document d_1 and d_3 are weighted with the same tf-idf-score meaning they have the same relevance (which is higher than document d_2). ~~Since cosine similarity uses the same tf-idf-score to create the vectors and normalizes them, the tf-idf score is the only relevant~~ The length ~~does~~ of the document doesn't come in to play since the vectors get normalized.