Information Retrieval: Assignment 1

Marcel Aguilar Garcia, Id: 20235620

October 13, 2021

Question 1

For this question, we have a collection $\mathcal{C}$ of three documents, D_1, D_2 and D_3, and we would like to find the similarity between the query Q = 'gold silver truck' and these documents. The scores given by the similarity can be used to rank these documents with respect to the query.

In order to calculate the similarity between the documents and the query, I will start representing the documents in a vector space model by using the tf-idf weighting scheme:

$$w_{i,j} = tf_{i,j} idf_j = f_{i,j} log(N/n_i)$$

Let V be the vocabulary of the collection $\mathcal{C}$. The VSM given by $\mathcal{C}$ is a $|V|$-dimensional vector space and each of the documents can represented as:

$$V_{D_j} = (w_{0,j}, w_{1,j}, ..., w_{|V|,j})$$

where $w_{i,j}$ is the weight of the term $t_i \in V$ with respect to D_j. Note that if t_i does not appear in D_j, then $tf_{i,j} = 0$ and therefore $w_{i,j} = 0$. On the other hand, if t_i appears in all documents, then idf_j = log(N/N) = 0 and therefore $w_{i,j} = 0$.

Let's start calculating $idf_i = log(3/n_i)$ for $i = 1, ..., |V|$ where n_i is the number of documents in which the term $t_i \in V$ appears.

term	word	idf
t_1	Shipment	log(3/2)
t_2	of	0
t_3	gold	log(3/2)
t_4	damaged	log(3)
t_5	in	0
t_6	a	0
t_7	fire	log(3)
t_8	Delivery	log(3)
t_9	silver	log(3)
t_{10}	arrived	log(3/2)
t_{11}	truck	log(3/2)

Inverse Document Frequency penalises words that are too common by giving them a lower value. In our example, common words such as prepositions ('of') have been given an idf value of zero as they appear in all documents.

On the other hand, the term frequency $tf_{i,j}$ is used to quantify the frequency of a term within a document. An easy way to calculate $tf_{i,j}$ is by counting the times that a term t_i occurs in the document D_j. However, a term that appears 4 times in a 50 words document would have the same term frequency than a term that appears 4 times in a 1000 words document. This can be fixed by normalising this value by the length of each document. This is $tf_{i,j} = \frac{n_{i,j}}{|D_j|}$ where $n_{i,j}$ is the number of times that t_i appears in D_j. In the following example, I have done this calculation for document D1, D2,

and D3. Note that to simplify, I am not showing the terms from $|V|$ that do not appear in a document (with term frequency equal to zero):

Shipment	of	gold	damaged	in	a	fire
1/7	1/7	1/7	1/7	1/7	1/7	1/7

Delivery	of	silver	arrived	in	a	truck
1/8	1/8	2/8	1/8	1/8	1/8	1/8

Shipment	of	gold	arrived	in	a	truck
1/7	1/7	1/7	1/7	1/7	1/7	1/7

Now, tf-idf can be calculated by multiplying tf and idf:

Shipment	gold	damaged	fire
log(3/2)/7	log(3/2)/7	log(3)/7	log(3)/7

Delivery	silver	arrived	truck
log(3)/8	2log(3)/8	log(3)/8	log(3/2)/8

Shipment	gold	arrived	truck
log(3/2)/7	log(3/2)/7	log(3/2)/7	log(3/2)/7

The same transformation can be done to the query Q = 'gold silver truck':

gold	silver	truck
log(3/2)/3	log(3)/3	log(3/2)/3

Finally, we can calculate the similarities between the documents and the query $\text{sim}(Q, D_j)$ for j=1,2,3. This similarity score should represent how *close* the query is to each of the documents. As we are in a VSM seems natural to use the cosine of the angle between these vectors. Note that following the same order of wording from the idf table (shipment,of,gold,damaged,in,a,fire,delivery,silver,arrived,truck) the whole representation of the vectors is:

$$D_1 = (\frac{\log 3/2}{7}, 0, \frac{\log 3/2}{7}, \frac{\log 3}{7}, 0, 0, \frac{\log 3}{7}, 0, 0, 0, 0)$$

$$D_2 = (0, 0, 0, 0, 0, 0, 0, \frac{\log 3}{8}, \frac{\log 3}{8}, 2\frac{\log 3}{8}, \frac{\log 3/2}{8})$$

$$D_3 = (\frac{\log 3/2}{7}, 0, \frac{\log 3/2}{7}, 0, 0, 0, 0, 0, 0, \frac{\log 3/2}{7}, \frac{\log 3/2}{7})$$

$$Q = (0, 0, \frac{\log 3/2}{3}, 0, 0, 0, 0, 0, \frac{\log 3}{3}, 0, \frac{\log 3/2}{3})$$

the cosine of the angle can be calculated as

$$\text{sim}(Q, D_j) = \frac{\sum_{i=1}^{11} w_{i,j} w_{i,Q}}{\sqrt{\sum_{i=1}^{11} w_{i,j}^2} \sqrt{\sum_{i=1}^{11} w_{i,Q}^2}}$$

Let's first calculate the euclidean norm of each vector:

$$\|D_1\| = \sqrt{(\frac{\log 3/2}{7})^2 + (\frac{\log 3/2}{7})^2 + (\frac{\log 3}{7})^2 + (\frac{\log 3}{7})^2} \approx 0.102$$

$$\|D_2\| = \sqrt{(\frac{\log 3}{8})^2 + (\frac{2\log 3}{8})^2 + (\frac{\log 3}{8})^2 + (\frac{\log 3/2}{8})^2} \approx 0.148$$

$$\|D_3\| = \sqrt{(\frac{\log 3/2}{7})^2 + (\frac{\log 3/2}{7})^2 + (\frac{\log 3/2}{7})^2 + (\frac{\log 3/2}{7})^2} \approx 0.05$$

$$\|Q\| = \sqrt{(\frac{\log 3/2}{3})^2 + (\frac{\log 3}{3})^2 + (\frac{\log 3/2}{3})^2} \approx 0.179$$

Using the previous values, we can calculate the similarities as:

$$\text{sim}(Q, D_1) = \frac{\log(3/2)^2/21}{0.102 \cdot 0.179} \approx 0.08$$

$$\text{sim}(Q, D_2) = \frac{2\log(3)^2/24 + \log(3/2)^2/24}{0.147 \cdot 0.179} \approx 0.77$$

$$\text{sim}(Q, D_3) = \frac{\log(3/2)^2/21 + \log(3/2)^2/21}{0.05 \cdot 0.179} \approx 0.329$$

As $\text{sim}(Q, D_2) > \text{sim}(Q, D_3) > \text{sim}(Q, D_1)$ we can say that the rank of these documents with respect to the query is D_2, D_3, D_1 respectively.

Question 2

For this question, we need to analyse the change of $\text{sim}(Q, D_1)$ for some augmentations of D_1. In general, we would expect the similarity to reduce if words that are not in Q are introduced, otherwise, if the words added are in Q, we would expect the similarity to increase.

In this case, the words added in each of the augmentations were already in D_1. This means that the idf values calculated in Question 1 should remain the same. On the other hand, the tf will increase for the term added and, due to the normalisation, will be reduced for all other terms.

When considering the angle between two vectors in a VSM, we would expect the vectors to get more distant for each added term that is not in Q, or closer otherwise.

Finally, as D_1 and Q have only one word in common, *gold*, the similarity can be simplified to:

$$\text{sim}(Q, D_1) = \frac{w_{gold,D_1} w_{gold,Q}}{\|D_1\|\|Q\|}$$

From now on I will denote by D_{1j}^A the augmentation of D_1 for the cases below j=1,2,3,4.

1)

D_{11}^A **= Shipment of gold damaged in a fire. Fire.**

Following the reasoning above, we would expect $\text{sim}(Q, D_{11}^A) < \text{sim}(Q, D_1)$. This is equivalent to:

$$\frac{w_{gold,D_{11}^A} w_{gold,Q}}{\|D_{11}^A\|\|Q\|} < \frac{w_{gold,D_1} w_{gold,Q}}{\|D_1\|\|Q\|}$$

Note that, as we are normalising the term frequency, $w_{gold,D_{11}^A} < w_{gold,D_1}$. Therefore, proving that

$$\|D_1\| \approx 0.08 < \|D_1^A\|$$

is enough to see that $\text{sim}(Q, D_{11}^A) < \text{sim}(Q, D_1)$.

$$\|D_{11}^A\| = \sqrt{((\log(3/2)/8)^2 + (\log(3/2)/8)^2 + (\log(3)/8)^2 + (2\log(3)/8)^2)} \approx 0.136$$

As expected, this augmentation decreases the similarity.

2)

D_{12}^A = **Shipment of gold damaged in a fire. Fire. Fire.**

As we are adding an additional term that is not in Q, we would expect the similarity to be lower than in the first case, this is equivalent to $\text{sim}(Q, D_{12}^A) < \text{sim}(Q, D_{11}^A) < \text{sim}(Q, D_1)$. Applying the same reasoning as in case (1) is enough to prove that

$$\|D_{11}^A\| \approx 0.136 < \|D_{12}^A\|$$

where

$$\|D_{12}^A\| = \sqrt{((\log(3/2)/9)^2 + (\log(3/2)/9)^2 + (\log(3)/9)^2 + (3\log(3)/9)^2)} \approx 0.170$$

As expected, this augmentation decreases similarity more than D_{11}^A

3)

D_{13}^A = **Shipment of gold damaged in a fire. Gold.**

As we are adding a term that appears in Q, we would expect the similarity to be greater. This is the same as $\text{sim}(Q, D_1) < \text{sim}(Q, D_{13}^A)$:

$$\frac{w_{gold,D_1} w_{gold,Q}}{\|D_1\|\|Q\|} < \frac{w_{gold,D_{13}^A} w_{gold,Q}}{\|D_{13}^A\|\|Q\|}$$

However, as we have increased the term frequency of *gold*, we would expect $w_{gold,D_1} < w_{gold,D_{13}^A}$. This means that proving

$$\|D_{13}^A\| < \|D_1\| \approx 0.102$$

is enough to prove that this augmentation increases the similarity.

$$\|D_{14}^A\| = \sqrt{((\log(3/2)/8)^2 + (2\log(3/2)/8)^2 + (\log(3)/8)^2 + (\log(3)/8)^2)} \approx 0.0972$$

Again, this result is aligned with our expectation.

4)

D_{14}^A = **Shipment of gold damaged in a fire. Gold. Gold.**

Following the same reasoning as before, we would expect $\text{sim}(Q, D_1) < \text{sim}(Q, D_{13}^A) < \text{sim}(Q, D_{14}^A)$. In order to prove this inequality, is enough to see that

$$\|D_{14}^A\| < \|D_{13}^A\| \approx 0.0977$$

$$\|D_{13}^A\| = \sqrt{((\log(3/2)/9)^2 + (3\log(3/2)/9)^2 + (\log(3)/9)^2 + (\log(3)/9)^2)} \approx 0.0972$$

which proves again that the similarity has increased.

Question 3

In the previous questions we have been using the weighting scheme tf-idf which is based on the frequency of the words in each document and through all documents from the collection. Within the context of all the scientific articles published in the Communications of the ACM, we may want to consider other features such as the citations and the date of the publication of the documents (scientific articles) returned .

Citations

A feature that could be taken into account to quantify the importance of an article would be the number of citations. The new weighting scheme could be adapted over tf-idf. In order to understand better how the number of citations of a document could be used, let's consider the simplest case in which a query returns two scientific articles, S_1 and S_2, with the same similarity score based on tf-idf, this is $\text{sim}(Q, S_1) = \text{sim}(Q, S_2)$. However, if we know that S_1 has more citations than S_2, it seems common sense to rank S_1 before S_2. The new similarity could be define as, e.g.,

$$\text{sim'}(Q, S) := W_1 \cdot \text{sim}(Q, D) + W_2 \cdot \text{citations}(S)$$

where citations(D) = number of citations from scientific article S and, W_1 and W_2, are the weights we want to give to them. In general, I would expect W_1 to be close to 1 as we would still want to have a document similarity based on the frequency of words and used the citation score to be make a difference in cases where scores are very similar.

Date Of Publication

In a similar way, we may want to give less weight to articles that were published long time ago. Using the same reasoning as before, let's consider a query that returns two scientific articles S_1 and S_2 with the same similarity score by using tf-idf weighting scheme, this is $\text{sim}(Q, S_1) = \text{sim}(Q, S_2)$. If we know that S_1 is newer than S_2, we may want to rank S_1 before S_2. A similar definition than before could be adapted to consider date of publication, e.g.,

$$\text{sim'}(Q, D) := W_1 \cdot \text{sim}(Q, S) + W_2 \cdot \text{years\_since\_publication}(S)$$

Note that this definitions are not suitable for the interpretation of an angle in a vector space model and that there is probably a better way of defining them.