Information Retrieval: Assignment 1

Adrian Langseth

September 2019

Task 1: Basic Definitions

Information Retrieval vs Data Retrieval

Data retrieval is based on searching for a document which includes a certain set of keywords. Data retrieval needs an exact match to return a result. This demands very strict semantics in which a single mistake might make the system fail.
Information retrieval has a different objective in its search for documents on a certain subject. Information retrieval uses data retrieval to accomplish this, but uses many other disciplines as well. Therefore we can say data retrieval is a subset of information retrieval. IR uses a partial match or best match type matching, and therefore the semantics and tolerance for errors are looser.

Structured Data vs Unstructured Data

Structured data is data which has a clearly defined data type, making it very efficient for retrieval. Unstructured data is the other type which is not as easily searchable. This is could be videos, audio clips, social media posts, pictures and so on.

Task 2: Term Weighting

Term Frequency

The frequency of which a term appears in a document.

Document Frequency

The frequency of which a term appears globally in documents. So it measures the global frequency of a term.

Inverse Document Frequency

The inverse of the frequency of which a term appears globally in documents. So it measures the global rarity of a term.

Why idf is important for term weighting?

Rare terms in the search query should have a bigger impact if it occurs in the document than a very common phrase such as ”and”. e.g. if a search is done on ”A Rhododendron” we would want documents which contain only the word ”Rhododendron” to appear before a document which contains the word ”a” but not ”rhododendron”.

Boolean Model and Vector Space Model

q1: doc1, doc3, doc4, doc6, doc8
q2: doc1, doc4, doc6, doc9
q3: doc1, doc2, doc3, doc4, doc5, doc6, doc7, doc8, doc10
q4: doc2, doc3, doc8
q5: doc1, doc3, doc4, doc6, doc8, doc9

2) The dimension of the vector space model in this example is 4, as this is the amount of terms (Big, small, cat, dog)

3)

	”Cat”	”Dog”	”Big”	”Small”
doc1	1	1	1	1
doc2	0	1	0	0
doc3	1	1	0	0
doc4	2	1	2	1
doc5	0	0	1	1
doc6	1	1	1	1
doc7	0	0	3	0
doc8	2	1	0	0
doc9	1	0	0	1
doc10	0	1	1	2
idf	0.737	0.515	0.737	0.737

4)

Similarity of doc2 and doc9:

$$\sqrt[4]{(0-1)^2+(1-0)^2+(0-0)^2+(0-1)^2} = \sqrt[4]{3} = \underline{\underline{1.316}}$$

Similarity of doc3 and doc9:

$$\sqrt[4]{(1-1)^2+(1-0)^2+(0-0)^2+(0-1)^2} = \sqrt[4]{2} = \underline{\underline{1.189}}$$

Similarity of doc5 and doc9:

$$\sqrt[4]{(0-1)^2+(0-0)^2+(1-0)^2+(1-1)^2} = \sqrt[4]{2} = \underline{\underline{1.189}}$$

Similarity of doc7 and doc9:

$$\sqrt[4]{(0-1)^2+(0-0)^2+(3-0)^2+(0-1)^2} = \sqrt[4]{12} = \underline{\underline{1.861}}$$

5)
q5 = ”cat”

$$
\begin{aligned}
sim(q5, doc1) &= \frac{1}{\sqrt{4} * \sqrt{1}} = 0.5 \\
sim(q5, doc2) &= 0 \\
sim(q5, doc3) &= \frac{1}{\sqrt{2} * \sqrt{1}} = 0.707 \\
sim(q5, doc4) &= \frac{2}{\sqrt{10} * \sqrt{1}} = 0.632 \\
sim(q5, doc5) &= 0 \\
sim(q5, doc6) &= \frac{1}{\sqrt{4} * \sqrt{1}} = 0.5 \\
sim(q5, doc7) &= 0 \\
sim(q5, doc8) &= \frac{2}{\sqrt{5} * \sqrt{1}} = 0.89 \\
sim(q5, doc9) &= \frac{1}{\sqrt{2} * \sqrt{1}} = 0.707 \\
sim(q5, doc10) &= 0
\end{aligned}
$$

Ranking:

1: Doc8

2: Doc3 and Doc9

3: Doc4

4: Doc6 and Doc1

5: Doc2, Doc5, Doc7 and Doc10

Probabilistic Models

1) Robertson-Jones uses statistics and odds, whilst BM25 uses term frequency and idf.

2) Rank the documents using the BM25 model

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{df_t} \right] \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b(L_d/L_{ave})) + tf_{td}}$$

We input for the given $k_1 = 1.2$ and $b = 0.75$, as well as we input idf for the logarithm as we will calculate it seperately:

$$RSV_d = \sum_{t \in q} idf_t \frac{2.2 * tf_{td}}{1.2((0.25) + 0.75(L_d/L_{ave})) + tf_{td}}$$

$$RSV_d = \sum_{t \in q} idf_t \frac{2.2 * tf_{td}}{0.3 + 0.9(L_d/L_{ave}) + tf_{td}}$$

We find the L_{ave} by finding the average length of the documents.

$$L_{ave} = \frac{31}{10} = 3.1$$

We update our formulae:

$$RSV_d = \sum_{t \in q} idf_t \frac{2.2 * tf_{td}}{0.3 + 0.9(L_d/3.1) + tf_{td}}$$

We also use our term document matrix for exercise 3.1
For q1 = "Cat Dog":
doc1:

$$\begin{aligned} RSV_1 &= 0.737\frac{2.2}{0.3 + 0.9(4/3.1) + 1} + 0.515\frac{2.2}{0.3 + 0.9(4/3.1) + 1} \\ &= \frac{1.6214}{2.4613} + \frac{1.133}{2.4613} \\ &= 0.6588 + 0.4603 \\ &= \underline{\underline{1.119}} \end{aligned}$$

doc2:

$$\begin{aligned} RSV_2 &= 0.515\frac{2.2 * 1}{0.3 + 0.9(1/3.1) + 1} \\ &= \frac{1.133}{1.590} \\ &= \underline{\underline{0.712}} \end{aligned}$$

doc3:

$$\begin{aligned} RSV_3 &= 0.737\frac{2.2}{0.3 + 0.9(2/3.1) + 1} + 0.515\frac{2.2}{0.3 + 0.9(2/3.1) + 1} \\ &= \frac{1.6214}{1.881} + \frac{1.133}{1.881} \\ &= \underline{\underline{1.465}} \end{aligned}$$

doc 4:

$$\begin{aligned}RSV_4 &= 0.737\frac{2.2*2}{0.3+0.9(6/3.1)+2}+0.515\frac{2.2*1}{0.3+0.9(6/3.1)+1}\\&=\frac{3.243}{4.042}+\frac{1.133}{3.042}\\&=\underline{\underline{1.175}}\end{aligned}$$

doc 5 and doc 7: 0, as it has no terms matching the query
doc 6: syntactically equal to doc 1, therefore gives same ranking score
doc 8:

$$\begin{aligned}RSV_8 &= 0.737\frac{2.2*2}{0.3+0.9(3/3.1)+2}+0.515\frac{2.2*1}{0.3+0.9(3/3.1)+1}\\&=\frac{3.243}{3.171}+\frac{1.133}{2.171}\\&=\underline{\underline{1.175}}\end{aligned}$$

doc 9:

$$\begin{aligned}RSV_9 &= 0.737\frac{2.2}{0.3+0.9(2/3.1)+1}\\&=\frac{1.6214}{1.881}\\&=\underline{\underline{0.862}}\end{aligned}$$

doc 10:

$$\begin{aligned}RSV_10 &= 0.515\frac{2.2*1}{0.3+0.9(4/3.1)+1}\\&=\frac{1.133}{2.461}\\&=\underline{\underline{0.460}}\end{aligned}$$

For q1 = "small":
Doc 2,3,7,8: These are 0 we know as they do not contain the search phrase
Doc 1:

$$\begin{aligned}RSV_1 &= 0.737\frac{2.2}{0.3+0.9(4/3.1)+1}\\&=\frac{1.6214}{2.4613}\\&=\underline{\underline{0.659}}\end{aligned}$$

Doc 4:

$$\begin{aligned} RSV_4 &= 0.737\frac{2.2}{0.3+0.9(6/3.1)+1} \\ &= \frac{1.6214}{3.042} \\ &= \underline{\underline{0.533}} \end{aligned}$$

doc 5:

$$\begin{aligned} RSV_3 &= 0.737\frac{2.2}{0.3+0.9(2/3.1)+1} \\ &= \frac{1.6214}{1.881} \\ &= \underline{\underline{0.862}} \end{aligned}$$

doc 9:

$$\begin{aligned} RSV_1 &= 0.737\frac{2.2*2}{0.3+0.9(4/3.1)+2} \\ &= \frac{3.2428}{3.4613} \\ &= \underline{\underline{0.937}} \end{aligned}$$

Finally, this gives us the following table of scores and ranks for both queries.

Doc	q1 BM25 score	q1 rank	q2 BM25 score	q2 rank
doc1	1.119	4t	0.659	4t
doc2	0.712	7	0	7t
doc3	1.465	2	0	7t
doc4	1.175	3	0.533	6
doc5	0	9t	0.862	2t
doc6	1.119	4t	0.659	4t
doc7	0	9t	0	7t
doc8	1.545	1	0	7t
doc9	0.862	6	0.862	2t
doc10	0.460	8	0.937	1