

Assignment 1

Task 1

1. Information retrieval is finding documents based on a search query where we are trying to find information about a subject. In this case we might need information about a subject a user would like to retrieve documents about a subject, documents about synonyms of the query and other information what may be useful to the user. Data retrieval on the other hand is more about getting the data exactly as searched for. With data retrieval you are only able to find text, but with IR you can also find images, videos, audio etc.
2. The difference between unstructured and structured data is how the data is stored and indexed. With structured data everything is stored with attributes and values so that querying this data is as simple as searching for it. With unstructured data this becomes much more difficult since none of the data is splitt into categories. Everything is stored as it is and there is no standarised format on how all the documents look. One document can have a title, author, and a lot of meta data but since this isn't categorized its much more difficult differentiation this vs just text in the document.

Task 2



1. **Term Frequency (tf):** This concept is used when trying to decide how useful this document is to explain the term. By checking how many times this term is repeated in a document it's much more likely that if the document refers to the term many times this term is described in detail and will be more useful for the enduser.
2. **Document Frequency (df):** This is on the other side of the matrix where you are finding how many documents a term is referred in. This can be useful when trying to find what terms are "popular".
3. **Inverse Document Frequency (idf):** Is used to find more rare terms in a document. This will measure how important the term is in the text document. The idea is that the less frequent the term appears in the collection, the more informative the term is.

4. **idf** is important for term weighing because it measures how rare the term is in the document. And based on this rarity you can also measure how informative this term is in the text documents. This is of course useful when trying to find what document can give most value to the user based on the search terms.


Task 3.1

Subtask 1

Document Collection

 Document	 Tags
<u>1</u>	Big,Cat,Small,Dog
<u>2</u>	Dog
<u>3</u>	Cat,Dog
<u>4</u>	Big,Cat,Big,Small,Cat,Dog
<u>5</u>	Big,Small
<u>6</u>	Small,Cat,Dog,Big
<u>7</u>	Big,Big,Big
<u>8</u>	Dog,Cat,Cat
<u>9</u>	Cat,Small
<u>10</u>	Small,Small,Big,Dog

Q1

 Document	<input checked="" type="checkbox"/> Cat	<input checked="" type="checkbox"/> Dog	Σ Valid Query
<u>1</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>2</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>3</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>4</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>5</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>6</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>7</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

<u>Aa</u> Document	<input checked="" type="checkbox"/> Cat	<input checked="" type="checkbox"/> Dog	Σ Valid Query
<u>8</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>9</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>10</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

As shown in the table above the query "Cat AND Dog" will return documents {1, 3, 4, 6, 8}. These documents contain both the term cat and dog.

Q2

<u>Aa</u> Document	<input checked="" type="checkbox"/> Cat	<input checked="" type="checkbox"/> Small	Σ Valid Query
<u>1</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>2</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>3</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>4</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>5</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>6</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>7</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>8</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>9</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>10</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

As shown in the table above the query "Cat AND Small" will return documents {1, 4, 6, 9}. These documents contain both the term cat and small.

Q3

<u>Aa</u> Document	<input checked="" type="checkbox"/> Dog	<input checked="" type="checkbox"/> Big	Σ Valid Query
<u>1</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>2</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>3</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>4</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>5</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>6</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

<u>Aa</u> Document	<input checked="" type="checkbox"/> Dog	<input checked="" type="checkbox"/> Big	Σ Valid Query
<u>7</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<u>8</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>9</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>10</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

As shown in the table above the query "Dog OR Big" will return all documents except for d9. These documents contain either the word dog or big.

Q4

<u>Aa</u> Document	<input checked="" type="checkbox"/> Dog	<input checked="" type="checkbox"/> Small	Σ Valid Query
<u>1</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>2</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>3</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>4</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>5</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>6</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>7</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<u>8</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<u>9</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<u>10</u>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

As shown in the table above the query "Dog NOR Small" will return documents {2, 3, 8}. These documents contains the term dog, but not the term small.

Q5

<u>Aa</u> Document	<input checked="" type="checkbox"/> Cat
<u>1</u>	<input checked="" type="checkbox"/>
<u>2</u>	<input type="checkbox"/>
<u>3</u>	<input checked="" type="checkbox"/>
<u>4</u>	<input checked="" type="checkbox"/>
<u>5</u>	<input type="checkbox"/>

<u>Aa</u> Document	<input checked="" type="checkbox"/> Cat
<u>6</u>	<input checked="" type="checkbox"/>
<u>7</u>	<input type="checkbox"/>
<u>8</u>	<input checked="" type="checkbox"/>
<u>9</u>	<input checked="" type="checkbox"/>
<u>10</u>	<input type="checkbox"/>

As shown in the table above the query "Cat" will return documents {1, 3, 4, 6, 8, 9}. These documents contains the term cat. Quite a simple query really.

Subtask 2

The vector model dimension is 4

Subtask 3

Term Frequency

<u>Aa</u> Document	# Cat	# Dog	# Big	# Small
<u>1</u>	1	1	1	1
<u>2</u>	0	1	0	0
<u>3</u>	1	1	0	0
<u>4</u>	2	1	2	1
<u>5</u>	0	0	1	1
<u>6</u>	1	1	1	1
<u>7</u>	0	0	2.58	0
<u>8</u>	2	1	0	0
<u>9</u>	1	0	0	1
<u>10</u>	0	1	1	2

Inverse Document Frequency

<u>Aa</u> Name	# Cat	# Dog	# Big	# Small
<u>N</u>	10	10	10	10
<u>df</u>	6	7	6	6

Aa Name	# Cat	# Dog	# Big	# Small
idf	0.74	0.51	0.74	0.74

TF-IDF = Term Frequency * Inverse Document Frequency

Aa Document	# Cat	# Dog	# Big	# Small
<u>1</u>	0.74	0.51	0.74	0.74
<u>2</u>	0	0.51	0	0
<u>3</u>	0.74	0.51	0	0
<u>4</u>	1.47	0.51	1.47	0.74
<u>5</u>	0	0	0.74	0.74
<u>6</u>	0.74	0.51	0.74	0.74
<u>7</u>	0	0	1.9	0
<u>8</u>	1.47	0.51	0	0
<u>9</u>	0.74	0	0	0.74
<u>10</u>	0	0.51	0.74	1.47

Subtask 4

Formula for Euclidean distance

$$d(d_1, d_2) = \sqrt{(d_{1,1} - d_{2,1})^2 + (d_{1,2} - d_{2,2})^2 + (d_{1,3} - d_{2,3})^2 + (d_{1,4} - d_{2,4})^2}$$

$$d(d_2, d_9) = \sqrt{(0 - 0.74)^2 + (0.51 - 0)^2 + (0 - 0)^2 + (0 - 0.74)^2} = \underline{\underline{1.16}}$$

$$d(d_3, d_9) = \sqrt{(0.74 - 0.74)^2 + (0.51 - 0)^2 + (0 - 0)^2 + (0 - 0.74)^2} = \underline{\underline{0.9}}$$

$$d(d_5, d_9) = \sqrt{(0 - 0.74)^2 + (0 - 0)^2 + (0.74 - 0)^2 + (0.74 - 0.74)^2} = \underline{\underline{1.05}}$$

$$d(d_7, d_9) = \sqrt{(0 - 0.74)^2 + (0 - 0)^2 + (1.9 - 0)^2 + (0 - 0.74)^2} = \underline{\underline{2.17}}$$

Subtask 5

$$similarity = \cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \ ||\vec{b}||}$$

$$\vec{q} = \{1, 0, 0, 0\}$$

$$\begin{aligned} d_8 &= \{2, 1, 0, 0\} \\ sim(d_8, q) &= \frac{(2 * 1 + 1 * 0 + 0 * 0 + 0 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{2^2 + 1^2 + 0^2 + 0^2}} = 0.89 \end{aligned}$$

$$\begin{aligned} d_3 &= \{1, 1, 0, 0\} \\ sim(d_3, q) &= \frac{(1 * 1 + 1 * 0 + 0 * 0 + 0 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{1^2 + 1^2 + 0^2 + 0^2}} = 0.71 \end{aligned}$$

$$\begin{aligned} d_9 &= \{1, 0, 0, 1\} \\ sim(d_9, q) &= \frac{(1 * 1 + 0 * 0 + 0 * 0 + 1 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{1^2 + 0^2 + 0^2 + 1^2}} = 0.71 \end{aligned}$$

$$\begin{aligned} d_4 &= \{2, 1, 2, 0\} \\ sim(d_4, q) &= \frac{(2 * 1 + 1 * 0 + 2 * 0 + 0 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{2^2 + 1^2 + 2^2 + 0^2}} = 0.66 \end{aligned}$$

$$\begin{aligned} d_1 &= \{1, 1, 1, 1\} \\ sim(d_1, q) &= \frac{(1 * 1 + 1 * 0 + 1 * 0 + 1 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{1^2 + 1^2 + 1^2 + 1^2}} = 0.5 \end{aligned}$$

$$\begin{aligned} d_6 &= \{1, 1, 1, 1\} \\ sim(d_6, q) &= \frac{(1 * 1 + 1 * 0 + 1 * 0 + 1 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{1^2 + 1^2 + 1^2 + 1^2}} = 0.5 \end{aligned}$$

$$\begin{aligned} d_2 &= \{0, 1, 0, 0\} \\ sim(d_2, q) &= \frac{(0 * 1 + 1 * 0 + 0 * 0 + 0 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{0^2 + 1^2 + 0^2 + 0^2}} = 0 \end{aligned}$$

$$\begin{aligned} d_5 &= \{0, 0, 1, 1\} \\ sim(d_5, q) &= \frac{(0 * 1 + 0 * 0 + 1 * 0 + 1 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{0^2 + 0^2 + 1^2 + 1^2}} = 0 \end{aligned}$$

$$\begin{aligned} d_7 &= \{0, 0, 3, 0\} \\ sim(d_7, q) &= \frac{(0 * 1 + 0 * 0 + 3 * 0 + 0 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{0^2 + 0^2 + 3^2 + 0^2}} = 0 \end{aligned}$$

$$d_{10} = \{0, 0, 2, 2\}$$

$$sim(d_{10}, q) = \frac{(0 * 1 + 0 * 0 + 2 * 0 + 2 * 0)}{\sqrt{1^2 + 0^2 + 0^2 + 0^2} * \sqrt{0^2 + 0^2 + 2^2 + 2^2}} = 0$$

Task 3.2

1. The main difference is that BM25 is a collection of many scoring features using many different formulas and parameters to weight documents. On the other hand the probabilistic model is mainly a model that should serve as a framework for future forms. The framework has several shortcomings that make it unsuitable as an actual algorithm for weighting documents.

Subtask 2

$$k = 1.2, b = 0.75$$

$$q_1 = \{Cat, Dog\}$$

$$q_2 = \{Small\}$$

▼ The score is ordered by query and by score

Query 1

$$score(d_8, q_1) =$$

$$(0.74 * \frac{2 * (1.2 + 1)}{2 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})} +$$

$$0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})})$$

$$= \underline{\underline{1.54}}$$

$$score(d_3, q_1) =$$

$$(0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} +$$

$$0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})})$$

$$= \underline{\underline{1.46}}$$

$$\begin{aligned}
& score(d_4, q_1) = \\
& (0.74 * \frac{2 * (1.2 + 1)}{2 + 1.2 * (1 - 0.75 + 0.75 * \frac{6}{3.1})} + \\
& 0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{6}{3.1})}) \\
& = \underline{\underline{1.17}}
\end{aligned}$$

$$\begin{aligned}
& score(d_1, q_1) = \\
& (0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} + \\
& 0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})}) \\
& = \underline{\underline{1.12}}
\end{aligned}$$

$$\begin{aligned}
& score(d_6, q_1) = \\
& (0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} + \\
& 0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})}) \\
& = \underline{\underline{1.12}}
\end{aligned}$$

$$\begin{aligned}
& score(d_9, q_1) = \\
& (0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} + \\
& 0.51 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})}) \\
& = \underline{\underline{0.87}}
\end{aligned}$$

$$\begin{aligned}
& score(d_2, q_1) = \\
& (0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{1}{3.1})} + \\
& 0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{1}{3.1})}) \\
& = \underline{\underline{0.71}}
\end{aligned}$$

$$\begin{aligned}
& score(d_{10}, q_1) = \\
& (0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} + \\
& 0.51 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})}) \\
& = \underline{\underline{0.46}}
\end{aligned}$$

$$\begin{aligned}
& score(d_5, q_1) = \\
& (0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} + \\
& 0.51 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})}) \\
& = \underline{\underline{0}}
\end{aligned}$$

$$\begin{aligned}
& score(d_7, q_1) = \\
& (0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})} + \\
& 0.51 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})}) \\
& = \underline{\underline{0}}
\end{aligned}$$

Query 2

$$score(d_{10}, q_2) = 0.74 * \frac{2 * (1.2 + 1)}{2 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} = \underline{\underline{0.94}}$$

$$score(d_5, q_2) = 0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} = \underline{\underline{0.87}}$$

$$score(d_9, q_2) = 0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} = \underline{\underline{0.87}}$$

$$score(d_1, q_2) = 0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} = \underline{\underline{0.66}}$$

$$score(d_6, q_2) = 0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{4}{3.1})} = \underline{\underline{0.66}}$$

$$score(d_4, q_2) = 0.74 * \frac{1 * (1.2 + 1)}{1 + 1.2 * (1 - 0.75 + 0.75 * \frac{6}{3.1})} = \underline{\underline{0.54}}$$

$$score(d_2, q_2) = 0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{1}{3.1})} = \underline{\underline{0}}$$

$$score(d_3, q_2) = 0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{2}{3.1})} = \underline{\underline{0}}$$

$$score(d_7, q_2) = 0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})} = \underline{\underline{0}}$$

$$score(d_8, q_2) = 0.74 * \frac{0 * (1.2 + 1)}{0 + 1.2 * (1 - 0.75 + 0.75 * \frac{3}{3.1})} = \underline{\underline{0}}$$