Norwegian University of Science and Technology

Assignment Title

Assignment 1

-

Basic Definitions, IR Models, Term Weighting

Course

TDT4117 Information Retrieval

Semester

FALL 2018

Date Submitted

30/09/2018

Submitted by

Dusan Jakovic

Task 1: Basic Definitions

Explain the main differences between:

1. Information Retrieval vs Data Retrieval

Information retrieval: the information is about a subject or topic, semantics is frequently loose, small errors are tolerated, produces multiple results with ranking (partial match is allowed).

Data retrieval (database management system): which documents that contains a set of keywords, well defined semantics, a single erroneous object implies failure, and it produces exact results or no results if no exact match is found.

2. Structured Data vs Unstructured Data

Structured Data: Clearly defined data types whose pattern makes them easily searchable. Used for relational databases and data warehouses, e.g: phone numbers, addresses, dates, credit card numbers, etc..

Unstructured Data: Data that is usually not as easily searchable. it is essentially everything else, e.g.: text files, website, media (images, video files, audio files), reports, etc..

Task 2: IR Models

Given the following document collection containing words from the set O = {Autumn,Winter, Spring}, answer the questions in subtasks 1.1 and 1.2

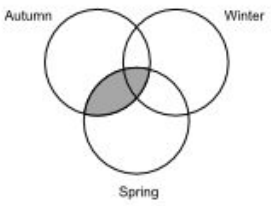
doc_1 = {Winter Autumn Spring Spring}
doc_2 = {Spring Autumn}
doc_3 = {Winter Winter Autumn}
doc_4 = {Autumn Autumn Winter Autumn Spring Winter Spring Autumn}
doc_5 = {Winter}
doc_6 = {Spring Spring}
doc_7 = {Winter Autumn Spring}
doc_8 = {Spring Spring Autumn }
doc_9 = {Spring Spring Winter}
doc_{10} = {Winter Autumn Winter Spring}

SubTask 1.1: Boolean Model and Vector Space Model

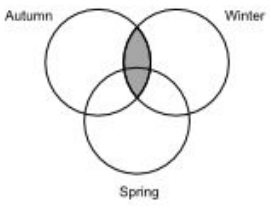
Given the following queries:

q_1 = "Spring AND Autumn"
q_2 = "Autumn AND Winter"
q_3 = "Winter OR Spring"
q_4 = "Autumn NOT Spring"
q_5 = "Winter"

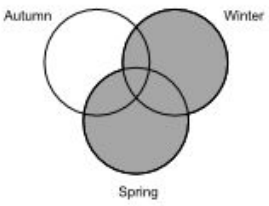
1. Which of the documents will be returned as the result for the above queries using the Boolean model? Explain your answers and draw a gure to illustrate.



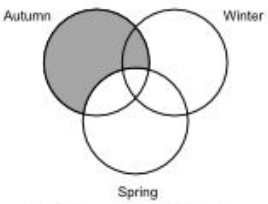
q1: Spring AND Autumn



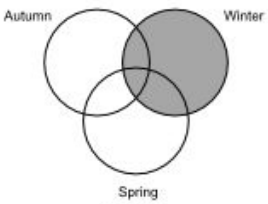
q2:Autumn AND Winter



q3:Winter OR Spring



q4:Autumn NOT Spring



q5: Winter

Following table will illustrate each which of the documents that will be returned.
'1' = returned
'0' = not returned

	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7	doc_8	doc_9	doc_{10}
q1	1	1	0	1	0	0	1	1	0	1
q2	1	0	1	1	0	0	1	0	0	1
q3	1	1	1	1	1	1	1	1	1	1
q4	0	0	1	0	0	0	0	0	0	0
q5	1	0	1	1	1	0	1	0	1	1

2. What is the dimension of the vector space representing this document collection when you use the vector model and how is it obtained?

There are three unique terms in the document collection, therefore the dimension is 3.

3. Calculate the weights for the documents and the terms using tf and idf weighting. Put these values into a document-term-matrix. (Tip: use the equations in the book and state which one you used.)

Term frequency tf: tf (1 + log2(fi,j))

#	term	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$f_{i,5}$	$f_{i,6}$	$f_{i,7}$	$f_{i,8}$	$f_{i,9}$	$f_{i,10}$
1	Autumn	1	1	1	4	0	0	1	1	0	1
2	Spring	2	1	0	2	0	2	1	2	2	1
3	Winter	1	0	2	2	1	0	1	0	1	2
Doc length		4	2	3	8	1	2	3	3	3	4

$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$	$TF_{i,5}$	$TF_{i,6}$	$TF_{i,7}$	$TF_{i,8}$	$TF_{i,9}$	$TF_{i,10}$
1	1	1	3	-	-	1	1	-	1
2	1	-	2	-	2	1	2	2	1
1	-	2	2	1	-	1	-	1	2

Inverse Document Frequency: *idf($log_2(N/df_t)$:*

#	term	n_i	IDF = $\log_2(N/n_i)$
1	Autumn	7	0.515
2	Spring	8	0.322
3	Winter	7	0.515

Term Frequency - Inverse Document Frequency: *TF-IDF = tf * idf*

#	Term	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
1	Autumn	0.515	0.515	0.515	1.545	-	-	0.515	0.515	-	0.515
2	Spring	0.644	0.322	-	0.644	-	0.644	0.322	0.644	0.644	0.322
3	Winter	0.515	-	1.03	1.03	0.515	-	0.515	-	0.515	1.03

4. Study the documents 1, 2, 4 and 10 and compare them to document 5. Calculate the similarity between document 5 and these four documents according to Euclidean distance. (Use tf -idf weights for your computations).

Euclidean distance (d(p,q) = $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$ **)**

$d(d_5,d_1) = \sqrt{(0 - 0.515)^2 + (0 - 0.644)^2 + (0.515 - 0.515)^2} = 0.825$

$d(d_5,d_2) = \sqrt{(0 - 0.515)^2 + (0 - 0.322)^2 + (0.515 - 0)^2} = 0.796$

$d(d_5,d_4) = \sqrt{(0 - 1.545)^2 + (0 - 0.644)^2 + (0.515 - 1.03)^2} = 0.972$

$d(d_5,d_{10}) = \sqrt{(0 - 0.515)^2 + (0 - 0.322)^2 + (0.515 - 1.03)^2} = 0.796$

5. Rank the documents for query q5 using cosine similarity.

q = q_5 = {Winter}

$$sim(dj, q) = \frac{\vec{dj} \bullet \vec{q}}{|\vec{dj}| \times |\vec{q}|}$$

Calculation example for the cosine similarity (d_1,q):

$$sim(d_1, q) = \frac{(1 * 0) + (2 * 0) + (1 * 1)}{\sqrt{1^2 + 2^2 + 1^2} * \sqrt{0^2 + 0^2 + 1^2}} = 0.408$$

$sim(d_j, q) : \downarrow$

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
$sim(d_j,q)$	0.408	0	0.894	0.408	1	0	0.577	0	0.447	0.816

#	d
1	d_5

#	d
2	d_3

#	#
3	d_{10}

#	d
4	d_7

#	d
5	d_9

#	d
6	d_1

#	d
7	d_4

#	d
8	d_2

#	d
9	d_6

#	d
10	d_8

SubTask 1.2: Probabilistic Models

1. What are the main differences between BM25 and the probabilistic model introduced by Robertson-Jones?

Roberson-Spark-Jones Probabilistic Model:
The probability that the user will find the document d_j relevant.
The model has some flaws:

- Relevance to the user might be affected by variables *outside* the system. Thus, the ideal answer set produced using information available to the system might not be ideal from the point of view of the user.
- The principle does not state explicitly how to compute the probabilities of relevance.
- Taking the *odds of relevance* as the rank minimizes the probability of an erroneous judgement

BM25 (Best Match 25)
The BM25 is a combination of the *BM11* and *BM15* ranking formulas. Contrary to the original conceptions in the classic probabilistic model, BM25 formula uses parameters and can be computed without any relevance information provided by the user, i.e., in fully automatic fashion.

2. Rank the documents using the BM25 model. Set the parameters to k = 1.2, b = 0:75 (Here we assume relevance information is not provided.)

Given the following queries:

q_1 = “Winter Spring
q_2 = “Autumn”

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i,D) * (k_1+1)}{f(q_i,D) + k_1 * ((1 - b + b * (\frac{|D|}{Lave}))}$$

L = 33, L_{ave}= 3.3.

Calculation example using BM25 model (d_1,q):

$$score(d_1,q_1) + score(d_1,q_1) = 0.515 * \frac{1 * (1.2+1)}{1 + (1.2 * (1-0.75) + 0.75 * \frac{4}{3.3})} + 0.322 * \frac{2 * (1.2+1)}{2 + (1.2 * (1-0.75) + 0.75 * \frac{4}{3.3})}$$

$$= 0.892$$

$$score(d_1,q_2) = 0.515 * \frac{1 * (1.2+1)}{1 + (1.2 * (1-0.75) + 0.75 * \frac{4}{3.3})} = 0.474$$

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
$sim(d_j,q_1)$	**0.892**	**0.384**	**0.535**	**0.936**	**0.720**	**0.498**	**0.867**	**0.534**	**0.989**	**0.770**

#	d
1	d_9

#	d
2	d_4

#	#
3	d_1

#	d
4	d_7

#	d
5	d_{10}

#	d
6	d_5

#	d
7	d_3

#	d
8	d_8

#	d
9	d_6

#	d
10	d_2

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
$sim(d_j,q)_2$	**0.474**	**0.614**	**0.727**	**0.506**	**0**	**0**	**0.534**	**0.454**	**0**	**0.668**

#	d
1	d_3

#	d
2	d_{10}

#	#
3	d_2

#	d
4	d_7

#	d
5	d_4

#	d
6	d_1

#	d
7	d_8

#	d
8	d_5

#	d
9	d_6

#	d
10	d_9

Task 3: Term Weighting

Explain:

1. Term Frequency (tf)
   The number of times a term occurring in the given document

2. Document Frequency
   The number of documents in the collection that contains a term

3. Inverse Document Frequency (idf)
   How important a term is.

4. Why *idf* is important for term weighting
   It diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.