

Bilkent University CS533: Information Retrieval Systems

Assignment – 2

Q1) To be clear in the a,b and c options of this question, all recall and precision values for Q1 and Q2 are shown in following tables:

Q1: r=5 (The total number of relevant document)

Doc.No.	1	2	3	4	5	6	7	8	9	10
Relevancy	+	-	+	-	+	+	-	-	+	-
Precision	1	1/2	2/3	2/4	3/5	4/6	4/7	4/8	5/9	5/10
Recall	1/5	1/5	2/5	2/5	3/5	4/5	4/5	4/5	5/5	5/5

Q2: r=5 (The total number of relevant document)

Doc.No.	1	2	3	4	5	6	7	8	9	10
Relevancy	-	+	+	-	-	+	-	-	+	-
Precision	0	1/2	2/3	2/4	2/5	3/6	3/7	3/8	4/9	4/10
Recall	0	1/5	2/5	2/5	2/5	3/5	/5	3/5	/5	4/5

Formulas:

Recall = (# of retrieved&relevant docs)/(#of relevant docs in collection)

Precision = (# of retrieved&relevant docs)/(#of retrieved docs)

are used for computation of the tables

a) R-precision = the precision at point r where r is the number of relevant docs = 5 for both queries

For Q1 -> R-prec = Precision@5 = 3/5 from the table

For Q2 -> R-prec = Precision@5 = 2/5 from the table

b) MAP (Mean Average Precision) for these queries:

MAP = $\frac{\sum_{k=1}^n p(k) * rel(k)}{\# \text{ of relevant docs}}$ where rel(k) stands for relevancy of k-th doc which is either 0 or 1 and p(k) stands for precision at position k

Then,

MAP-Q1 = (1+2/3+3/5+4/6+5/9)/5 = 0.6977

MAP-Q2 = (1/2+2/3+3/6+4/9)/5 = 0.4222

c) Using TP, FP, FN and TN -> **Precision** = $\frac{TP}{TP+FP}$ **Recall** = $\frac{TP}{TP+FN}$

where TP = retrieved&relevant

FP = retrieved but not relevant

FN = relevant but not retrieved. Then,

For Q1:

$$\text{Precision@10} = 5/(5+5) = 0.5$$

$$\text{Recall@10} = 5/5 = 1$$

For Q2:

$$\text{Precision@10} = 4/(4+6) = 0.4$$

$$\text{Recall@10} = 4/(4+1) = 0.8$$

Q2) We know that $\text{Precision}(P) = \frac{TP}{TP+FP}$ $\text{Recall}(R) = \frac{TP}{TP+FN}$ $F = \frac{2PR}{P+R}$

$$\text{Then, } F = \frac{2PR}{P+R} = \frac{2 * \frac{TP}{TP+FP} * \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} = \frac{2 * TP * TP}{(TP+FP) * (TP+FN)} * \frac{(TP+FP) * (TP+FN)}{TP(TP+FN+TP+FP)} = \frac{2TP}{2TP+FP+FN}$$

Q3)

a) As S is a mxm matrix with having all diagonal as 1 ($S_{ii} = 1$) and the simetric values are equal to each other since it represents the similarity ($S_{ij} = S_{ji}$), we only need to calculate upper (or lower) triangular. Thus, for m=6 documents, we need to calculate $5+4+3+2+1 = 15$ values (or simply, $m(m-1)/2 = 6*5/2 = 15$)

b) We can derive the term inverted list from the given documents x terms matrix as:

$$t_1 \rightarrow \langle 1,1 \rangle, \langle 3,1 \rangle, \langle 4,1 \rangle$$

$$t_2 \rightarrow \langle 2,1 \rangle$$

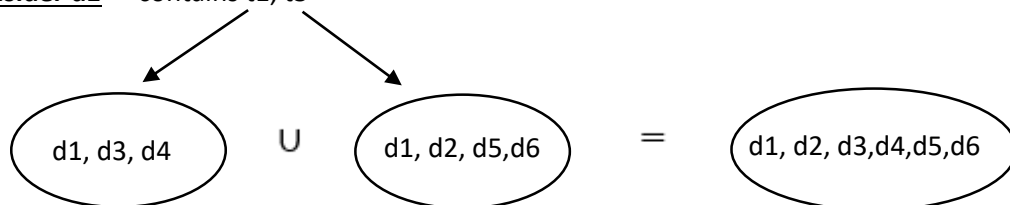
$$t_3 \rightarrow \langle 3,1 \rangle, \langle 5,1 \rangle$$

$$t_4 \rightarrow \langle 2,1 \rangle, \langle 3,1 \rangle, \langle 4,1 \rangle$$

$$t_5 \rightarrow \langle 1,1 \rangle, \langle 2,1 \rangle, \langle 5,1 \rangle, \langle 6,1 \rangle$$

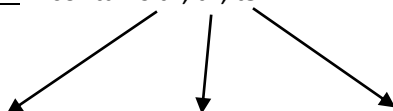
$$t_6 \rightarrow \langle 5,1 \rangle, \langle 6,1 \rangle \quad \text{where } \langle x,y \rangle \text{ represents } x\text{-th document and } y \text{ number of occurrence}$$

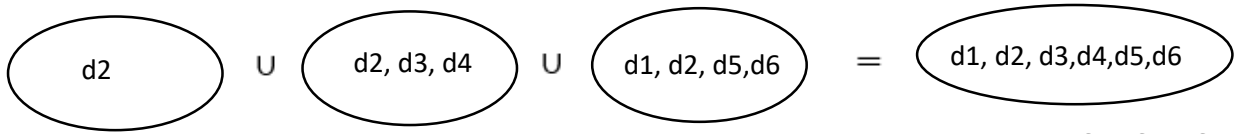
Consider d1 -> contains t1, t5



Thus; calculate $S_{12}, S_{13}, S_{14}, S_{15}, S_{16}$

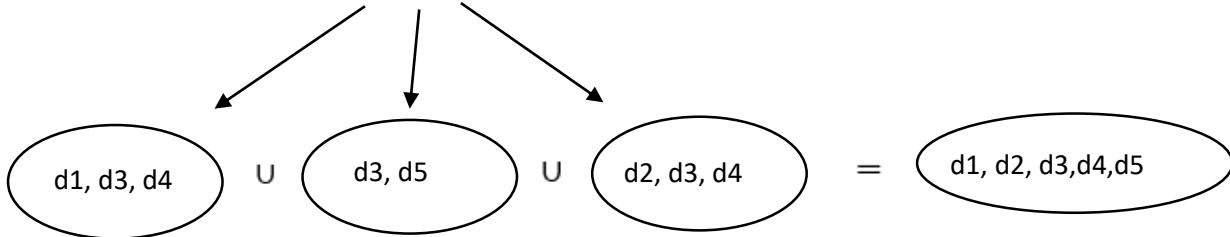
Consider d2 -> contains t2, t4, t5





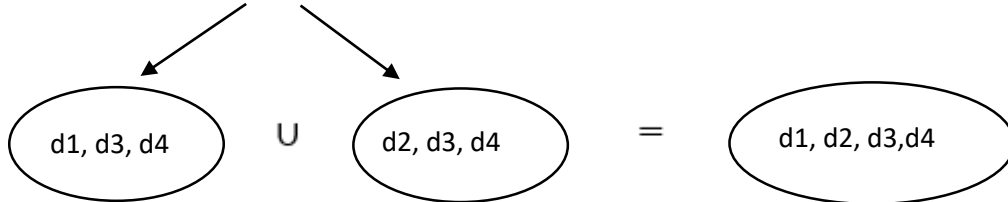
Thus; calculate $S_{23}, S_{24}, S_{25}, S_{26}$

Consider d3 -> contains t1, t3, t4



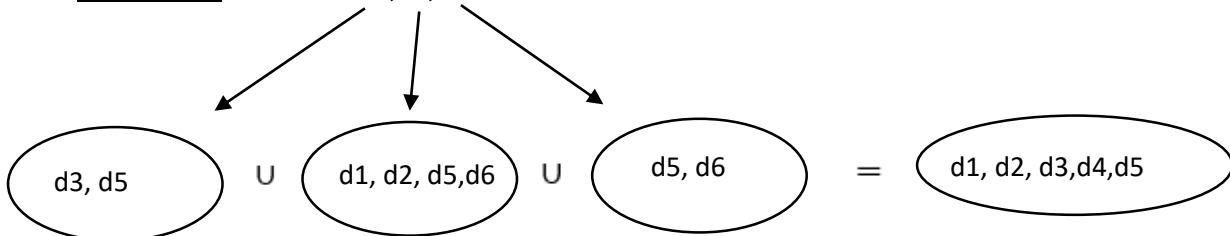
Thus; calculate S_{34}, S_{35}

Consider d4 -> contains t1, t4



No calculation needed.

Consider d5 -> contains t3, t5, t6



Thus; calculate S_{56}

Consider d6 -> no calculation needed.

Total: $S_{12}, S_{13}, S_{14}, S_{15}, S_{16}, S_{23}, S_{24}, S_{25}, S_{26}, S_{34}, S_{35}, S_{56}$ -> **12 number of calculation needed.**

c) Dice coefficient = $\frac{2|X \cap Y|}{|X| + |Y|}$ is used in this option.

We already found the necessary S_{ij} entries in option b. We need document length information which is:

d1	d2	d3	d4	d5	d6
2	3	3	2	3	2

In each step in the previous operation (option b) we get the following similarities:

For d1 ->

X	1(due to t5)	1(due to t1)	1(due to t1)	1(due to t5)	1(due to t5)
---	--------------	--------------	--------------	--------------	--------------

Final similarity value using Dice coefficient for d1 ->

S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
X	$\frac{2 * 1}{2 + 3} = \frac{2}{5}$	$\frac{2 * 1}{2 + 3} = \frac{2}{5}$	$\frac{2 * 1}{2 + 2} = \frac{2}{4}$	$\frac{2 * 1}{2 + 3} = \frac{2}{5}$	$\frac{2 * 1}{2 + 2} = \frac{2}{4}$

For d2 ->

X	X	1(due to t4)	1(due to t4)	1(due to t5)	1(due to t5)
---	---	--------------	--------------	--------------	--------------

Final similarity value using Dice coefficient for d2 ->

S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}
X	X	$\frac{2 * 1}{3 + 3} = \frac{2}{6}$	$\frac{2 * 1}{3 + 2} = \frac{2}{5}$	$\frac{2 * 1}{3 + 3} = \frac{2}{6}$	$\frac{2 * 1}{3 + 2} = \frac{2}{5}$

For d3 ->

X	X	X	2(due to t1 and t4)	1(due to t4)	0
---	---	---	---------------------	--------------	---

Final similarity value using Dice coefficient for d3 ->

S_{31}	S_{32}	S_{33}	S_{34}	S_{35}	S_{36}
X	X	X	$\frac{2 * 2}{3 + 2} = \frac{4}{5}$	$\frac{2 * 1}{3 + 3} = \frac{2}{6}$	0

For d4 -> X (No calculation needed).

For d5 ->

X	X	X	X	X	2(due to t5 and t6)
---	---	---	---	---	---------------------

Final similarity value using Dice coefficient for d5 ->

S_{51}	S_{52}	S_{53}	S_{54}	S_{55}	S_{56}
----------	----------	----------	----------	----------	----------

X	X	X	X	X	$\frac{2 * 2}{3 + 2} = \frac{4}{5}$
---	---	---	---	---	-------------------------------------

Thus, our similarity matrix is:

$$S = \begin{bmatrix} 1 & 0.4 & 0.4 & 0.5 & 0.4 & 0.5 \\ X & 1 & 0.33 & 0.4 & 0.33 & 0.4 \\ X & X & 1 & 0.8 & 0.33 & 0 \\ X & X & X & 1 & 0 & 0 \\ X & X & X & X & 1 & 0.8 \\ X & X & X & X & X & 1 \end{bmatrix}$$

Q4)

a) term a: <1, 2> <3, 2> <9, 5> <10, 3> <12, 4> <17, 4> <18, 3>, <22, 2> <24, 4> <33, 4> <38, 5>

<43, 4> <55, 3><64, 2> <68, 4> <72, 3> <75, 1> <88, 2>

term-b: <12, 7> <22, 7><45, 3> <66, 1> -> **CORRECTION:** I ignored the first <66,3> term since it should be in sorted order

Without skipping: Using the algorithm that we discussed in class;

For <12,7> in term b -> compare <1, 2> <3, 2> <9, 5> <10, 3> <12, 4> -> 5 comparisons (and increase both x and y index)

For <22,7> in term b -> compare <17, 4> <18, 3>, <22, 2> -> 3 comparisons (and increase both x and y index)

For <45,3> in term b -> compare <24, 4> <33, 4> <38, 5> <43, 4> <55, 3>-> 5 comparison (and increase y index)

For <66,1> in term b -> compare <55, 3><64, 2> <68, 4> -> 3 comparison (and increase y index, end)

Thus, total of 5+3+5+3 = 16 comparisons without skipping.

With skipping structure given in the question:

Chunk1 : <1, 2> <3, 2> <9, 5> <10, 3> <12, 4>

Chunk2 : <17, 4> <18, 3>, <22, 2> <24, 4> <33, 4>

Chunk3 : <38, 5> <43, 4> <55, 3><64, 2> <68, 4>

Chunk4 : <72, 3> <75, 1> <88, 2>

For <12,7> in term b -> Is it in chunk1? Yes -> 1comparison

To find <12,4> -> 5 more comparison

For <22,7> in term b -> Is it in chunk1? No -> 1comparison

Is it in chunk2? Yes -> 1comparison

To find <22,2> -> 3 more comparison

For <45,3> in term b -> Is it in chunk1? No -> 1comparison

Is it in chunk2? No -> 1comparison

Is it in chunk3? Yes -> 1comparison

Go through until <55,3> -> 3 more comparison

For <66,1> in term b -> Is it in chunk1? No -> 1comparison

Is it in chunk2? No -> 1comparison

Is it in chunk3? Yes -> 1comparison

Go through all chunk until <68,4> -> 5 more comparison

In total = 6+5+6+8 = 25 comparisons needed with skips

	Large Skips	Small Skips
Advantages	-The total number of chunks decreases -The number of comparisons with chunk descriptors decreases	-More chunks can be skipped -Less comparisons within the chunk since number of docs in the chunk decreases

Disadvantages	- The number of comparison within the chunk increases	-More comparison with the chunk descriptors since number of chunks increases.
----------------------	---	---

Information from [1] is used to construct the above table.

b) No, I think we can't take the advantage of using skipping for disjunctive (OR) queries. Because in disjunctive queries; for example, while looking at (term-a || term-b), we only need union of the posting lists because we need the documents in which either term a or term b takes place, thus; looking for common terms is useless in disjunctive queries.

c) a) ordered by $f_{d,t}$ (the frequency of term t in document d)

term a: <9, 5>, <38, 5>, <12, 4>, <17, 4>, <24, 4> <33, 4>, <43, 4>, <68, 4>, <10, 3>, <18, 3>, <55,3>, <72, 3>, <1, 2>, <3, 2>, <22, 2>, <64,2>, <88, 2>, <75, 1>

b) ordered by frequency information in prefix form

term a: <5:2: 9, 38> <4:6: 12, 17, 24, 33, 43, 68> <3:4: 10,18, 55,72> <2:5: 1, 3, 22, 64,88> <1:1: 75>

	$f_{d,t}$	ordered by frequency information in prefix form
Advantages	-Query processing time is improved -Can increase performance if a frequency threshold is used	-Query processing time is improved again - Saves spaces (no need to keep same frequency information)
Disadvantages	-More comparisons to find the document	-More comparisons to find the document again -Difficult to construct

Q5) The components of information retrieval test collection according to [2] are:

-Document collection

-A set of queries

- Relevance information about each document with respect to each query

In the pooling approach, as its name suggests, top p ranked documents by assessors (gathered from different IR systems) are collected (pooled) for identification of the relevant documents for each query. Thus, documents in the pool are treated as relevant documents. Here, p being the pool-depth, Zobel has some criticism about the pooling approach.

Firstly, Zobel mentions that pooling approach can be biased giving the example of fixed-depth pools; it can show a new system that is combination of two approaches as better system. Thus, Zobel emphasizes the importance of decreasing the bias when pooling approach is used in this paper. He also mentions that when pool-depth increases the possibility “to obtain useful estimates of the likely numbers of new relevant documents that can be discovered for each query” without introducing bias. Secondly, Zobel thinks pooling as a fair approach because each system has same number of documents for assessment. Lastly, he states some disadvantages of pooling approach such as being not very useful for systems aiming to maximize recall. Also if the pool size is not appropriate he states the system reinforcement (the documents retrieved by two systems can reinforce each other and result in underestimation of the effectiveness of the third system) and system omission (a technique that didn’t contribute to the pool may be found as ineffective while it may be).

Q6) a) Stages of clustering defined in [3]:

- 1) *Pattern representation*: It refers to the number of classes/available patterns, and also the number, type, scale of the features available for clustering algorithm. This step may include feature extraction (transformations of the input features are used to produce new remarkable features) and/or selection (selection of the most effective/useful features)
- 2) *Definition of a pattern proximity measure appropriate to the data domain*: Using various techniques to define pattern similarity.
- 3) *Clustering/Grouping*: This is the most important step in clustering (according to me) since it is the stage where actual clustering takes place. The authors state different grouping and clustering methods and their different outputs while explaining this step.
- 4) *Data Abstraction* (if needed): As the name suggests this step is for extraction of representatives (cluster prototypes or centroids) from data set.
- 5) *Assessment of output* (if needed): This stage is for understanding if the output is useful and to what degree it is effective to use that clustering algorithm, thus assessment of the output. The authors mention the cluster validity analysis which are external assessment, internal assessment and relative

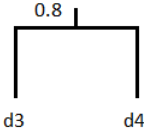
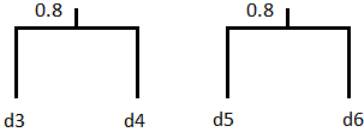
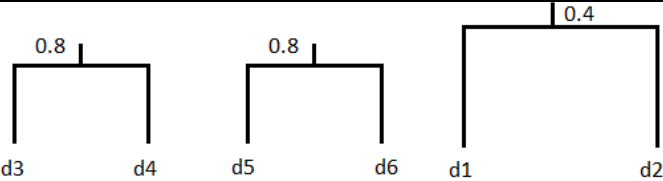
test. Also, the definition of clustering tendency is firstly stated in assessment of output stage which will be given in detail in part b of this question.

b) Clustering tendency refers to examination of input data to see if its suitable to a cluster analysis. In other words, real dataset may suggest only 2 clusters or no clusters at all but a clustering algorithm may give lots of meaningless clusters. Thus, it is useful to use clustering tendency in some stages of clustering since it may stop the stages if there is no meaningful cluster to make. I would suggest the use of similarity matrix in some random stages of clustering algorithm by setting a threshold value to check in similarity matrix and reorder the similarity matrix as we cluster; I borrowed this idea from VAT algorithm [4] by literature search in which dissimilarity matrix (DM) is used but I replaced it with similarity matrix and setting threshold value.

Q7) To apply the complete link, we first need the pair similarity values as sorted:

Pair	Similarity
d3-d4	0.8
d5-d6	0.8
d1-d4	0.5
d1-d6	0.5
d1-d2	0.4
d1-d3	0.4
d1-d5	0.4
d2-d4	0.4
d2-d6	0.4
d2-d3	0.33
d2-d5	0.33
d3-d5	0.33
d3-d6	0
d4-d5	0
d4-d6	0

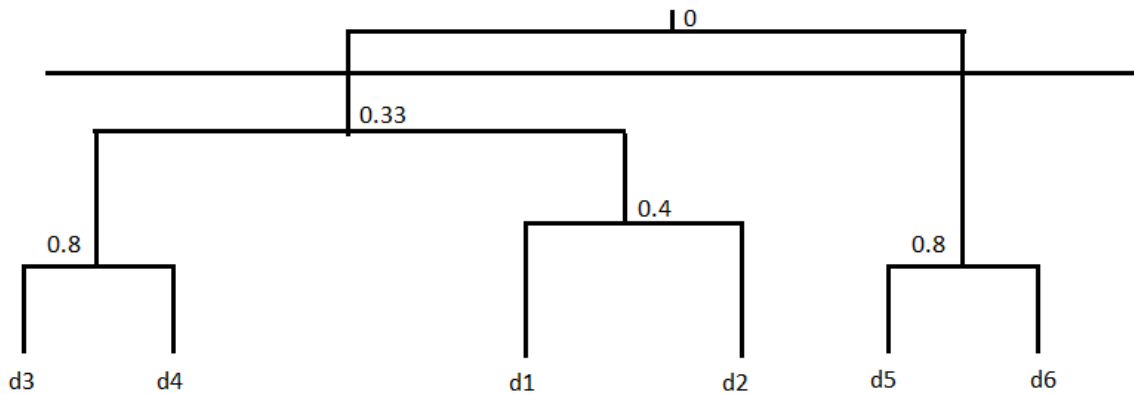
Now, using complete-link;

Step	Pair	Similarity	Complete-link structure	Pairs covered
1	d3-d4	0.8		(d3-d4)
2	d5-d6	0.8		(d3-d4), (d5-d6)
3	d1-d4	0.5	Too early to connect, we don't know (d1-d3)	(d3-d4), (d5-d6), (d1-d4)
4	d1-d6	0.5	Too early to connect, we don't know (d1-d5)	(d3-d4), (d5-d6), (d1-d4), (d1-d6)
5	d1-d2	0.4		(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2)
6	d1-d3	0.4	Too early to connect, we don't know (d2-d3) and (d2-d4)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3)
7	d1-d5	0.4	Too early to connect, we don't know (d2-d5) and (d2-d6)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5)
8	d2-d4	0.4	Too early to connect, we don't know (d2-d3)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4)
9	d2-d6	0.4	Too early to connect, we don't know (d2-d5)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6)

10	d2-d3	0.33		(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3)
11	d2-d5	0.33	Too early to connect, we don't know (d3-d5), (d3-d6), (d4-d5) and (d4-d6)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3), (d2-d5)
12	d3-d5	0.33	Too early to connect, we don't know (d3-d6), (d4-d5) and (d4-d6)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3), (d2-d5), (d3-d5)
13	d3-d6	0	Too early to connect, we don't know (d4-d5) and (d4-d6)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3), (d2-d5), (d3-d5), (d3-d6)
14	d4-d5	0	Too early to connect, we don't know (d4-d6)	(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3), (d2-d5)

				(d2-d5), (d3-d5), (d3-d6), (d4-d5)
15	d4- d6	0		(d3-d4), (d5-d6), (d1-d4), (d1-d6), (d1-d2), (d1-d3), (d1-d5), (d2-d4), (d2-d6), (d2-d3), (d2-d5), (d3-d5), (d3-d6), (d4-d5), (d4-d5)

Q8) Yes, we can obtain a partitioning clustering structure with two cluster from the above dendrogram by using threshold value between 0 and 0.33 and cut from that threshold. Below diagram shows how we get two cluster by cutting the dendrogram:



Q9) We know that $n_c = \sum_{i=1}^m c_{ii}$ (sum of diagonal in C matrix)

We also have $c_{ij} = \alpha_i * \sum_{k=1}^n d_{ik} * \beta_k * d_{jk}$ and $c'_{ij} = \beta_i * \sum_{k=1}^m d_{ki} * \alpha_k * d_{kj}$

Then, for $n_c = \sum_{i=1}^m c_{ii}$

$$\begin{aligned}
 &= \sum_{i=1}^m \alpha_i * \sum_{k=1}^n d_{ik} * \beta_k * d_{ik} \\
 &= \sum_{i=1}^m \sum_{k=1}^n \alpha_i * \beta_k * d_{ik} * d_{ik} \\
 &= \sum_{k=1}^n \sum_{i=1}^m \alpha_i * \beta_k * d_{ik} * d_{ik}
 \end{aligned}$$

$$= \sum_{k=1}^n \beta_k \sum_{i=1}^m \alpha_i * d_{ik} * d_{ik}$$

$$= \sum_{i=1}^n c'_{ii}$$

Hence, number of clusters implied by C and C' is the same.

Q10)

Double stage probability experiment

For d1 → {t1, t5} prob = 1/2

→ t1 covers {d1, d3, d4} prob = 1/3

→ t5 covers {d1, d2, d5, d6} prob = 1/4

$$C_{11} = \frac{1}{2} * \frac{1}{3} + \frac{1}{2} * \frac{1}{4} = 0.291$$

$$C_{12} = \frac{1}{2} * \frac{1}{4} = 0.125$$

$$C_{13} = \frac{1}{2} * \frac{1}{3} = 0.166$$

$$C_{14} = \frac{1}{2} * \frac{1}{3} = 0.166$$

$$C_{15} = \frac{1}{2} * \frac{1}{4} = 0.125$$

$$C_{16} = \frac{1}{2} * \frac{1}{4} = 0.125$$

For d2 → {t2, t4, t5} prob = 1/3

→ t2 covers {d2 } prob = 1

→ t4 covers {d2, d3, d4 } prob = 1/3

→ t5 covers {d1, d2, d5, d6} prob = 1/4

$$C_{21} = \frac{1}{3} * \frac{1}{4} = 0.083$$

$$C_{22} = \frac{1}{3} + \frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{4} = 0.527$$

$$C_{23} = \frac{1}{3} * \frac{1}{3} = 0.111$$

$$C_{24} = \frac{1}{3} * \frac{1}{3} = 0.111$$

$$C_{25} = \frac{1}{3} * \frac{1}{4} = 0.083$$

$$C_{26} = \frac{1}{3} * \frac{1}{4} = 0.083$$

For d3 → {t1, t3, t4} prob = 1/3

→ t1 covers {d1, d3, d4} prob = 1/3

- ➔ t3 covers {d3,d5} prob = 1/2
- ➔ t4 covers {d2, d3, d4 } prob = 1/3

$$C_{31} = \frac{1}{3} * \frac{1}{3} = 0.111$$

$$C_{32} = \frac{1}{3} + \frac{1}{3} = 0.111$$

$$C_{33} = \frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * \frac{1}{3} = 0.388$$

$$C_{34} = \frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{3} = 0.222$$

$$C_{35} = \frac{1}{3} * \frac{1}{2} = 0.166$$

$$C_{36} = 0$$

For d4 ➔ {t1, t4} prob = 1/2

- ➔ t1 covers {d1,d3,d4} prob = 1/3
- ➔ t4 covers {d2, d3, d4 } prob = 1/3

$$C_{41} = \frac{1}{2} * \frac{1}{3} = 0.166$$

$$C_{42} = \frac{1}{2} * \frac{1}{3} = 0.166$$

$$C_{43} = \frac{1}{2} * \frac{1}{3} + \frac{1}{2} * \frac{1}{3} = 0.333$$

$$C_{44} = \frac{1}{2} * \frac{1}{3} + \frac{1}{2} * \frac{1}{3} = 0.333$$

$$C_{45} = 0$$

$$C_{46} = 0$$

For d5 ➔ {t3, t5, t6} prob = 1/3

- ➔ t3 covers {d3,d5} prob = 1/2
- ➔ t5 covers {d1,d2, d5,d6} prob = 1/4
- ➔ t6 covers {d5,d6} prob = 1/2

$$C_{51} = \frac{1}{3} * \frac{1}{4} = 0.083$$

$$C_{52} = \frac{1}{3} * \frac{1}{4} = 0.083$$

$$C_{53} = \frac{1}{3} * \frac{1}{2} = 0.166$$

$$C_{54} = 0$$

$$C_{55} = \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * \frac{1}{4} + \frac{1}{3} * \frac{1}{2} = 0.416$$

$$C_{56} = \frac{1}{3} * \frac{1}{4} + \frac{1}{3} * \frac{1}{2} = 0.25$$

For d6 → { t5, t6} prob = 1/2

→ t5 covers {d1,d2, d5,d6} prob = 1/4

→ t6 covers {d5,d6} prob = 1/2

$$C_{61} = \frac{1}{2} * \frac{1}{4} = 0.125$$

$$C_{62} = \frac{1}{2} * \frac{1}{4} = 0.125$$

$$C_{63} = 0$$

$$C_{64} = 0$$

$$C_{65} = \frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{1}{2} = 0.375$$

$$C_{66} = \frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{1}{2} = 0.375$$

Hence, C matrix is;

$$C = \begin{bmatrix} 0.291 & 0.125 & 0.166 & 0.166 & 0.125 & 0.125 \\ 0.083 & 0.527 & 0.111 & 0.111 & 0.083 & 0.083 \\ 0.111 & 0.111 & 0.388 & 0.222 & 0.166 & 0 \\ 0.166 & 0.166 & 0.333 & 0.333 & 0 & 0 \\ 0.083 & 0.083 & 0.166 & 0 & 0.416 & 0.25 \\ 0.125 & 0.125 & 0 & 0 & 0.375 & 0.375 \end{bmatrix}$$

nc = 0.291 + 0.527 + 0.388 + 0.333 + 0.416 + 0.375 ≈ 2 clusters to generate.

Seed power of di:

$$P_1 = 0.291 * (1 - 0.291) * 2 \approx 0.413$$

$$P_2 = 0.527 * (1 - 0.527) * 3 \approx 0.748$$

$$P_3 = 0.388 * (1 - 0.388) * 3 \approx 0.712$$

$$P_4 = 0.333 * (1 - 0.333) * 2 \approx 0.444$$

$$P_5 = 0.416 * (1 - 0.416) * 3 \approx 0.729$$

$$P_6 = 0.375 * (1 - 0.375) * 2 \approx 0.469$$

Thus, our cluster seeds are d2 and d5 -> Cluster1: d1,d2, d4 (As $C_{12} > C_{15}$ and $C_{42} > C_{45}$)

-> Cluster2: d3, d5, d6 (As $C_{35} > C_{32}$ and $C_{65} > C_{62}$)

References

- [1] Zobel, J. and Moffat, A. (2006). Inverted files for text search engines. ACM Computing Surveys, 38(2), p.6-es.
- [2] Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98.
- [3] Jain, A., Murty, M. and Flynn, P. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), pp.264-323.
- [4] Assessing clustering tendency: A vital issue - Unsupervised Machine Learning - Easy Guides - Wiki - STHDA. [online] Available at: <http://www.sthda.com/english/wiki/assessing-clustering-tendency-a-vital-issue-unsupervised-machine-learning#why-assessing-clustering-tendency> [Accessed 2 Nov. 2016].