1) Explain the main differences between:

1. Information Retrieval vs Data Retrieval

IR, Information Retrieval, handler om å finne informasjon som er relevant for brukeren. IR ønsker å bestemme hvilke dokumenter som ofte nok inneholder nok nøkkelord til å tilfredsstille informasjonsbehovet til brukeren. Vi kan kalle IR søket for noe unøyaktig og kan kalles en suksess selv med mindre feil i søket.

Data Retrieval, DR, handler om å finne nøyaktig matching med søket. Ved bruk av DR er tilstedeværelse av feildata sett på som en «failure». DR systemer består av velstrukturerte datastrukturer som ofte kan finnes med spørrespråk (SQL)

2. Structured Data vs Unstructured Data

Strukturert data refererer til informasjon som er lagret på en strukturert og organisert måte, som f.eks. en relasjonsdatabase. Strukturert data kan effektivt søkes i igjennom spørrespråk og søkemotorer.

Ustrukturert data referer til informasjon som ikke er organisert etter noe fast struktur. Et relevant eksempel på dette kan være informasjon fra mail eller andre uformelle brev. Ettersom disse ofte har skrivefeil, lite brukte ord/synonymer og «slang» fra talespråk kan det være utfordrende for søkemotorer og maskiner å finne den informasjonen som brukeren vil finne relevant.

2) Explain:

1. Term Frequency (tf)

Term frequency refererer til hvor ofte en term (ord/begrep) dukker opp i et dokument, ofte notert som $tf_{t,d}$. Det er viktig å merke seg at relevans ikke er proporsjonal med frekvensen av et ord i dokumentet. En term som gjentas 10 ganger i et dokument gir ikke 10 ganger så mye relevans som om den kun oppstod 1 gang.

2. Document Frequency (df)

Document frequency refererer til hvor ofte en term finnes i en kolleksjon av dokumenter. Ofte brukte termer som «i», «av», «til» osv. (stop words) vil ha en høy dokumentfrekvens og vil ofte ikke ha høy relevans til den informasjonen vi er ute etter.

3. Inverse Document Frequency (idf)

idf er en vektingsmetode for termer i et dokument. Idf finnes ved formelen $IDF_i = log\frac{N}{n_i}$ der N er alle dokumentene og n_i er antall dokumenter som inneholder termen vi er interessert i

4. Why idf is important for term weighting

Idf er viktig for term vekting ettersom det kan gi oss en indikator på hvor sannsynlig det er at et termen vi er ute etter er inneholdt i dokument-universet vi søker i. Vi ser også at veldig mange vektings-modeller bruker idf, eller en variant, i utregningen sin.

3) IR Models:

Given the following document collection containing words from the set O = {Big, Cat, Small, Dog}, answer the questions in subtasks 3.1 and 3.2.
doc1 = {Big Cat Small Dog}
doc2 = {Dog}
doc3 = {Cat Dog}
doc4 = {Big Cat Big Small Cat Dog}
doc5 = {Big Small}
doc6 = {Small Cat Dog Big }
doc7 = {Big Big Big}
doc8 = {Dog Cat Cat }
doc9 = {Cat Small }
doc10 = {Small Small Big Dog}

SubTask 3.1: Boolean Model and Vector Space Model

Given the following queries:
q1 = "Cat AND Dog"
q2 = "Cat AND Small"
q3 = "Dog OR Big"
q4 = "Dog NOT Small"
q5 = "Cat"

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 figure to illustrate.

q1: {doc1, doc3, doc4, doc6, doc8} will be returned as relevant

q2: {doc1, doc4, doc6, doc9} will be returned as relevant

q3: {doc1, doc2, doc3, doc4, doc5, doc6, doc7, doc8, doc10} will be returned as relevant

q4: {doc2, doc3, doc8} will be returned as relevant

q5: {doc1, doc3, doc4, doc6, doc8, doc9} will be returned as relevant

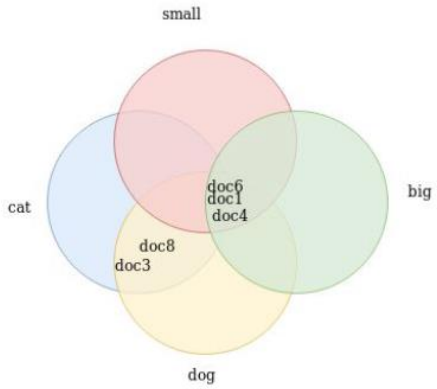

diagram for q1

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

The keyword set is {Big, Cat, Dog, Small} = {t1, t2, t3, t4}

	{t1	t2	t3	t4}
d1:	**[1**	**1**	**1**	**1]**
d2:	**[0**	**0**	**1**	**0]**
d3:	**[0**	**1**	**1**	**0]**
d4:	**[2**	**2**	**1**	**1]**
d5:	**[1**	**0**	**0**	**1]**
d6:	**[1**	**1**	**1**	**1]**
d7:	**[3**	**0**	**0**	**0]**
d8:	**[0**	**2**	**1**	**0]**
d9:	**[0**	**1**	**0**	**1]**
d10:	**[1**	**0**	**1**	**2]**

By using gaussian elimination we will conclude that the vector space has 4 dimensions

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.)

we use the following formula form the book:

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Term Frequency for each document, (not weighted)

	j	1	2	3	4	5	6	7	8	9	10
i	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	Big	1	0	0	2	1	1	3	0	0	1
2	Cat	1	0	1	2	0	1	0	2	1	0
3	Dog	1	1	1	1	0	1	0	1	0	1
4	Small	1	0	0	1	1	1	0	0	1	2

Here we see the occurrence of each term in each document, in our next matrix we will apply the formula mentioned above for each $f_{i,j}$

TF weighting:

	j	1	2	3	4	5	6	7	8	9	10
i	Term	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$	$TF_{i,5}$	$TF_{i,6}$	$TF_{i,7}$	$TF_{i,8}$	$TF_{i,9}$	$TF_{i,10}$
1	Big	1	0	0	2	1	1	2.585	0	0	1
2	Cat	1	0	1	2	0	1	0	2	1	0
3	Dog	1	1	1	1	0	1	0	1	0	1
4	Small	1	0	0	1	1	1	0	0	1	2

1 + $log_2(1)$ = 1 + 0 = 1
1 + $log_2(2)$ = 1 + 1 = 2
1 + $log_2(3)$ = 1 + 1.585 = 2.585

IDF weighting:

i	Term	n_i	$IDF_i = log\frac{N}{n_i}$
1	Big	6	0.737
2	Cat	6	0.737
3	Dog	7	0.515
4	Small	6	0.737

N = 10 og er hvor mange dokumenter vi har og n_i er antall dokumenter som inneholder termen vi er interessert i

$log\frac{10}{6}$ = 0.737, $log\frac{10}{7}$ = 0.515

TF-IDF weighting:

	j	1	2	3	4	5	6	7	8	9	10
i	Term	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$	$TF_{i,5}$	$TF_{i,6}$	$TF_{i,7}$	$TF_{i,8}$	$TF_{i,9}$	$TF_{i,10}$
1	Big	0.737	0	0	1.474	0.737	0.737	1.905	0	0	0.737
2	Cat	0.737	0	0.737	1.474	0	0.737	0	1.474	0.737	0
3	Dog	0.515	0.515	0.515	0.515	0	0.515	0	0.515	0	0.515
4	Small	0.737	0	0	0.737	0.737	0.737	0	0	0.737	1.474

Calculated by: $\left(1 + log\left(f_{i,j}\right)\right)$ X $log\frac{N}{n_i}$

4. Study the documents 2, 3, 5 and 7 and compare them to document 9. Calculate the similarity between document 9 and these four documents according to Euclidean distance. (Use tf -idf weights for your computations).

TF weighting for documents 2, 3, 5, 7 and 9
d2 = [0, 0, 0.515, 0]
d3 = [0, 0.737, 0.515, 0]
d5 = [0.737, 0, 0, 0.737]
d7 = [1.905, 0, 0, 0]
d9 = [0.737, 0, 0.515, 1.474]

$S_{9,2}$: $\sqrt{(0.737 - 0)\text{^}2 + (0 - 0)\text{^}2 + (0.515 - 0.515)\text{^}2 + (1.474 - 0)\text{^}2}$ **= 1.648**

$S_{9,3}$: $\sqrt{(0.737 - 0)\text{^}2 + (0 - 0.737)\text{^}2 + (0.515 - 0.515)\text{^}2 + (1.474 - 0)\text{^}2}$ **= 1.805**

$S_{9,5}$: $\sqrt{(0.737 - 0.737)\text{^}2 + (0 - 0)\text{^}2 + (0.515 - 0)\text{^}2 + (1.474 - 0.737)\text{^}2}$ **= 0.899**

$S_{9,7}$: $\sqrt{(0.737 - 1.905)\text{^}2 + (0 - 0)\text{^}2 + (0.515 - 0)\text{^}2 + (1.474 - 0)\text{^}2}$ **= 1.950**

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

q5 = "Cat"

$$sim(d_j, q) = \frac{\vec{d_j} * \vec{q}}{|\vec{d_j}||\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2}\sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

d_j is the document we are comparing, q = [0, 1, 0, 0] = {cat} = q5 is the query

The keyword set is {Big, Cat, Dog, Small}

$sim(d_1, q5) = \frac{[1*0+1*1+1*0+1*0]}{\sqrt{1^2+1^2+1^2+1^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{1}{\sqrt{4}\sqrt{1}} = \frac{1}{\sqrt{4}} = \frac{1}{2} = \mathbf{0.5}$

$sim(d_2, q5) = \frac{[0*0+0*1+1*0+0*0]}{\sqrt{0^2+0^2+1^2+0^2}*\sqrt{0^2+1^2+0^2+0^2}} = \mathbf{0}$

$sim(d_3, q5) = \frac{[0*0+1*1+1*0+0*0]}{\sqrt{0^2+1^2+1^2+0^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{1}{\sqrt{2}\sqrt{1}} = \frac{1}{\sqrt{2}} = \mathbf{0.707}$

$sim(d_4, q5) = \frac{[2*0+2*1+1*0+1*0]}{\sqrt{2^2+2^2+1^2+1^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{2}{\sqrt{10}\sqrt{1}} = \frac{2}{\sqrt{10}} = \mathbf{0.632}$

$sim(d_5, q5) = \frac{[1*0+0*1+0*0+1*0]}{\sqrt{1^2+0^2+0^2+1^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{0}{\sqrt{2}\sqrt{1}} = \mathbf{0}$

$sim(d_6, q5)$ has the same terms as $sim(d_1, q5)$, $sim(d_6, q5)$ **= 0.5**

$sim(d_7, q5) = \frac{[3*0+0*1+0*0+0*0]}{\sqrt{3^2+0^2+0^2+0^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{0}{\sqrt{9}\sqrt{1}} = \mathbf{0}$

$sim(d_8, q5) = \frac{[0*0+2*1+1*0+0*0]}{\sqrt{0^2+2^2+1^2+0^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{2}{\sqrt{5}\sqrt{1}} = \frac{2}{\sqrt{5}} = \mathbf{0.894}$

$sim(d_9, q5) = \frac{[0*0+1*1+0*0+1*0]}{\sqrt{0^2+1^2+0^2+1^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{1}{\sqrt{2}\sqrt{1}} = \frac{1}{\sqrt{2}} = \mathbf{0.707}$

$sim(d_{10}, q5) = \frac{[1*0+0*1+1*0+2*0]}{\sqrt{1^2+0^2+1^2+2^2}*\sqrt{0^2+1^2+0^2+0^2}} = \frac{0}{\sqrt{6}\sqrt{1}} = \mathbf{0}$

SubTask 3.2: Probabilistic Models

Given the following queries:

q1 = "Cat Dog"

q2 = "Small"

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

$$sim_{BM25}(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \quad (3.41)$$

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{p_{iR}}{1 - p_{iR}} \right) + \log \left(\frac{1 - q_{iR}}{q_{iR}} \right) \quad (3.15)$$

$sim(d_j, q)$ is the probabilistic model and is based on trying to estimate a "answer set". This requires some relevance information provided by the user (interaction with the user)

$sim_{\mathrm{BM25}}(d_j, q)$ is the BM25 model and is an extension of the probabilistic model which incorporates the 3 main main features from the vector model (inverse document frequency, term frequency and document length normalization). Whereas $sim(d_j, q)$ only covers the principle of inverse document frequency.

An important difference is that $sim_{\mathrm{BM25}}(d_j, q)$ can be computed in fully automatic fashion (requires no relevance information by user) as vector model theory, and derived different variables in the formula, replaces the need for it.

2. Rank the documents using the BM25 model. Set the parameters to k = 1:2 and b = 0:75. (Here we assume relevance information is not provided.)

2. Rank the documents using the BM25 model. Set the parameters to $k = 1.2$ and $b = 0.75$. (Here we assume relevance information is not provided.)
Hint: To avoid getting negative numbers, you need to use $idf = \log\left[\frac{N}{df_t}\right]$ in the BM25 model.

$$L_{ave} = (4 + 1 + 2 + 6 + 2 + 4 + 3 + 3 + 2 + 4)/10 = 3.1$$

Formelen fra boka:

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen}\right] + f_{i,j}} \quad (3.40)$$

$$sim_{BM25}(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \quad (3.41)$$

Formelen fra slides:

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

Documents for q1:

$$\mathrm{RSV_{d1}} = (0.737 + 0.515) * \frac{(1.2 + 1) * 1}{1.2\left((1 - 0.75) + 0.75 \times (4/3.1)\right) + 1} =$$

$$\mathrm{RSV_{d1}} = (1.252) * \frac{2.2}{1.2(0.25 + 0.968) + 1} = \mathbf{1.119}$$

Note: this works because tf_{td} = 1 for “cat” and “dog” in d1

$$\mathrm{RSV}_{d2} = (0.737) * \frac{(1.2+1)*0}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+0}$$

$$+\ (0.515) * \frac{(1.2+1)*1}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+1}$$

$$=\ 0.515 * \frac{(1.2+1)*1}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+1} = \mathbf{0.460}$$

RSV_{d3} has the same case as RSV_{d1}, RSV_{d3} = $\mathbf{1.119}$

$$\mathrm{RSV}_{d4} = (0.737) * \frac{(1.2+1)*2}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+2}$$

$$+\ (0.515) * \frac{(1.2+1)*1}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+1}$$

$$=\ (0.737)*(1.271) + (0.515)*(0.460) = \mathbf{1.174}$$

d5 has neither of the terms in the query, $\mathrm{RSV}_{d5} = \mathbf{0}$

$\mathrm{RSV}_{d6} =$ **1.119**, as we have the same situation as for RSV_{d1}

d7 has neither of the terms in the query, $\mathrm{RSV}_{d7} = \mathbf{0}$

$\mathrm{RSV}_{d8} =$ **1.174**, as we have the same situation as for RSV_{d4}

$\mathrm{RSV}_{d9} = (0.737) * \frac{(1.2+1)*1}{1.2\big((1-0.75)+0.75\times(4/3.1)\big)+1} = \mathbf{0.568}$

$\mathrm{RSV}_{d10} =$ **0.460**, as we have the same situation as for RSV_{d2}

Documents for q2:

$$\mathrm{RSV_{d1}} = (0.737) * \frac{(1.2+1)*1}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+1} = \mathbf{0.656}$$

$$\mathrm{RSV_{d2}} = (0.737) * \frac{(1.2+1)*0}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+0} = \mathbf{0}$$

$$\mathrm{RSV_{d3}} = (0.737) * \frac{(1.2+1)*0}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+0} = \mathbf{0}$$

$$\mathrm{RSV_{d4}} = (0.737) * \frac{(1.2+1)*1}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+1} = \mathbf{0.656}$$

$$\mathrm{RSV_{d5}} = (0.737) * \frac{(1.2+1)*1}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+1} = \mathbf{0.656}$$

$$\mathrm{RSV_{d6}} = (0.737) * \frac{(1.2+1)*1}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+1} = \mathbf{0.656}$$

$$\mathrm{RSV_{d7}} = (0.737) * \frac{(1.2+1)*0}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+0} = \mathbf{0}$$

$$\mathrm{RSV_{d8}} = (0.737) * \frac{(1.2+1)*0}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+0} = \mathbf{0}$$

$$\mathrm{RSV_{d9}} = (0.737) * \frac{(1.2+1)*1}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+1} = \mathbf{0.656}$$

$$\mathrm{RSV_{d10}} = (0.737) * \frac{(1.2+1)*2}{1.2\left((1-0.75)+0.75\times(4/3.1)\right)+2} = \mathbf{0.937}$$

IDF weighting used :

i	Term	n_i	$IDF_i = log\frac{N}{n_i}$
1	Big	6	0.737
2	Cat	6	0.737
3	Dog	7	0.515
4	Small	6	0.737