which will rea..n.
who.

22nd June, 2020.
17k-3730.
E GR-1.

INFORMATION RETRIEVAL
FINAL EXAM - 2020

Question # 01:

(a)

1. Limitations of boolean retreival model:

Some limitations of boolean retreival model are

- Exact matching → boolean retreival model only allows exact matching and doesn't support the retreival based on partial matching.
- The results are not based on scoring → all documents have equal importance which is not a very good approach in some cases.
- Hard query formation/translation into boolean expression.

2. Major problems observed in VSM:

Words are ~~put~~ kept in a "bag of word" which means ~~they~~ their order is lost and they are treated in statistically independent manner. There can be alot of zero elements in the documents vectors which can lead to large amount of data sparsity in vectors, which wastes large amount of

problem of symentic 17k-3730

memory. The ~~problem of~~ symentic ambiguity may occur. Longer documents may have higher frequency.

3- False negatives in a VSM

False negatives are where the ~~document~~ model incorrectly predicts negative class In a VSM, sometimes the word substrings/ stemmed similar words (i.e: dolled up or doll become doll after stemming) might result is false positive match. The search keywords should, therefore, match precisely to doc. terms.

(dolled up → getting ready; doll → toy)

4- Drawback of VSM from human standpoint:
From a human perspective, the sentence and documents have a context and the words are connected with each other having some context while, in a VSM terms are treated statistically independent. The problem of synonym and polysemy may also occur which is also a drawback.
Statistically independent refers that order of the terms is lost.

5.

(a)

It's one of the assumptions of PRP that if the word is not present in query it's equally likely to occur in R and NR populations. Assuming this, we calculate the probabilities of the common words of query and docs.

(b)

$d_1 = <0.04, 0.04, 0, 0.06, 0, 0, 0>$

$d_2 = <0, 0.11, 0.16, 0, 0, 0.07, 0>$

$d_3 = <0.01, 0.07, 0, 0, 0, 0, 0.03>$

Query vector:

$q = <0, 0, 1, 1, 0, 0, 1>$

$$\cos(d_1, q) = \frac{d_1 \cdot q}{|d_1||q|} = \frac{0.06}{(0.08)(1.73)} = 0.43.$$

$$\cos(d_2, q) = \frac{d_2 \cdot q}{|d_2| \cdot |q|} = \frac{0.16}{(0.21)(1.73)} = 0.44$$

$$\cos(d_3, q) = \frac{d_3 \cdot q}{|d_3| \cdot |q|} = \frac{0.03}{(0.07)(1.73)} = 0.24$$

Document ranking:

D_2, D_1, D_3.

Question #02

(a)

~~Getting~~ The two quantities, precision and recall clearly trade off against one another. If we talk about the IR system development perspective, there can be two types of applications; One system that is precision sensitive ie web search and one that is recall sensitive i.e: legal/ patent seach. In these type of IR systems having precision or recall greater than each other is tolerable to a certain percentage inorder to serve the demand/need. F-measure can be used, which is HM of precision and recall.

(b)

Break even point in IR systems are evaluation measure that returns a list of documents based on their order of relevance score according the user's requirement/need.

There has to be a breakeven point between precision and recall. This can be explained as:

If ranked element is NR, then tp=0

which will result in trivial break-even point where precision = recall. If highest relevant element is ranked and Relevant docs are found in list, then fn decreases and false positive doesn't change. Therefore the document list is ordered based on the false positives and false negatives i.e

$fp < fn$

$fn < fp$

→ so, there has to be a breakeven point in the list.

(c)

1) Precision $= \frac{tp}{tp + fp}$

$tp = 5$, $fp = 7$, $fn = 8 - 5 = 3$

$$\text{precision} = \frac{5}{7+5} = \frac{5}{12}$$

precision = 0.41667

2) finding recall to get the F_1 score.

$$\text{Recall} = \frac{tp}{fn + fp} = \frac{5}{3+5} = \frac{5}{8}$$

Recall = 0.625

now, finding f_1 s

PERS12.

$$F_1 = \frac{2PR}{P+R} = \frac{2 \times 0.4166 \times 0.625}{0.41667 \times 0.625}$$

$$\boxed{F_1 = 0.5}$$

3) $$MAP = \frac{1}{5}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{9} + \frac{5}{12}\right) = 0.6922$$

$$\boxed{MAP = 0.6922}$$

4) Largest possible MAP would be:

$$MAP = \frac{1}{8}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{9} + \frac{5}{12} + \frac{6}{13} + \frac{67}{14} + \frac{8}{15}\right)$$

$$\boxed{MAP = 0.619}$$

5) Smallest possible MAP:

→ when remaining 3 docs retrieved from collection (8-5=3).

$$MAP = \frac{1}{8}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{9} + \frac{5}{12} + \frac{6}{998} + \frac{7}{999} + \frac{8}{1000}\right)$$

$$\boxed{MAP = 0.435}$$

QUESTION # 03

	D1	D2	D3	D4	Test D5
Taipei	1	0	0	0	0
Taiwan	1	1	0	1	2
Macao	0	1	0	0	0
Shanghai	0	1	0	0	0
Japan	0	0	1	0	0
Sapporo	0	0	1	1	1
Osaka	0	0	0	1	0

a) $\cos(d_1, d_5) = \frac{\vec{d_1} \cdot \vec{d_5}}{|d_1||d_5|}$

$= \frac{\langle 1,1,0,0,0,0,0 \rangle \langle 0,2,0,0,0,1,0 \rangle}{\sqrt{2} \cdot \sqrt{5}}$

$= \frac{2}{\sqrt{10}} = 0.6324$

b)

$= \frac{\langle 0,1,1,1,0,0,0 \rangle \langle 0,2,0,0,0,1,0 \rangle}{\sqrt{3} \cdot \sqrt{5}}$

$= \frac{2}{\sqrt{15}} = 0.5164$

c)

$$= \frac{\langle 0,0,0,0,1,1,0\rangle \langle 0,2,0,0,0,1,0\rangle}{\sqrt{2} \cdot \sqrt{5}}$$

$$= \frac{1}{\sqrt{10}}$$

$$= 0.3162.$$

d)

$$= \frac{\langle 0,1,0,0,0,1,1\rangle \langle 0,2,0,0,0,1,0\rangle}{\sqrt{3} \cdot \sqrt{5}}$$

$$= \frac{3}{\sqrt{15}}$$

$$= 0.7746$$

(i) for 3-NN, the query doc (D_5) belongs to class "Yes".

(ii) Rocchio's Algorithm:

Centroids:

a) $$\mu_{china=yes} = \frac{1}{D_{china=yes}} \sum_{d_j \in D_{china=yes}} \vec{v}(d_j)$$

$$= \frac{1}{2} (\langle 1,1,0,0,0,0,0\rangle \langle 0,1,1,1,0,0,0\rangle)$$

$$= \frac{1}{2} \langle 1,2,1,1,0,0,0\rangle$$

$= \langle 0.5, 1, 0.5, 0.5, 0, 0, 0 \rangle$

b) $\mu_{china=no} = \frac{1}{D_{china=no}} \sum_{dj \in D_{china=no}} \vec{v}(dj)$

$= \frac{1}{2}(\langle 0,1,0,0,0,1,1 \rangle + \langle 0,0,0,0,1,1,0 \rangle)$

$= \frac{\langle 0,1,0,0,1,2,1 \rangle}{2}$

$= \langle 0, 0.5, 0, 0, 0.5, 1, 0.5 \rangle$

for Doc 5 :

$|\vec{\mu}_{yes} - \vec{d_5}| = \sqrt{(0.5)^2 + (1-2)^2 + 0.5^2 + 0.5^2 + 0 + 1 + 0}$

$= \sqrt{0.25 + 1 + 0.25 + 0.25 + 0 + 1 + 0}$

$= 1.658.$

$|\vec{\mu}_{no} - \vec{d_5}| = \sqrt{0 + (0.5-2)^2 + 0 + 0 + 0.5^2 + 0 + 0.5^2}$

$= \sqrt{0 + 2.25 + 0 + 0 + 0.25 + 0 + 0.25}$

$= 1.6583$

min (1.65, 1.65)

Result can be 'yes' or 'no'

→ If the result is same the higher prior class is usually considered.

(iii) Multinomial NB

Priors would be:

$P(Yes) = 2/4 = 0.5$

$P(No) = 2/4 = 0.5$

$P(w|c) = \frac{count(w,c) + 1}{count(c) + |V|}$

- $P(Taiwan|Yes) = 2+1/5+7 = 3/12$

  $P(Sapparo|Yes) = 0+1/5+7 = 1/12$

- $P(Taiwan|No) = 1+1/12 = 2/12$

  $P(Sapparo|No) = (1+2)/12 = 3/12$

So $P(y|d_5) = 0.5 \times \frac{3}{12} \times \frac{3}{12} \times \frac{1}{12}$

$= 2.60 \times 10^{-3}$

$P(no|d_5) = 0.5 \times \frac{2}{12} \times \frac{2}{12} \times \frac{3}{12}$

$= 0.264$

d_5 has the label 'no'

Q4

(a)

Let, $c_1 = d_2$ and $c_2 = d_5$

taking the c_1 and c_2 as initial cluster

so

for d1 :

distance $(c_1, d_1) = \sqrt{(1-2)^2 + (4-4)^2} = 1$

& $(c_2, d_1) = \sqrt{(1-2)^2 + (4-1)^2} = 3.16$

$c_1 < c_2$ so it belongs to c_1

for d3 :

distance $(c_1, d_3) = (4-2)^2 + (4-4)^2 = 2$

& $(c_2, d_3) = (4-2)^2 + (4-1)^2 = 3.6$

$c_1 < c_2$ so it belongs to c_1

for d4 :

$d(c_1, d_4) = (1-2)^2 + (1-4)^2 = 3.16$

$d(c_2, d_4) = (1-2)^2 + (1-1)^2 = 1$

$c_2 < c_1$ so belongs to c_2

for d6 :

$d(c_1, d_6) = 3.6$

$d(c_2, d_6) = 2$

$c_2 < c_1$

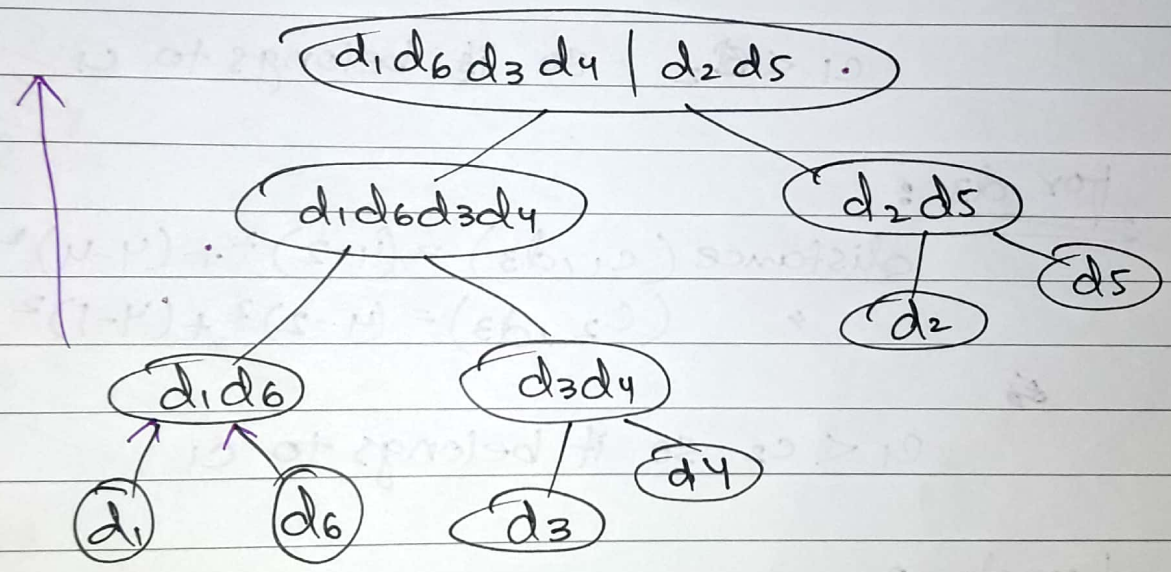
so c_2.

PEB3/2.

for k-mean the time complexity is $O(n^2)$ which can be improved through various techniques i.e: choosing the right stopping criteria.

(b)

HAC approach



	d_1	d_2	d_3	d_4	d_5	d_6
d_1	0	1	3	3	3.16	4.24 → max
d_2	1	0	2	3.16	3	3.6
d_3	3	2	0	4.24	3.6	3
d_4	3	3.16	4.24	0	1	3
d_5	3.16	3	3.6	1	0	2
d_6	4.2	3.6	3	3	2	0

	d_1d_6	d_2	d_3	d_4	d_5
1	2.12	2.3	3	3	2.55
2	2.3	0	2	3.16	3
3	3	2	0	4.24	3.6
4	3	3.16	(4.24)	0	1
5	2.56	3	3.6	1	0

max.

	d_1d_6	d_2	d_3d_4	d_5
d_1d_6	2.1	2.3	3	2.58
d_2	2.3	0	2.58	3
d_3d_4	3	2.58	2.12	2.12
d_5	2.58	(3)	2.12	0

max

	$d_1d_6d_3d_4$	d_2	d_5
$d_1d_6d_3d_4$	2.56	2.4	2.56
d_2	2.4	0	3
d_5	2.5	(3)	0

max

Complexity : n^3.

(c) No, we donot expect same result in part a, b - ~~As in k-mean centroid~~.

Q5

(a)

Type of web queries:

a) Navigational → user seach for single entity or directly go to the page i.e state, classroom, flex

b) Transactional → user go for online payments and transactions.

c) ~~Navigatio~~ Informational → user seach for specific topic or general information of specific topic.

(b)

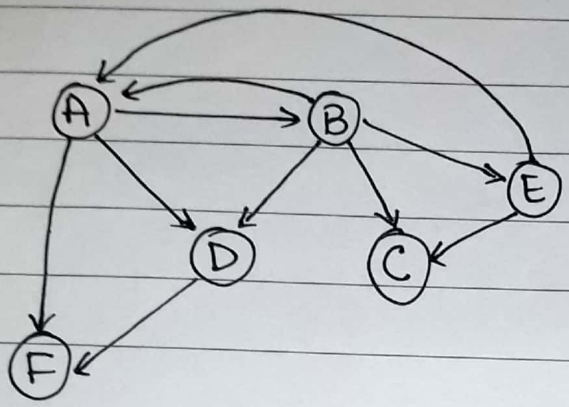
politeness : No. of searched web pages without permission

Solution: Policies for regulating the crawler

Freshness: The crawler should crawl page with the rate of change of that page.

Extensible: Crawler should be extensible to handle newly fetched protocols, new format etc.

Q6



In the given web graph edge C and F have no outgoing edge. Therefore, it's not always possible to redirect ~~from the page F and C~~ from all the pages (due to F and C), also from D there is no way to go to any page other than F.

	A	B	C	D	E	F
A	0	1/4	0	0	1/2	0
B	1/3	0	0	0	0	0
C	0	1/4	0	0	1/2	0
D	1/3	1/4	0	0	0	0
E	0	1/4	0	0	0	0
F	1/3	0	0	1	0	0

$$\begin{bmatrix} 0.85 \\ 0.85 \\ 0.85 \\ 0.85 \\ 0.85 \\ 0.85 \end{bmatrix}$$