Name— Ayesha Asaf
Roll no— 17K-3867
Section— GR2
Course— Information Retrieval

QUESTION 1

Part (a)

(1) Limitations are: (1) It doesn't allow more flexible matching operation. The exact matching either retrieves too many or less documents. (2) All terms are equally weighted which gives flat results & doesn't allow ranked retrieval. (3) It is difficult to formulate a query & process large data.

(2) We assume Theoretically assume that the terms are statistically independent which is not always ~~true~~ true in practice. The order in which the term appears in document is lost in VSM representation.

(3) A few important factors are: (1) The keywords that are searched should precisely match the document term, else the substring words can result in false negative match. (2) The synonyms (ie. semantically sensitive words that have same context (meaning) but different spelling (terms) might also result in false negative as they don't match the vocabulary.

(3) A few important factors are: (1) The keywords that are searched should precisely match the document term, else the substring words can result in false negative match. (2) The synonyms (ie. semantically sensitive words that have same context (meaning) but different spelling (terms) might also result in false negative as they don't match the vocabulary.

(4) In VSM, the features are spases. It can not deal with lexical (vocabulary) ambiguity therefore if we want to work on domain specific words, similarity would be close to null resulting in meaningless data.

(5) ~~Since we assume that terms are independent of each other & ~~ we don't need to estimate as we are only ranking documents. Naive Bayes Independence assumption makes it easier that absence & presence of one word is independent to another. & since each term is either present or not therefore & we assume & it is likely to happen that there's an equal chance.

QUESTION 1 (part b)

$d_1 = \langle 0.04, 0.04, 0, 0.06, 0, 0, 0 \rangle$

$d_2 = \langle 0, 0.11, 0.16, 0, 0, 0.09, 0 \rangle$

$d_3 = \langle 0.01, 0.07, 0, 0, 0, 0, 0.03 \rangle$

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

Time complexity $= O(n^2)$

$$\cos(d_1, q) = \frac{d_1 \cdot q}{|d_1||q|} = \frac{0.06}{0.08 \times 0.73} = 0.43$$

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

$$\cos(d_3, q) = \frac{0.03}{0.09 \times 1.73} = 0.24$$

rank : d_2, d_1, d_3.

Name – Ayesha Ausaf
Roll no – 17K-3867
Section – GR2
course – Information Retrieval Date

QUESTION 2

Part (a)

An IR system can give 100% precision when a single related doc is retrieved & it can also give 100% recall when all docs are returned for some query. Whereas, when used together, they can alter the values.

Part (b)

Breakeven point is used to defined equality relationship holding. It is evaluate retrieval system that returns a list of docs ordered by their supposed relevance need.

Precision ~= recall iff fp = fn or tp = 0. If the highest ranked element is not relevant then tp = 0 and that is trivial breakeven point. If highest ranked element is relevant then the no. of false +ve increases as you go down the list & the no. of false -ve decreases. As we go down the list, if item is R, then fn decreases & fp doesnot change. If item is N, then fp increase & fn doesnot change. At the beginning of list fp < fn & at the end fp > fn; Thus, there has to be a break even point It is related to f1 because it can be used to evaluate classification f model

Name – Ayesha Ausaf
Roll no – 17k-3867
Section – GR2
Course – Information Retrieval Date ____ 20 ____

QUESTION 2

Part (c)

(a) Precision = 5/12 = 0.416

(b) Recall = 5/8 = 0.625

$$F_1 = \frac{2PR}{P+R} = 0.499$$

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

= 0.692

(d) MAP $_{\text{largest}}$ = $\frac{1}{8} \times \left(1 + 1 + \frac{3}{5} + \frac{4}{9} + \frac{5}{12} + \frac{6}{13} + \frac{7}{14} + \frac{8}{15}\right)$

= 0.61

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

= 0.43

Name – Ayesha Ausaf
Roll no – 17K-3867
Section – GR2
Course – Information Retrieval

20

QUESTION 3

Part

(i) KNN — k=3.

Dictionary = {Japan, Macao, Osaka, Sapporo, Shangai, Taipei, Taiwan}

$d_1 = <0,0,0,0,0,1,1>$
$d_2 = <0,1,0,0,1,0,1>$
$d_3 = <1,0,0,1,0,0,0>$
$d_4 = <0,0,1,1,0,0,1>$
$d_5 = <0,0,0,1,0,0,2>$

$\cos(d1, d5) = (d1 \cdot d5) \div |d1| \times |d5|$
$= (0+0+0+0+0+0+2) \div \sqrt{|1+1|} \times \sqrt{|1+4|}$
$= 2/\sqrt{2} * \sqrt{5}$

$= ~~0.547~~ \; 0.632$

$\cos(d2, d5) = (d2 \cdot d5) \div |d2| \times |d5|$
$= 2 \div ((\sqrt{3})(\sqrt{5}))$
$= 0.516$

$\cos(d3, d5) = 1 \div \sqrt{2}\sqrt{5}$
$= 0.316$

$\cos(d4, d5) = (1+2) \div \sqrt{3}\sqrt{5}$
$= 0.77$

for 3-NN : 0.77, 0.632, 0.516

since d1, d2 belong to yes ∴

d5 belongs to yes

Page No.

Bright

Teacher's Signature

Name — Ayesha Ausaf
Roll no — 17K-3867
section — GR2
course — Information Retrieval

part (iii)

$$\hat{P}(c) = \frac{N_c}{N} \quad ; \quad \hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$$

① Prior Probability

$P(yes) = 2/4 = 1/2$

$P(no) = 2/4 = 1/2$

② conditional probability

$P(Japan\|yes) = 0+1/5+7$	$= 1/12$	$P(Japan\|no) = 0+1/5+7$	$= 1/6$
$P(Macoa\|yes) = 1+1/5+7$	$= 1/6$	$P(Macao\|no) = 0+1/5+7$	$= 1/12$
$P(Osaka\|yes) = 0+1/5+7$	$= 1/12$	$P(Osaka\|no) = 1+1/5+7$	$= 1/6$
$P(Sapporo\|yes) = 0+1/5+7$	$= 1/12$	$P(Sapporo\|no) = 2+1/5+7$	$= 1/4$
$P(Shangai\|yes) = 1+1/5+7$	$= 1/6$	$P(Shangai\|no) = 0+1/5+7$	$= 1/12$
$P(Taipei\|yes) = 1+1/5+7$	$= 1/6$	$P(Taipei\|no) = 0+1/5+7$	$= 1/12$
$P(Taiwan\|yes) = 2+1/5+7$	$= 1/4$	$P(Taiwan\|no) = 1+1/5+7$	$= 1/6$

③ Choosing class

$$P(yes|d5) = 1/4 \times 1/4 \times 1/12 \times 1/2$$
$$= 2.6 \times 10^{-3}$$

$$P(no|d5) = 1/6 \times 1/6 \times 1/4 \times 1/2$$
$$= 3.47 \times 10^{-3}$$

since 3.47×10^{-3} is greater, Thus doc 5 belongs to class **no**

Roll no— 17L-3867
Section— GR2
Course— Information Retrieval

Date _____ 20 _____

QUESTION 4 (part a)

let $c_1 = d2$ & $c_2 = d5$.

c_1 & c_2 are initial clusters. Deciding to which cluster each doc belongs to.

For d1 :

Distance $(c_1, d1) <$ Distance $(c_2, d1)$

∴ d1 belongs to C1

For d3 :

Dist $(c_1, d3) <$ Dist $(c_2, d3)$

∴ d3 belongs to C1

For d4 : Dist $(c_1, d4) >$ Dist $(c_2, d4)$

∴ d4 belongs to C2

For d6 : Dist $(c_1, d6) >$ Dist $(c_2, d6)$

∴ d6 belongs to C2

∴ d1, d2 and d3 are in C1. d4, d5, d6 are in C2.

K mean converge to local maximum, in order to find optimal clustering one need to produce all possible clustering arrangements

Name— Ayesha Ausaf
Roll no— 17K-3867
Section— GR2
Course— Information Retrieval Date ____ 20 ____

QUESTION 5 (part a)

There are 4 type of queries
→ Informational query
→ Navigational query
→ Transactional query
→ Faceted query (extra)

Navigational → some user might put youtube or instagram in search bar rather than adding a URL.

Informational → some user searching about IR or AI to learn about the domains of CS.

Transactional → some user might search to download some app, or know price of ~~some~~ iphone X.

Faceted → A query on drugs for prevention of osteoporosis would include
→ osteoporosis OR boneloss
→ prevention OR cure

QUESTION 5 (b)

(i) Politeness

Web servers have both implicit & explicit ~~policies~~ policies regulating the rate at which a crawler can visit them. These politeness policies must be respected.

~~(ii) Validity of URL/spectr~~

Challenge: URL frontier — containing URL yet to be fetched in current crawl

Solution: decrease speed by which it visits sites.

(ii) Freshness

In many app, crawler should operate in continuous mode. It should obtain fresh copies of previously fetched pgs.

Challenge: Refreshing page & bringing new page is a challenge.

Sol: don't overload servers & use network to maximum extent.

(iii) Extensible

Crawler should be designed to be extensible in many ways — to cope with new data format, new fetch protocol etc.

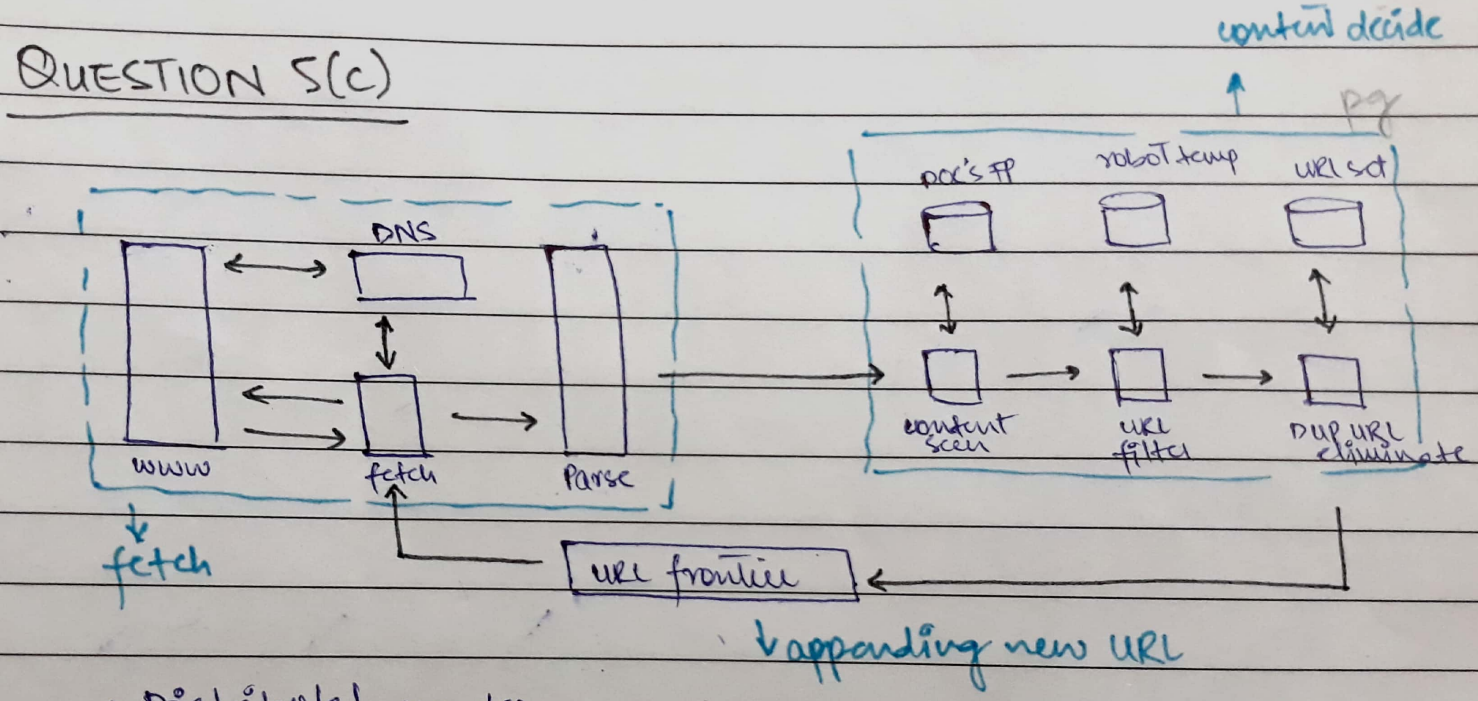
Crawler architect should be be modular

Challenge: Data with unknown format

Sol: Decode the format to extract data.

content decide

QUESTION 5(c)



→ Distributed crawler

Threads in a crawler can run under different process each at a node of a distributed crawling sys. It is essential for scaling & can also be good for geographically known where each node crawls hosts 'near it' Partitioning can be done by hash function.

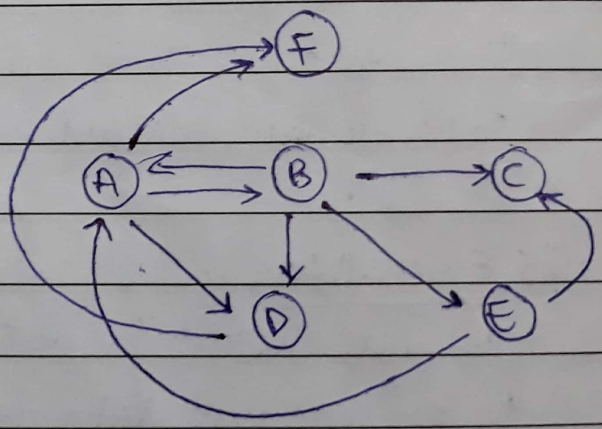
→ ~~Distributed crawler for bar~~

Name — Ayesha Ausaf
Roll no — 17k-3867
Section —
Course —

Date

QUESTION 6 (part a)

	A	B	C	D	E	F
A	0	1	0	1	0	1
B	1	0	1	1	1	0
C	0	0	0	0	0	0
D	0	0	0	0	0	1
E	1	0	1	0	0	0
F	0	0	0	0	0	0



Name – Ayesha Ausaf
Roll no – 17K-3867
Section –
Course –
Date

QUESTION 6 (part b)

$a^1 = A^T \cdot h^{(0)}$

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^t \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Pg

$$\begin{bmatrix} 2 \\ 1 \\ 2 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

$$= \sqrt{2^2 + 1^2 + 2^2 + 2^2 + 1^2 + 2^2}$$

$$= \sqrt{4 + 1 + 4 + 4 + 1 + 4}$$

$$= \sqrt{18}$$

$$= \begin{bmatrix} 0.471 \\ 0.236 \\ 0.471 \\ 0.471 \\ 0.236 \\ 0.471 \end{bmatrix}$$

Name — Ayesha Ansab
Roll no — 17k-3867
Section —
Course —
Date ____ 20

$h^{(0)} = A \cdot a^0$

$$\begin{bmatrix} 0&1&0&1&0&1 \\ 1&0&1&1&1&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&1 \\ 1&0&1&0&0&0 \\ 0&0&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0.471 \\ 0.236 \\ 0.471 \\ 0.471 \\ 0.236 \\ 0.471 \end{bmatrix}$$

$$\begin{bmatrix} 1.178 \\ 1.169 \\ 0 \\ 0.471 \\ 0.942 \\ 0 \end{bmatrix}$$

$$a^2 = \begin{bmatrix} 0&1&0&0&1&0 \\ 0&0&0&0&0&0 \\ 0&1&0&0&1&0 \\ 1&1&0&0&0&0 \\ 0&1&0&0&0&0 \\ 1&0&0&1&0&0 \end{bmatrix}$$

pg

$$= \begin{bmatrix} 0.47 \\ 0.23 \\ 0.47 \\ 0.55 \\ 0.31 \\ 0.31 \end{bmatrix} \quad h^{(1)} \Rightarrow \begin{bmatrix} 1.25 \\ 1.8 \\ 0 \\ 0.31 \\ 0.94 \\ 0 \end{bmatrix}$$