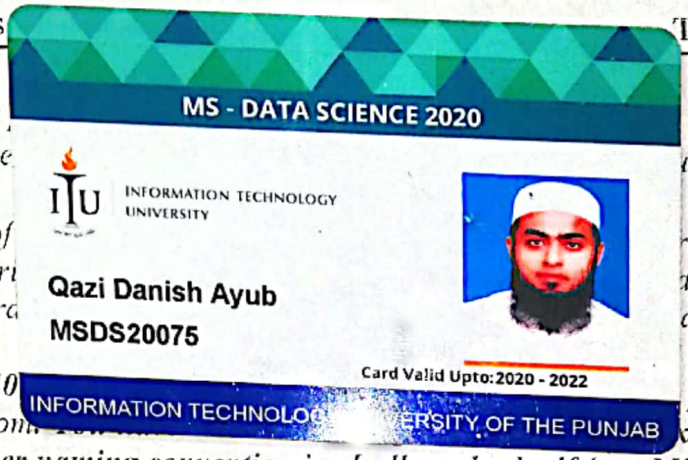
MS Data Science: Information Retrieval and Text Mining (Fall 2020)

Mid-Term Exam **Instructor:** Dr. Saeed Ul Hassan

Name Qazi Danish Ayub **Roll No.** MSDS 20075

Date 08-Dec-2020

Time Allowed: 2 Hours **Total Marks:** 60



- *Use plan A4 size*
- *Please follow the* *uestion number on the sheet.*
- *The first page of* *roll number. For that, students must wri* *dent ID card (or CNIC if student ID card* *ake sure all the details are visible*
- *An additional 40* *pdf file and submit on google classroom* *xam on the classroom* ***following proper naming convention i.e. [rollnumber].pdf (e.g. MSDS19016.pdf).***
- ***Please note no extra time will be given and wrong submissions will not be considered.***
- *Best of luck!*

Q No 1

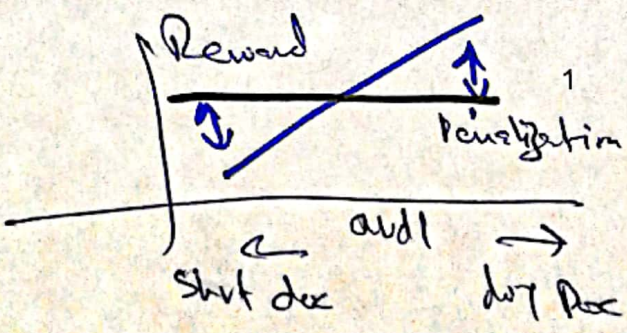
(4)

Long Doc — better chance to match any Query

Need avoid over Penalization

So Pivoted Length normalizer : take average doc length as "Pivot"

Norm = 1 if |d| = ~~pivot~~, (and 1)



be [0, 1]

Scanned with CamScanner

Q No 1

(a)

Answer:

Pseudo Relevance Feedback:

A method for automatic local analysis on manual part of relevance feedback such that the user gets improved retrieval performance without extra interaction. It is also known as blind relevance feedback. * A unsupervised technique to some extent.

Relevance feedback:

The involvement of User in the Process for improving the result such that the user give feedback and the System computes better representation of that feedback is called Relevance feedback.

* A Pure supervised technique.

Scanned with CamScanner

(2)

Ans:.

→ Stemming increases the total recall

Reason:.

According to the concept of stemming User entered term more documents are matched and as alternately words forms for a user entered terms are matched as well thus which produces the increase in recall - Although also reduces the precision -

(3)

Ans:.

Index Searcher:- Is a Gateway to Searching an index - All searches come through an Index Searcher instance using any of the several overloaded methods -

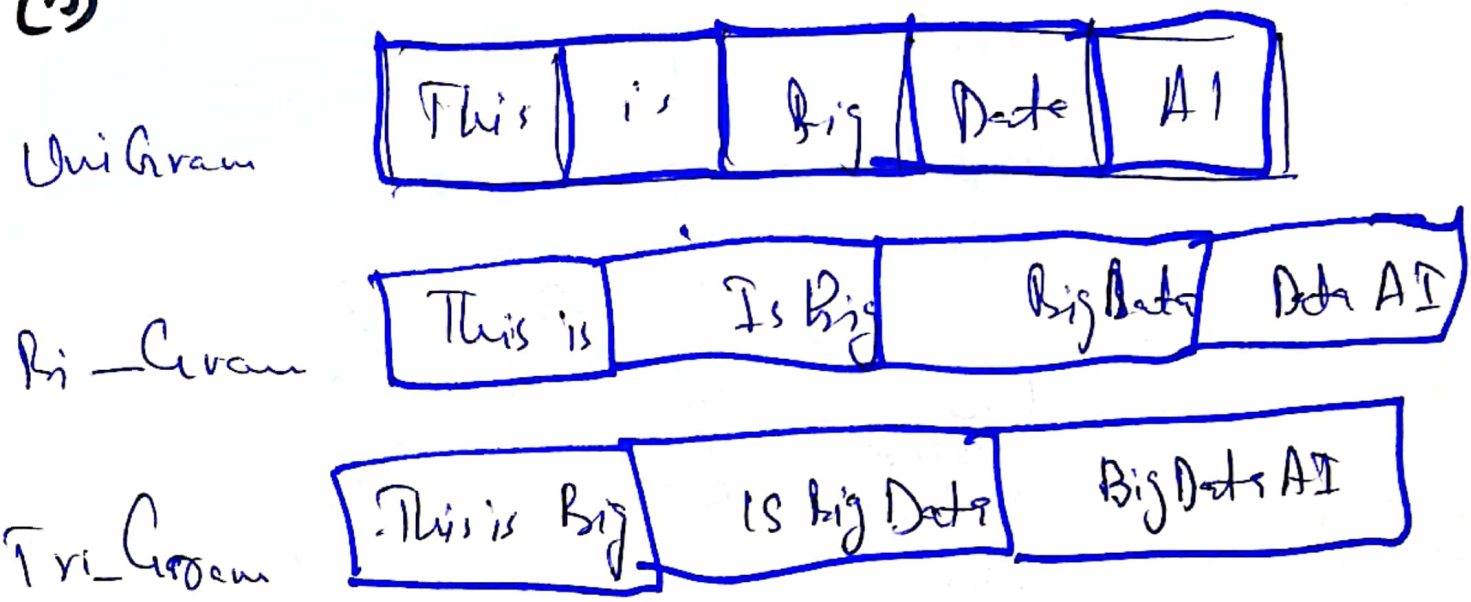
QueryParser:- Processes a human-entered (and readable) expression into a concrete Query Object.

Scanned with CamScanner

Q.N.1

(5)

This is big Data AI



Scanned with CamScanner

Q2: Long Questions (50 Marks)

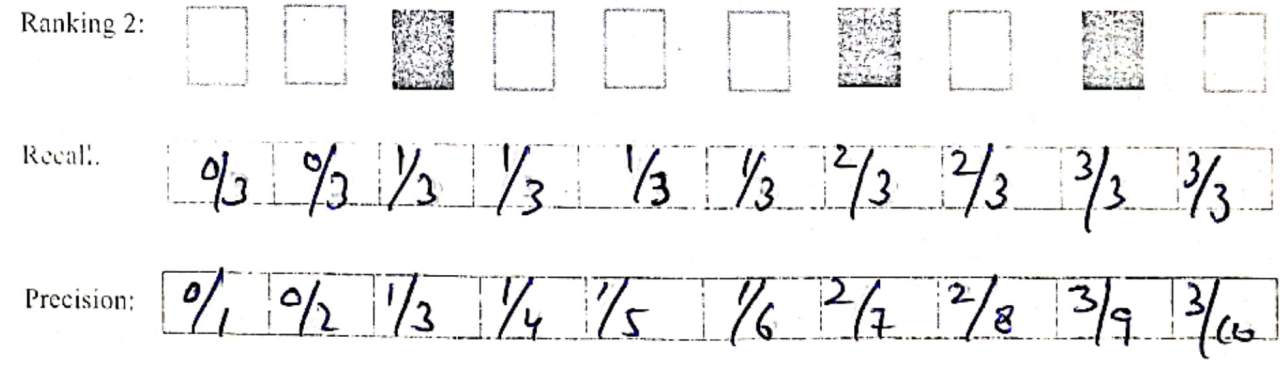
1. Provide answers for a, b, c and d parts (10 marks)

a) Fill the recall and precision values for the below given documents w.r.t relevant documents found for each query? Draw similar table on your paper for both Query 1 and Query 2 separately. (3)

= Relevant Documents for **Query 1**

Ranking 1:										
Recall:	1/5	1/5	1/5	2/5	3/5	3/5	4/5	4/5	4/5	5/5
Precision:	1/1	1/2	1/3	2/4	3/5	3/6	4/7	4/8	4/9	5/10

= Relevant Documents for Query 2



Ranking 2:										
Recall:	0/3	0/3	1/3	1/3	1/3	1/3	2/3	2/3	3/3	3/3
Precision:	0/1	0/2	1/3	1/4	1/5	1/6	2/7	2/8	3/9	3/10

b) Find the Average Precision of Query 1 (2.5)

c) Find the Average Precision for Query 2 (2.5)

(b) $\frac{\frac{1}{1} + \frac{2}{4} + \frac{3}{5} + \frac{4}{7} + \frac{5}{10}}{5} = \frac{3.17}{5} = 0.6342$

(c) $\frac{\frac{1}{3} + \frac{2}{7} + \frac{3}{9}}{3} = \frac{0.952}{3} = 0.317$

(d) $MAP = \frac{0.6342 + 0.317}{2} = 0.4756$

Scanned with CamScanner

Lay

Q No 2

(2) (a)

Boolean Model:

Deals with the the Document selection and thus Does not contribute in Ranking as much [0,1]

Vector Space Model:

Deals with Term Frequency Inverse Document Frequency and thus also contribute is Document Ranking also values is not boolean.

(b) Pseudo Code for removing stopword punctuation and case sensitiveness

1. Dictionary – Stopwords
2. Dictionary.lower()
3. Dictionary.removePunctuation()

Scanned with CamScanner

	TF_D1	TF_D2	TF_D3	doc_freq	IDF
representation	0	0	1	1	2.0
important	1	1	0	2	1.0
realizing	1	0	0	1	2.0
sub	0	0	1	1	2.0
semantic	1	1	0	2	1.0
knowledge	0	0	1	1	2.0
owl	0	1	2	2	1.0
Language	0	0	1	1	2.0
Idf	0	1	0	1	2.0
technologies	1	1	0	2	1.0
several	1	0	1	2	1.0
Language	0	0	1	1	2.0
web	1	1	0	2	1.0
amongst	0	1	0	1	2.0

Scanned with CamScanner

(c)

TF-IDF_1	TF-IDF_2	TF-IDF_3
0	0	2.0
1.0	1.0	0
2.0	0	0
0	0	2
1.0	1.0	0
0.0	0	2
0.0	1.0	2
0.0	2.0	2
0.0	1.0	0
1.0	0	0
1.0		1.0
0.0	0	2.0
1.0	1.0	0
0.0	2.0	0

D_1

$D_3 \longrightarrow 2.0$

$D_2 \longrightarrow 1.00$

$D_1 \longrightarrow 0$

D_2

$D_3 \longrightarrow 2.0$

$D_1 \longrightarrow 0$

$D_2 \longrightarrow 0$

Scanned with CamScanner

Long
QNo3

$$V_{iu} = \alpha \vec{v} + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} d_j - \frac{\gamma}{|D_r|} \frac{\sum d_j}{\forall d_j \in D_N}$$

Calculation Term Frequency

$$TF = \frac{\text{no of rep of word}}{\text{no of wor in doc}}$$

D_1 { 2, 1, 1, 1}

the dog attached girl

D_2 = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1}

Using Form

+ D_1 { 0.4, 0.2, 0.2, 0.2}

- D_2 { 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1}

$\alpha = \gamma = 1$, $\beta = 0.5$

{1 + 0.2 × 0.5 − 0.1, 1 * 0.2 × 0.5 − 0.1}

{1, 1} Ans.

Scanned with CamScanner

Q No 7.

Range:

5 ↑ Relevant

4

3

2 ↓ Non-R

1

$$Q_M = (a \cdot \vec{Q_0}) + b \cdot \frac{1}{|D_r|} \sum_{\vec{D_j} \in D_r} \vec{D_j} - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\vec{D_k} \in D_{nr}} \vec{D_k} \right)$$

1 = Non-Relevant $\vec{P_k}$

5 = Relevant $\vec{P_j}$

Scanned with CamScanner