# Information Retrieval

## Mid Semester Examination - Online Mode (Set 1)

(For Students having Roll Nos. ending with EVEN numbers)

**Date: 24 Sept 2020** **Time: 09:00 AM - 10:00 AM**

**Instructions:**

a) Read all questions carefully and answer them in A4 Sheets.
b) Answer all questions (no choice, unless otherwise mentioned) and avoid unnecessary/ trivial explanations.
c) Calculators / Electronic gadgets are NOT permitted during the examination.
d) Most importantly, NO answer should be written in Pencil. Final answers should be written using either a BALLPOINT pen or INK pen.
e) On the top right-hand corner of every A4 sheet (Answer Sheet), write your Roll No, Name, and keep the page number encircled.
f) Sheets having missing student details would not be evaluated.
g) At the end of the examination, you will get a link to upload the scanned copy of the answer script in a single PDF format.
h) Late Submissions will not be accepted under any circumstances.
i) Most importantly, this is a proctored examination, and students are advised to keep their videos on during the examinations.
j) Students must use the meet link sent to them and they should not use the meet link sent to others.
k) The duration of the descriptive exam will be 60 minutes.

**Descriptive Questions:**

1) State the tasks involved in preprocessing text data with suitable examples and briefly describe the necessity of these tasks in building a scalable Information Retrieval system

2) Create a positional index of the following documents:
d1 = “shipment of platinum damaged in Delhi, India”
d2 = “delivery of a parcel for Delhi arrived in a new truck”
d3 = “shipment of my new parcel arrived in a truck”
d4 = “platinum and parcels are the shipments today from new Delhi”

3) How are the enumerated terms used in wild-card queries? Compare it with the permuterm index and justify the best approach in terms of the computational cost (time and space)

4) Explain Bi-Word indexing with an example. Describe at least 4 advantages and disadvantages of this approach in detail.

5) Assume that the collection has 25,431 terms and the following terms are given:

| Term | Term Frequency | X\|W | P(X\|W) |
|---|---|---|---|
| actress | 321 | c\|ct | 0.000117 |

| | | | |
|---|---|---|---|
| cress | 80 | a\|# | 0.000144 |
| caress | 121 | ac\|ca | 0.000164 |
| access | 527 | r\|c | 0.000209 |
| across | 803 | e\|o | 0.00093 |
| acres | 412 | es\|e | 0.00321 |
| acres | 412 | ss\|s | 0.00342 |

Now user types “acress” wrongly while searching for information. Now perform noisy channel modeling for independent word spelling correction

6) State and describe 5 term weighting methods each with an example