CMI/Aug-Nov 2020/Assignment 1 Aug 2020

Chennai Mathematical Institute

Information Retrieval Deadline: Sep 07, 2020 11:59 PM IST. Max Marks: 10.

Roll No.: _____

Name: _____

Question 1 [10 Marks]: Take any three of your favorite movie names from any source. Each movie name should at least have three words of three or more characters in it. There must be **at least** one **common** word between at least any two of these movie names after *case folding* and *stop word* removal.

Assuming that these movie names form documents in your collection, answer the following questions:

(1) Apply case-folding to each movie name.
(2) To the resulting movie name after case folding, apply stop-word removal.
(3) Use the resulting movie names after case-folding and stop-word removal to draw the positional inverted index.
(4) For the same movie names after case-folding and stop-word removal, draw a 2-gram non-positional inverted index.
(5) Find a query which will result in a false-positive when fired on your 2-gram non-positional index. Explain the reason for our simple retrieval system (as discussed in the class) to result in a false-positive for this query on your 2-gram non-positional index.

Answer:

Movie Name 1:
After Case Folding:
After Stop-word Removal:

Movie Name 2:
After Case Folding:
After Stop-word Removal:

Movie Name 3:
After Case Folding:
After Stop-word Removal:

Common Word(s) that occurred in movie names before case-folding and stop-word removal:

Positional Inverted Index:

2-gram Non-positional Inverted Index:

Query:

Intent:

Why does this query result in a false-positive on a 2-gram non-positional inverted index?