Chennai Mathematical Institute

Information Retrieval Deadline: Aug 23, 2019. Max Marks: 10.

Roll No.: _____

Name: _____

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

Assuming that these movie names form documents in your collection, draw the corresponding 3-gram non-positional inverted index derived from that collection.

Also draw the permuterm non-positional inverted index for the same collection.

Compare these two indexes using couple of wild-card queries. Discuss which index is better in the context of wild-card queries and why.

Answer:

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

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

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

Common Word(s):

3-gram Non-positional Inverted Index:

Permuterm Non-positional Inverted Index:

Wild-Card Query 1:
Wild-Card Query 2:

Using these two queries as example, explain what happens during wild-card query processing? Which index is better for wild-card query processing and why?