Group members (Name, Sjsu ID) :

HW 3 : EXPERIMENTS

URLS from homework 1

We have the following three URLs :-

1. https://www.yahoo.com/
2. https://www.imdb.com/
3. https://en.wikipedia.org/wiki/Canada

Queries from homework

Topic 1 – Yahoo Landing Page

1. New developments in AI industry
2. Recent activities about climate change.
3. Actors who got MTV VMA awards

Topic 2 - Internet Movie Database Landing Page

4. New releases this month
5. Top most watched favourite movies
6. Latest movies released by Marvel

Topic 3 - Wikipedia Canada Page

7. Location of Canada in America
8. Total population of Canada
9. Neighbouring countries of Canada

We created stemmed, char-gramed and none(neither stemmed nor char-graming) corpus of all these three URLs and computed TF-IDF score using disjunctive cosine ranking.

For None(neither stemmed nor char-grammed) corpus:-

Score of Query 1 : New developments in AI industry ("_OR _OR New Developments AI")

DocId Score

2	0.0
1	0.0

Score of Query 2: Recent activities about climate change(“_OR _OR recent activity climate”)

DocId	Score
-	-

Score of Query 3: Actors who got MTV VMA awards (“_OR _OR actors MTV VMA”)

DocId	Score
-	-

Score of Query 4: New releases this month (“_OR _OR new releases month”)

DocId	Score
2	0.0
1	0.0

Score of Query 5: Top most watched favourite movies (“_OR _OR most watched movies”)

DocId	Score
2	0.19

Score of Query 6: Latest movies released by Marvel (“_OR _OR latest released marvel”)

DocId	Score
2	0.11

Score of Query 7: Location of Canada in America (“_OR _OR location Canada America”)

DocId	Score
3	0.13

Score of Query 8: Total population of Canada (“_OR _OR total population Canada”)

DocId	Score
3	0.18

Score of Query 9: Neighbouring countries of Canada (“_OR _OR neighbour countries Canada")

DocId	Score
3	0.14

For Stemmed corpus:-

Score of Query 1 : New developments in AI industry (“_OR _OR New Develop AI”)

DocId	Score
2	0.04
1	0.0

Score of Query 2: Recent activities about climate change (“_OR _OR recent activ climat”)

DocId	Score
-	-

Score of Query 3: Actors who got MTV VMA awards (“_OR _OR actor MTV VMA”)

DocId	Score
2	0.04

Score of Query 4: New releases this month (“_OR _OR new releas month”)

DocId	Score
2	0.04
1	0.0

Score of Query 5: Top most watched favourite movies (“_OR _OR most watch movi”)

DocId	Score
2	0.23

Score of Query 6: Latest movies released by Marvel (“_OR _OR latest releas marvel”)

DocId	Score
2	0.11

Score of Query 7: Location of Canada in America (“_OR _OR locat Canada America”)

DocId	Score
3	0.14

Score of Query 8: Total population of Canada (“_OR _OR total popul Canada”)

DocId	Score
3	0.19

Score of Query 9: Neighbouring countries of Canada (“_OR _OR neighbour countri Canada ")

DocId	Score
3	0.14

For Char-grammed corpus:-

Score of Query 1 : New developments in AI industry

DocId	Score
2	0.09
3	0.0
1	0.0

Score of Query 2: Recent activities about climate change

DocId	Score
2	0.08
3	0.0

Score of Query 3: Actors who got MTV VMA awards

DocId	Score
2	0.03

Score of Query 4: New releases this month

DocId	Score
2	0.05
1	0.0

Score of Query 5: Top most watched favourite movies

DocId	Score
2	0.23

Score of Query 6: Latest movies released by Marvel

DocId	Score
2	0.12
1	0.01
3	0.0

Score of Query 7: Location of Canada in America

DocId	Score
3	0.26
2	0.01

1 0.0

Score of Query 8: Total population of Canada

DocId Score

3 0.28

2 0.02

1 0.0

Score of Query 9: Neighbouring countries of Canada

DocId Score

3 0.16

1 0.02

2 0.02

Relevant Documents:

Topic 1 = For query 1 to 3: Doc 1

Topic 2 = For query 4 to 6: Doc 2

Topic 3 = For query 7 to 9: Doc 3

Evaluating Precision for queries :-

1. None (Neither stemmed nor char-grammed) scenario

   Query 1: 1/2
   Query 2: --
   Query 3: --
   Query 4: 1/2
   Query 5: 1
   Query 6: 1
   Query 7: 1
   Query 8: 1
   Query 9: 1

2. Stemmed scenario

   Query 1: 1/2

Query 2: --
Query 3: 0
Query 4: 1/2
Query 5: 1
Query 6: 1
Query 7: 1
Query 8: 1
Query 9: 1

3. Char-grammed scenario

Query 1: 1/3
Query 2: 0
Query 3: 0
Query 4: 1/2
Query 5: 1
Query 6: 1/3
Query 7: 1/3
Query 8: 1/3
Query 9: 1/3

Average Precision for queries:-

1. None (Neither stemmed nor char-grammed) scenario

Query 1: 1/1*(1/2) = 1/2
Query 2: --
Query 3: --
Query 4: 1/1*(1/1) = 1/1
Query 5: 1/1*(1/1) = 1/1
Query 6: 1/1*(1/1) = 1/1
Query 7: 1/1*(1/1) = 1/1
Query 8: 1/1*(1/1) = 1/1
Query 9: 1/1*(1/1) = 1/1

Average Precision for topic 1: 0.16

Average Precision for topic 2: 1

Average Precision for topic 3: 1

2. Stemmed scenario

Query 1: 1/1*(1/2) = 1/2
Query 2: --

Query 3: 1/1*(0/1) = 0
Query 4: 1/1*(1/1) = 1/1
Query 5: 1/1*(1/1) = 1/1
Query 6: 1/1*(1/1) = 1/1
Query 7: 1/1*(1/1) = 1/1
Query 8: 1/1*(1/1) = 1/1
Query 9: 1/1*(1/1) = 1/1

Average Precision for topic 1: 0.16

Average Precision for topic 2: 1

Average Precision for topic 3: 1

3. Char-grammed scenario

   Query 1: 1/1*(1/3) = 1/3
   Query 2: 0
   Query 3: 0
   Query 4: 1/1*(1/1) = 1/1
   Query 5: 1/1*(1/1) = 1/1
   Query 6: 1/1*(1/1) = 1/1
   Query 7: 1/1*(1/1) = 1/1
   Query 8: 1/1*(1/1) = 1/1
   Query 9: 1/1*(1/1) = 1/1

   Average Precision for topic 1: 0.11

   Average Precision for topic 2: 1

   Average Precision for topic 3: 1

Evaluating MAP score for each three corpus:

1. None (Neither stemmed nor char-grammed) corpus:

   MAP = (0.16 + 1 + 1)/3 = 0.72

2. Stemmed corpus:

   MAP = (0.16 + 1 + 1)/3 = 0.72

3. Char-grammed corpus:

MAP = (0.11 + 1 + 1)/3 = 0.703

Conclusion:

From the above MAP results, char-gramming score is lower than stemming. This is because char-gramming fetches more non-relevant documents than stemming. So, stemming seems to be more effective in comparative to char-gramming, as it fetches less non-relevant documents to a query. Though stemming fetches more non-relevant documents as comparative to none case scenario.

Comparing inverted indexes:

On running program on these three case scenarios, we found the following results:-

Stemming reduces the inverted index as comparative to normal (none case) inverted index because it reduces the term to its root terms. For e.g., both terms ‘example’ and ‘examples’ refer to term ‘exampl’ in the stemmed inverted index.

Whereas char-gramming increases the size of inverted index as comparative to both normal and stemmed inverted index. Since in char-gramming we divide every character into small characters of length four each, thus increasing the size of index. For e.g., term ‘example’ converts to ‘exam xamp ampl mple’. This also increase the number of query terms.