

REPORT

IMPLEMENTATION

- 1) Library used nltk, re, csv.
- 2) Text cleaning and punctuation removal is done with re and the tokenization is done with nltk.
- 3) In part 1 an inverted index is built from the corpus in the form of a dictionary. And the doc id along with its respective title is stored in another array.
- 4) The processed corpus is converted to Inverted index and stored in a csv file which is then exported using csv library.
- 5) The doc ids and titles are exported to a text file.
- 6) The csv file is then imported and used for building the vector space model where then the inputted query is processed from the built vector space model, giving out results.
- 7) The same processing file which constructs the vector space model is again used to implement the issues correction we have noticed within the code of the implementation

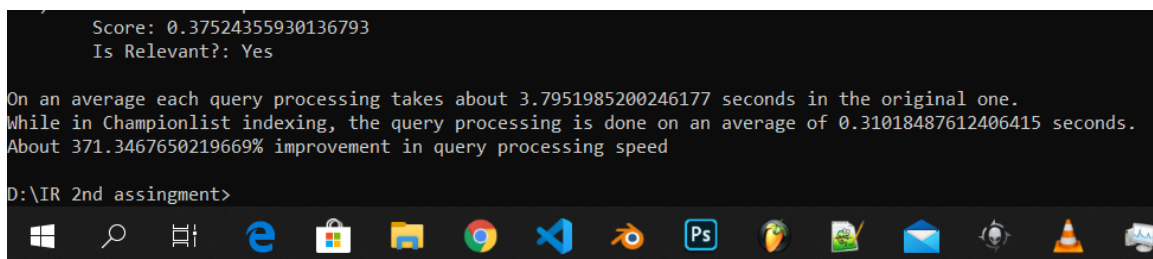
IMPROVEMENTS IN PART 1

ISSUE 1

- 1) HIGH LATENCY - Score computation is a large (10s of %) fraction of the, CPU work on a query, Generally, we have a tight budget on latency We'll look at ways of cutting CPU usage for scoring, without compromising the quality of results (much) .
- 2) CHAMPION LIST-For every term (t), store a list of r documents that have the highest score for term t, The score we are using weight score.r is fixed at

the index creation time, thus it's possible that $r < K$. The set of r documents are called the champion list for term t . Now, for a query, create a set of documents A from the champion list of all the terms in the query.

- 3) Decrease in the number of documents will result in low latency.
- 4) This is an inexact way of finding top r documents, corner cases might exist. example- weights are almost equal in more than r elements.
- 5) With the same queries in the original and modified ones, we have calculated their processing speed and checked out its improvement ratio from its predecessor.



```
Score: 0.37524355930136793
Is Relevant?: Yes

On an average each query processing takes about 3.7951985200246177 seconds in the original one.
While in Championlist indexing, the query processing is done on an average of 0.31018487612406415 seconds.
About 371.3467650219669% improvement in query processing speed

D:\IR 2nd assingment>
```

It had a 371.3 % improvement in processing speed

ISSUE 2

- 1) High frequency of some words such as stop words will increase the score of some documents ,since stop words are not the prime indicators of relevance of a document, the weight of stop words should be lesser than rare words.
- 2) In query, calculate IDF of every term, store the maximum IDF value, ignore the terms with IDF value lesser than half of the maximum value of IDF .
- 3) Less frequent words are given more importance than more frequent words, it will increase the probability of extracting more relevant documents .
- 4) Ignoring some words might cause some inconsistency in search.
- 5)

```

Query: Acalolepta Species
1.) Title: Acalolepta affinis
   Score: 0.37524355930136793
   Is Relevant?: Yes
2.) Title: Acalolepta andamanica
   Score: 0.37524355930136793
   Is Relevant?: Yes
3.) Title: Acalolepta arrowi
   Score: 0.37524355930136793
   Is Relevant?: Yes
4.) Title: Acalolepta basioplgiata
   Score: 0.37524355930136793
   Is Relevant?: Yes
5.) Title: Acalolepta bicolor
   Score: 0.37524355930136793
   Is Relevant?: Yes
6.) Title: Acalolepta bispinosa
   Score: 0.37524355930136793
   Is Relevant?: Yes
7.) Title: Acalolepta blairi
   Score: 0.37524355930136793
   Is Relevant?: Yes
8.) Title: Acalolepta borneensis
   Score: 0.37524355930136793
   Is Relevant?: Yes
9.) Title: Acalolepta buruensis
   Score: 0.37524355930136793
   Is Relevant?: Yes
10.) Title: Acalolepta celebensis
   Score: 0.37524355930136793
   Is Relevant?: Yes

```

```

C:\WINDOWS\System32\cmd.exe
D:\IR 2nd assingment>python high_idf.py
Query: ministry of zambia
1.) Title: Ministry of Commerce, Trade and Industry
   Score: 0.19973533396113868
   Is Relevant?: Yes
2.) Title: National Sports Council Malaysia
   Score: 0.08497779457594039
   Is Relevant?: No
3.) Title: District Development Committee
   Score: 0.04887611714171734
   Is Relevant?: No
4.) Title: Amirul Islam Chowdhury
   Score: 0.044988244187262556
   Is Relevant?: No
5.) Title: Rumen Radev
   Score: 0.044988244187262556
   Is Relevant?: No
6.) Title: Stéphane Bahier
   Score: 0.04110037123280776
   Is Relevant?: No
7.) Title: Mohamed Gueddiche
   Score: 0.03554626701215807
   Is Relevant?: No
8.) Title: Petre V. Haneş
   Score: 0.03443544616802813
   Is Relevant?: No
9.) Title: National Political Union (England)
   Score: 0.033880035745963155
   Is Relevant?: No
10.) Title: Medical Institute of North Caucasian Huma
   Score: 0.03165839405770328
   Is Relevant?: No

```

```

C:\WINDOWS\System32\cmd.exe
Score: 0.03165839405770328
Is Relevant?: No
Query: Lowpel and Geowaltek Nigeria Limited founder
1.) Title: Ibibia Walter
   Score: 0.173378429604093
   Is Relevant?: Yes
2.) Title: Unoneme Charles Chukwaku
   Score: 0.055704364441947055
   Is Relevant?: No
3.) Title: Dell Client Solutions Group
   Score: 0.04803516567843683
   Is Relevant?: No
4.) Title: Sizwe Nxasana
   Score: 0.0472573308243361
   Is Relevant?: No
5.) Title: Graham Bleathman
   Score: 0.04692685997941067
   Is Relevant?: No
6.) Title: Saif Sporting Club
   Score: 0.043291680685230975
   Is Relevant?: No
7.) Title: Leo's Room
   Score: 0.03569085125194615
   Is Relevant?: No
8.) Title: Byju Raveendran
   Score: 0.033281507648631234
   Is Relevant?: No
9.) Title: Mount Jagged, South Australia
   Score: 0.03139473026791559
   Is Relevant?: No
10.) Title: A Travel Duet
   Score: 0.028134882754513
   Is Relevant?: No

```

Sample queries.

ISSUE 3

- 1) The IR system in part 1 has no way to check whether the entered term of the query is correct in spelling. This results in unexpected consequences in the final output.
- 2) A spell corrector can be implemented using many dependable open source libraries. Each term of the query undergoes correction if the term is not available in the dictionary.
- 3) The imported libraries correct the spelling of the term and undergoes the query processing to correct results.
- 4) It does not always yield the expected correct term of its respective incorrectly entered term. This might cause the IR system to give an overall different result to the query.

5)

Query: Person behind wildliffe protaction act		Query: extaction of marinee reserrve	
1.)	Title: Virgin complex Score: 0.12697019152398964 Is Relevant?: No	1.)	Title: List of Governors of Paraná Score: 0.429 Is Relevant?: No
2.)	Title: Abdus Salam Pintu Score: 0.06819241425480319 Is Relevant?: No	2.)	Title: List of Albanian swimmers Score: 0.416 Is Relevant?: No
3.)	Title: City of London Electric Lighting Company Limited Score: 0.054078472468700955 Is Relevant?: No	3.)	Title: List of lighthouses in Montenegro Score: 0.384 Is Relevant?: No
4.)	Title: Abraham Kibiwott Score: 0.05303854442040248 Is Relevant?: No	4.)	Title: List of lighthouses in Romania Score: 0.384 Is Relevant?: No
5.)	Title: Artiifact Score: 0.05075632864155649 Is Relevant?: No	5.)	Title: Zachary Mills Score: 0.368 Is Relevant?: No
6.)	Title: List of Mayors of Pakistan Score: 0.05075632864155649 Is Relevant?: No	6.)	Title: Timeline of Augusta, Georgia Score: 0.365 Is Relevant?: No
7.)	Title: Mayor of Abbottabad Score: 0.04500045632137998 Is Relevant?: No	7.)	Title: List of Sigma Delta Tau Chapters Score: 0.362 Is Relevant?: No
8.)	Title: El Culebro: La historia de mi papá Score: 0.043946222519762046 Is Relevant?: No	8.)	Title: Eselsbach Score: 0.339 Is Relevant?: No
9.)	Title: Andrew Small Score: 0.040915448552881914 Is Relevant?: No	9.)	Title: Forellenbach Score: 0.339 Is Relevant?: No
10.)	Title: Alfonso Angelini Score: 0.03940006156944184 Is Relevant?: No	10.)	Title: Holovanivsk Score: 0.337 Is Relevant?: No

```

Query: psinteer famous for landscpse
1.) Title: Artie Green
Score: 0.09102955157056383
Is Relevant?: Yes
2.) Title: La segunda muerte
Score: 0.08638212356104466
Is Relevant?: Yes
3.) Title: Manhattan Vocational and Technical High School
Score: 0.08366694070246615
Is Relevant?: No
4.) Title: Elia Locardi
Score: 0.08111926732895373
Is Relevant?: No
5.) Title: Mathys Schoevaerdt
Score: 0.0713466897256598
Is Relevant?: No
6.) Title: Ahmed Omar Bin Fareed
Score: 0.05714815231804525
Is Relevant?: No
7.) Title: Caribbeing
Score: 0.05196585474712666
Is Relevant?: No
8.) Title: Colchicum brachyphyllum
Score: 0.05119917266867332
Is Relevant?: No
9.) Title: Where were you / Kga mi or
Score: 0.05106848744058014
Is Relevant?: No
10.) Title: Eugene Coste
Score: 0.04870238629292792
Is Relevant?: No

```

```

Query: Person behind wildliffe protactionm act
1.) Title: MK Ranjitsinh Jhala
Score: 0.12673457603493538
Is Relevant?: Yes
2.) Title: Washington State Department of Fish and Wildlife
Score: 0.10934298560205222
Is Relevant?: No
3.) Title: Virgin complex
Score: 0.09137245565758226
Is Relevant?: No
4.) Title: BC SPCA
Score: 0.07751434370748549
Is Relevant?: No
5.) Title: City of London Electric Lighting Company Limited
Score: 0.05243681674356045
Is Relevant?: No
6.) Title: Abdus Salam Pintu
Score: 0.04907378868136323
Is Relevant?: No
7.) Title: Mount Jagged, South Australia
Score: 0.042813154954715826
Is Relevant?: No
8.) Title: Abraham Kibiwott
Score: 0.038168502307726955
Is Relevant?: No
9.) Title: Artiifact
Score: 0.03652613524857166
Is Relevant?: No
10.) Title: List of Mayors of Pakistan
Score: 0.03652613524857166
Is Relevant?: No

```

```

Query: psinteer famous for landscpse
1.) Title: Artie Green
Score: 0.14099107023321444
Is Relevant?: No
2.) Title: Manhattan Vocational and Technical High School
Score: 0.12958749449221887
Is Relevant?: No
3.) Title: Ahmed Omar Bin Fareed
Score: 0.08851388387787537
Is Relevant?: No
4.) Title: Colchicum brachyphyllum
Score: 0.0792998100624026
Is Relevant?: No
5.) Title: Where were you / Kga mi or
Score: 0.07909739831968138
Is Relevant?: No
6.) Title: Eugene Coste
Score: 0.0754326638754067
Is Relevant?: No
7.) Title: La segunda muerte
Score: 0.071106324190653
Is Relevant?: No
8.) Title: Multan Cantt
Score: 0.07059759391097956
Is Relevant?: No
9.) Title: List of Super Smash Bros. for Nintendo 3DS and Wii U tournaments
Score: 0.06769894011675606
Is Relevant?: No
10.) Title: List of Super Smash Bros. Brawl tournaments
Score: 0.06464485259269187
Is Relevant?: No

```

```

Query: extaction of marinee reserrve
1.) Title: Cuinarana Marine Extractive Reserve
Score: 0.1694018661102809
Is Relevant?: Yes
2.) Title: Mestre Lucindo Marine Extractive Reserve
Score: 0.15899653651952428
Is Relevant?: Yes
3.) Title: Cuinarana River
Score: 0.14131318209432184
Is Relevant?: Yes
4.) Title: Mocajuba River
Score: 0.1191556457577246
Is Relevant?: Yes
5.) Title: Barreto River
Score: 0.1165399312087821
Is Relevant?: Yes
6.) Title: Mãe Grande de Curuçá Extractive Reserve
Score: 0.11486641522982216
Is Relevant?: Yes
7.) Title: Curuçá River (Pará)
Score: 0.07129969549040166
Is Relevant?: No
8.) Title: Captain Kidd (pub)
Score: 0.054897697663445634
Is Relevant?: No
9.) Title: Rumen Radev
Score: 0.04291205013874812
Is Relevant?: No
10.) Title: Arthur Giesl von Gieslingen
Score: 0.03446352797918181
Is Relevant?: No

```

Before correction
correction

After

Thank You