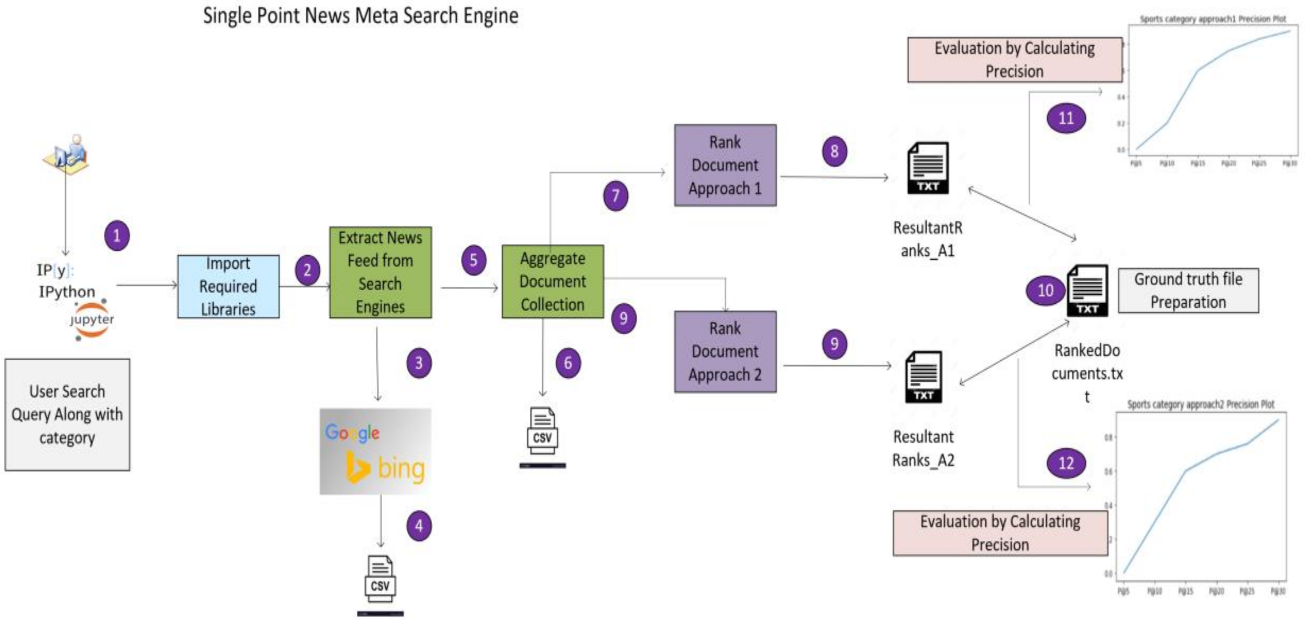
Design and Approach Document

Problem Statement:

Implement a simple “SinglePointNews” Meta-Search engine system and its ranking and evaluation system. As mentioned in the Assignment Task 1

Design / Flow Diagram:


Single Point News Meta Search Engine
1
IP[y]:
IPython
jupyter
User Search Query Along with category
Import Required Libraries
2
Extract News Feed from Search Engines
3
Google
bing
4
CSV
5
Aggregate Document Collection
6
CSV
7
9
Rank Document Approach 1
8
TXT
ResultantRanks_A1
Rank Document Approach 2
9
TXT
Resultant Ranks_A2
10
TXT
Ground truth file Preparation
RankedDocuments.txt
11
Evaluation by Calculating Precision
Sports category approach1 Precision Plot
0.0
0.2
0.4
P@5
P@10
P@15
P@20
P@25
P@30
12
Evaluation by Calculating Precision
Sports category approach2 Precision Plot
0.0
0.4
0.6
0.8
P@5
P@10
P@15
P@20
P@25
P@30

Approach and Assumptions:

1. Considered the Google and Bing news search engines to extract the news.
2. As mentioned in the Assignment statement fixed the date and categories.
3. We need design a meta search engine we have considered the rss feed from both the search engines to extract the news feed-based search query.
4. RSS feed considered as we only need to deal with Title, Summary/Description, Date and link as suggested in the document.
5. Extracted news from both the search engines for a given Category + Query and saved under **/Data/{Category}/** path for a specific Category. With .csv file extension Title, Summary/Description, Category, Date and Link.
6. Aggregated unique documents from extracted documents from both search engines as per step5, written into different file under **Data/{Category}** and stored as uniquenews.csv file.

7. Ranked the documents based on latest extracted based on datetime as search sites were not providing any default ranking. Hence proceeded with this assumption. This satisfies the Task 3 approach 1 strategy.
8. Created a document name to above file along with ranking as stated in step 7. Stored the file under **/Data/{Category}/Rank/ResultantRanks_A1.txt**
9. For Task 3 approach 2 ranking considered **TF-IDF** and **cosine similarity** measures to find the similar documents with rest to query and document collection. Went through the text processing stages to clean the document collection.
10. Once step 9 is complete results get stored under **Data/{Category}/Rank/ResultantRanks_A2.txt.**
11. Constructed the ground truth file based on manual ranking process after going through the details of aggregated unique documents created under step 6. Store the file under **Data/{Category}/Rank/RankedDocuments.txt**
12. Evaluated the retrieval accuracy by calculating the precision with respect to **RankedDocuments.txt vs ResultantRanks_A1.txt .** Plotted the precision accuracy for top5, top10 tp15 etc documents as mentioned in the Assuagement statement.
13. Evaluated the retrieval accuracy by calculating the precision with respect to **RankedDocuments.txt vs ResultantRanks_A2.txt .** Plotted the precision accuracy for top5, top10 tp15 etc documents as mentioned in the Assuagement statement.

14. All Extracted documents are stored under /Data/Sports/ for sports category as .csv file extension.

15. All Extracted documents are stored under /Data/Science/ for sports category as .csv file extension.

16. Aggregated documents are stored under /Data/Sports/ for sports category with.csv file extension.

17. All Ranked documents are stored under /Data/Sports/Rank for sports category with ResultantRanks_A1.txt and ResultantRanks_A2.txt

18. Ground truth file is stored under /Data/Sports/Rank for sports category with RankedDocuments.txt file extension.

19. For better visibility can open .csv (extracted news file) file using excel.

20. For this assignment we have considered Sports and Science Categories.

21. For Sports category date filter **Fri, 01 Jan 2021** is applied to extract news providing corpus in /Data/Sports folder. Results are valid for this date only.

22. For Science category date filter **Mon, 04 Jan 2021** is applied to extract news providing corpus in /Data/Science folder. Results are valid for this date only.

23. As we have considered the Rss news feed from Google and Bing for the data extraction purpose, running the ExtractNews method again will override the Data folder and results may vary as feed varies from date to date. Accordingly, RankedDocuments.txt (Ground truth file) should be constructed manually.