READ ME

Name : Arjee Jacob Jacob
BITS ID : 2019HT12111
Subject Code : SSZG537

Minimum Requirements

Software:
Python Version : 3.7.5
Jupyter Notebook

Hardware:
Windows 10 OS
8-GB Ram
Inte i5 Processor

Note:

This Program was executed on a workstation system with 7th Generation i7 Processor and 16GB of Ram, with SSD Hard Disk.

Hence the EXECUTION time may be different or more for other non-workstation systems

Brief Description

This Program(Python Notebook) is an IR System that reads a Wiki File as an input, generates tokens, pre-processes them using stemming, case-folding etc and generates an inverted index.

The Program will then take a User Query as an input, after which the program will return the top 10 most relevant Documents back to the user.

Here the output will be the top 10 document IDs along with their corresponding Document-Query score.

Note: The Design document will have some repeat information as in README file as well.

Files Provided

IR Assignment - Domain Specific IR System.ipynb
README.pdf
Domain Specific IR SYSTEM Design and Documentation.pdf

IR Assignment - Domain Specific IR System.ipynb
This file needs to be opened and run using Jupyter Notebook program

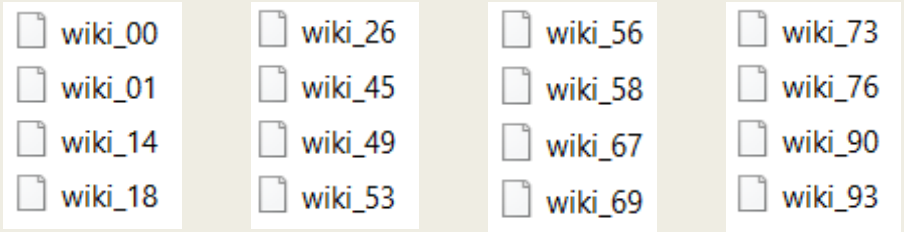
Note: Sample input files wiki_00 and wiki_01 have also been provided. Full list of input wiki files is given in a separate link, which will be given in the following slides.

INPUT SOURCE

The input source is a wiki file. One wiki file will contain a collection of multiple documents along with their document ids in the form of an XML.

Wiki Files have been uploaded to the following google drive location
https://drive.google.com/drive/folders/1i54s7ouUQUkNIX1R65TRjUO3WwLoM3SO?usp=sharing

The Following Wiki Files have been Provided to try



In case the link doesn't work, 2 sample files have been attached to the zip file as well

EXECUTION TIME

- Reading the Wiki File : 3-5 seconds
- Tokenization, Stemming and Inverted Index Construction : 40-43 seconds
- Considering other Fields to execute, Consider Total Time to be close to 1 min.
- Keep an Eye out on the following 2 Code sections as they take time to execute.

START OF MAIN PROGRAM

First Tokenize the documents from the input file

```
1 print("Reading Wiki File. This will take a few seconds...")
2 soup = BeautifulSoup(open(wikiPath, encoding="utf8"), "html.parser")
3 print("Reading Complete. Generated Beautiful soup object from Wiki File")
```

```
Reading Wiki File. This will take a few seconds...
Reading Complete. Generated Beautiful soup object from Wiki File
```

Note : Executing the following will take a while. Please be patient.

```
1 #Initializing the Inverted Index Structure
2 InvertedIndex = {'documentCount' : 0 , 'terms': {}}
3
4 #Create Porter Stemmer object for Stemming
5 ps = PorterStemmer()
6
7 print("Generating Tokens, Stemming and creating the Inverted Index.")
8 print("This will take a while. Please Wait till it says Complete...")
9
```

How to execute part 1 - Input

- First Provide the file path of “any one” of the Wiki Files in the **wikiPath** variable

```
#Provide the path of the wiki file here
wikiPath = "C:/Users/dnv786/Desktop/Zebra/Personal/Mtech/Semester2/IR/Assignment/Sem2Assignment/AA/wiki_18"
```

Each file contains a set of documents with **doc ids**, in xml format

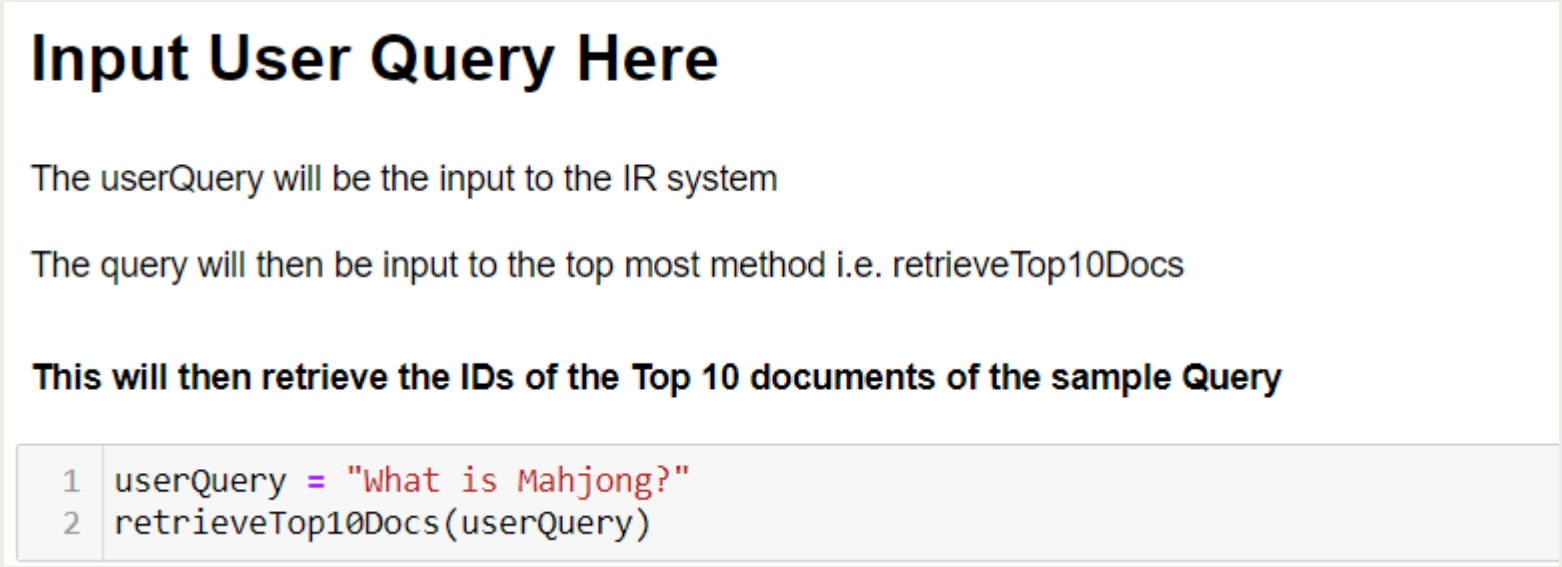
```
1 <doc id="12" url="https://en.wikipedia.org/wiki?curid=12"
  title="Anarchism">
2 Anarchism
3
4 Anarchism is a <a href="political%20philosophy">political
  philosophy</a> that advocates <a
  href="self-governance">self-governed</a> societies based
  on voluntary institutions. These are often described as
  <a href="stateless%20society">stateless societies</a>,
  although several authors have defined them more
  specifically as institutions based on non-<a
  href="Hierarchy">hierarchical</a> <a
  href="Free%20association%20%28communism%20and%20anarchism%2
  9">free associations</a>. Anarchism considers the <a
```

Note:

An **InvertedIndex.json** will get generated in the location where the program is run. It is only an intermediate **output** that is generated for the user to inspect.

How to execute part 2 - Input

- Provide your required Query at the Bottom of the Python *.ipynb* Document



How to execute part 3 - Output

After providing the wiki file path & user query, start executing from the 1st cell to the last cell in sequence. The OUTPUT RESULT will be returned as below in the last cell.

```
1 userQuery = "What is the meaning of Anarchism?"
2 retrieveTop10Docs(userQuery)
```

```
Query Tokens       : ['what', 'is', 'the', 'mean', 'of', 'anarch']
Query Token Count  : {'what': 1, 'is': 1, 'the': 1, 'mean': 1, 'of': 1, 'anarch': 1}
Query Vector : {'what': 0.34516395356044266, 'is': 0.03557615426119612, 'the': 0.03451818910486234, 'mean': 0.2681487692693255
7, 'of': 0.031359669860032684, 'anarch': 2.1712387562612694}
Top 10 Documents that match the sample input query, with corresponding scores :
12 0.1306231786481284
1023 0.1093620818053134
339 0.08474952699982599
696 0.0661941066714508
1176 0.06263976031918828
1212 0.05136417468626282
1158 0.05088670516230048
643 0.04972585571201736
1160 0.04948656879034355
1309 0.0447406502511525
```

OUTPUT EXPLANATION

- The first column shows the doc id in the wiki file

The second column shows the cosine score of the corresponding doc id.

Here the doc ids **12** and **1023** have high scores of **0.13** and **0.10** respectively which makes them more relevant to the query that the user had provided as input

```
Top 10 Documents that match the sample input query, with corresponding scores :
12 0.1306231786481284
1023 0.1093620818053134
339 0.084749526999825 99
696 0.06619410667145 08
1176 0.06263976031918828
1212 0.05136417468626282
1158 0.05088670516230048
643 0.04972585571201736
1160 0.04948656879034355
1309 0.0447406502511525
```

EXAMPLE QUERY REPORT 1

Query	Wiki File	Top 10 Document IDs returned	Document-Query Score	Search Relevance
Who is Jimi Hendrix?	wiki_14	15181	0.157144018	Yes
		15268	0.083441986	No
		15521	0.075422853	No
		15624	0.075214989	No
		15422	0.072996678	No
		15210	0.071610995	No
		15695	0.069093887	No
		15125	0.06869318	No
		15782	0.067903674	No
		15447	0.064475963	No

EXAMPLE QUERY REPORT 2

Query	Wiki File	Top 10 Document IDs returned	Document-Query Score	Search Relevance
What is Mahjong?	wiki_18	19496	0.177541521	Yes
		19328	0.046862029	No
		19327	0.045480866	No
		20041	0.039803807	No
		19581	0.031829796	No
		19958	0.031362642	No
		19372	0.029993652	No
		19738	0.029105519	No
		19719	0.028389043	No
		20088	0.025843326	No

OTHER QUERIES YOU CAN TRY

What is the meaning of Anarchism? – Wiki_00

What is the meaning of Anarchism? – Wiki_00

What is an Analog signal? – Wiki_00

Who is Abner Doubleday? – Wiki_00