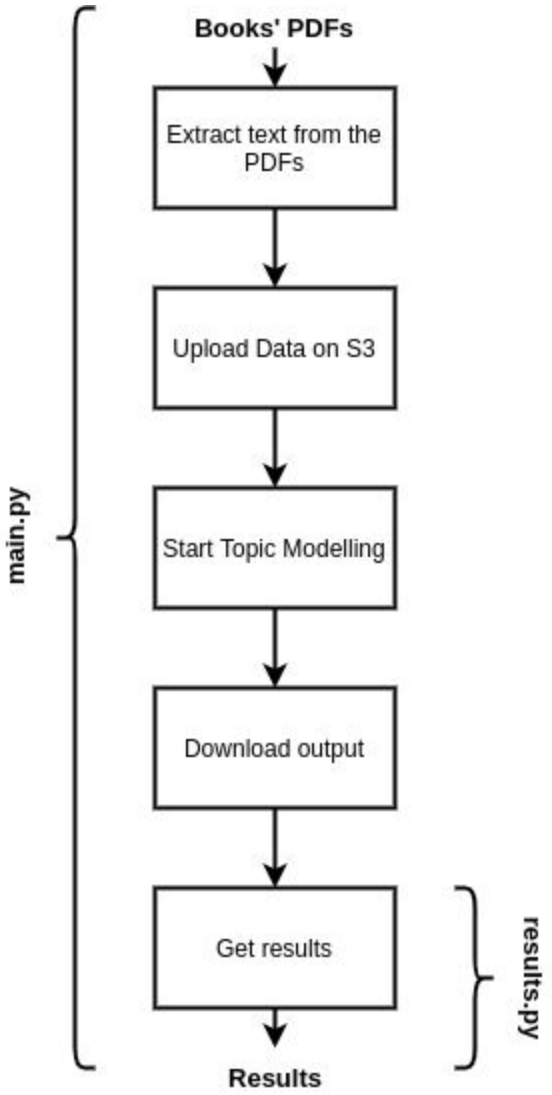
CS607: Contemporary Computing Platforms
Project 1

Submitted by:
Harshit Malik
2017CSB1078

Note: This program has been developed in python language and has been successfully tested with python version 3.8.2 in the Linux machine.

Overview of the program



Required libraries:

- boto3
- botocore
- pymongo
- pandas
- pdfplumber

To install these libraries, run the following command in the project directory

```
$ pip3 install -r requirements.txt
```

How to run the program:

➔ **For generating final results**

◆ Execute the results.py by the following command

```
$ python3 results.py
```

➔ **For complete program run**

◆ Provide the below AWS credentials in main.py

```
ACCESS_KEY = ''
SECRET_KEY = ''
REGION = ''
```

◆ Execute the main.py by the following command

```
$ python3 main.py
```

Results:

- Final results are saved at ‘Results/results.csv’
- Each line of the file denotes one similar topic belonging to different books
- Each column under book names denote the page number of the book where that topic has been identified
- The last column denotes the analogous terms belonging to different books under the same identified topic

- Similar pages belonging to the same topic are also saved in the ‘Results/similar_texts’ directory for easy reference

Example:

Gita	Quran	Taoist	Bible	GuruGranth	analogous_terms
[]	[29]	[155, 58]	[1453, 1433, 1420, 1425, 1436]	[]	['wicked', 'heart', 'good', 'evil', 'righteous', 'soul', 'hath', 'mouth', 'wise', 'eye']
[25, 32]	[]	[100]	[226, 2363, 150, 285, 317]	[]	['day', 'ye', 'night', 'israel', 'house', 'month', 'eat', 'sabbath', 'lord', 'pass']
[]	[]	[145]	[]	[114, 514, 1049, 568, 1043]	['true', 'lord', 'guru', 'shabad', 'nanak', 'naam', 'mehl', 'gurmukh', 'obtain', 'love']

The first row depicts that page 29 of the Quran, pages 58 and 155 of the Taoist, and pages 1420,1425,1433,1436 and 1453 of the Bible are similar and for this similarity, analogous terms have been provided in the last column. These pages are saved at ‘Results/similar_texts/topic_1’.