BDA Mini Project Report

News article scraping and keyword extraction

Sheetal Thakkar 1521002
Farheen Kamal 1521008

I. Introduction

Scraping in simple terms mean to remove an unwanted covering or a top layer from something. There is a wide range of applications of Nearest Neighbour Search, one of which is document analysis. Document analysis is a form of qualitative research in which documents are interpreted by the researcher to give voice and meaning around an assessment topic. There are primarily three types of documents used for analysis viz. Public records, personal records and physical evidences. Document Analysis holds its application in Plagiarism detection, News Aggregator, Text Summarization and so on. The idea for this project was incepted considering the applications and interest in this domain.

II. Problem Definition

This project aims at creating a script to scrape news articles, provided the URL of the article. It also gives keywords and summary of the content using NLP functions provided by Newspaper library. The script also extracts keywords using MapReduce strategy and both these results are compared to get an insight of similarity.

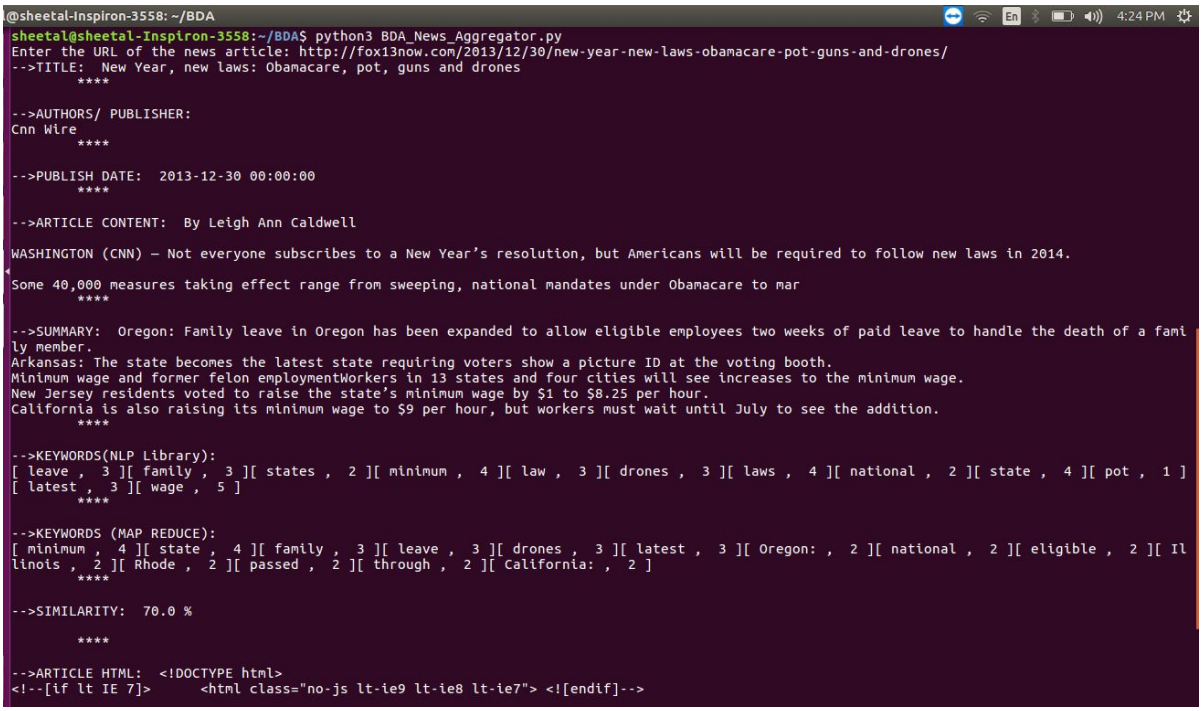
III. Implementation details

This project uses Newspaper, a Python 3 library for extracting and curating articles. Keyword extraction is done using the library functions and also using MapReduce, results from both these approaches are then compared. Text summarization is also implemented, besides, this work also extracts the title, top images and videos from the articles.

Steps for implementation are:

1. Implement mapper and reducer to truncate the common 2 letter and 3 letter words and get the count of frequency of each word in the content. Arrange these words in descending order, and get the top ten keywords.
2. Execute the mapper and reducer using shell script.
3. In the main Python script, import all the necessary libraries and scrape the article using Newspaper.

IV. Result and Analysis



There is around 60% of similarity in the keywords extracted using both the approaches. We also get the summary and HTML script of the article.

V. Application

1. News Aggregator: News Aggregator, is client software or a web application which aggregates syndicated web content such as online newspapers, blogs, podcasts, and video blogs (vlogs) in one location for easy viewing. These applications combine articles from different sources to avoid redundancy of information displayed on their portal. Hence, extracting keywords from the articles and comparing these keywords to check if certain percent of similarity exists can help us in combining articles from different sources.

2. Plagiarism detection: It is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet has made it easier to plagiarize the work of others. This work holds application in this area as well.

VI. Conclusion

We understood the importance and application of Big Data Analytics. We implemented different approaches to extract keyword and did a comparative study.

VII. References

[1] https://github.com/codelucas/newspaper
[2] http://newspaper.readthedocs.io/en/latest/
[3] https://stackoverflow.com/
[4] https://www.python.org/