|  |  |
| --- | --- |
| **Ambadipudi, Chaitra (223009560)** | **Sharma, Aastikta (823007872)** |

Detecting gender bias in news articles

Overview:

This report aims at highlighting the choice of dataset, methodologies used, challenges faced and milestones achieved for task one of the project.

Choice of Dataset: New York Times

The choice of dataset for this project is inspired from the fact that there are several articles being published on a daily basis in different categories and New York Times is one of the most widely acceptable source of latest news. It is the source of data extraction through publicly available API. The online version of the newspaper provides several APIs that are available here: <http://developer.nytimes.com/docs>.

We chose “Sports” and “Movies” category to crawl articles. Our choice of categories was influenced by the fact that most of the articles published in these categories tend to get biased towards one gender. We assume in “Sports” it is male biased and in “Movies” female biased. But, in order to rectify the results we have to perform further analysis. Due to our assumption, we identified these categories as the most biased ones and chose to proceed further in the project with these. We might extend it to other categories if time permits.

Data Crawling:

We created our own crawler (crawler.py) to extract data from the New York Times API. It crawls data from one year and one of the categories (in our case, Sports and Movies) up to 100 pagination which is the limitation of the API. For every crawl we are able to gather at a maximum of 1000 articles for one month and one category.

Data Statistics:

Over a period of almost 10 years (2005-2016), we were able to crawl 103,185 data files for Sports and Movies category.

We noticed that for the year 2005, there were very few articles available and as the year increased number of articles increased exponentially. From a mere 25 articles in 2005, we were able to crawl 11,162 articles for 2015 under Sports category.

Challenges and Solutions:

One of the initial challenges, was during the crawling of data. Since we didn’t come across any documentation regarding the article limit per page, we were able to get only 10 articles per month. Later we identified that every page has a display limit of 10 articles and we had to increase pagination to collect all possible articles.

The other challenge we faced was because of the restriction of the API pagination limit of 100. Since, every month had a pagination limit of 100 pages we couldn’t get any data beyond 1000 articles.

Since we couldn’t resolve this, we chose to increase our area of focus from Sports to Movies and collect more articles from “Movies” category as well.

Future Tasks:

In future tasks, we plan to pre-process the crawled data and using Solr index them to further perform query processing. We also plan to use word cloud and graphical visualizations for the results.