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

DETECTING GENDER BIAS IN NEWS ARTICLES

Newspapers are a strong source of data over the years, and more importantly, a timeline of opinions on how the society is functioning. They also help readers form opinions based on what they interpret from the readings. Within such an intricate web of articles, are there any unconscious biases hidden underneath, which may then be unconsciously perceived by the readers? This is what we aim to track. More specifically, we are looking at gender bias, and how it has changed over the years.

For this purpose, we have chosen to gather articles published over a decade by the New York Times. The online version of the newspaper provides several APIs that are available here: http://developer.nytimes.com/docs. Although gender bias can be detected over various topics, we will initially start with the Sports category, since we expect significant differences in this area. Primarily, we want to track, through visualizations:

1. How gender bias has changed over a 10-year period and
2. Within the United States, how does this bias differ with respect to each state?

To achieve this, we will first categorize *who* each article is talking about. For this, we will use Named Entity Recognition. There are several Named Entity Recognition tools available, out of which the most famous is the one provided by Stanford NLP (developed in Java). Once we know who the article is about, we will then use gender labeling to distinguish the gender of the person. As [1] mentions, this can be done by training the extracted names against the *Popular Baby Names* provided by the US Social Security Administration Database.

[1] http://chenlab.ece.cornell.edu/people/Andy/Andy_files/1424CVPR08Gallagher.pdf

For searching and indexing, we will use the recommended Solr framework. For the front-end, we plan to use a combination of Bootloader and the Bananajs framework. For specific visualizations, we will make use of the D3.js charting framework. However, the tools that we propose to use are currently tentative, and may change depending on the requirements or issues we face later on in the project.

We have also considered evaluation methods that can be applied to the project. There are two major checks to be done:

1. How accurate is the gender labeling?
2. Are the results we achieved comparable to existing research?

The accuracy of gender labeling can be evaluated by us by testing against the Google Freebase data dumps. For part (2), we also plan to compare our results to a similar research done by Ali et al [2], particularly for the Sports category.

In summary, our project aims to give a visual perspective into how gender bias in news articles has changed (or not) through a 10-year period, further broken down by state.

Future work:

1. Additionally, if the proposed work is completed before the estimated time frame, we may add more 'stereotypical' categories such as business, politics, and entertainment, and 'neutral' categories like top stories, or front page stories.
2. With respect to the visualizations, we could add word clouds. It will be interesting to see what words are associated in articles that speak about females, versus articles that speak about males.
    3. Although this is not currently being considered for our project (since it does not directly relate to gender bias), another demographic to consider in the visualizations would be age. How frequently are middle-aged men and women talked about as compared to older men and women.

[2] http://jmlr.csail.mit.edu/proceedings/papers/v11/ali10a/ali10a.pdf