Citation Prediction Bioinformatics
Luke Davis, Mouna Kalidindi, Darpan Jhawar, Charles Stamey, Logan Hornbuckle
CSC 505-01 Fall 2018 UNCG

After the 3rd presentation, we started working towards applying Machine Learning on our dataset.

We first had to perform NLP to convert all textual columns to numeric. We had 3 such columns. We first removed all stop words, performed stemming and then used TFIDF. We used 50 topics to train the model. For each row we used only the topic which had the highest probability. We also added a column where we mentioned the number of topics we got for each row.

The next task was to decide which Machine Learning algorithms to use. We decided to use both classification and regression models. We figured out that for our dataset Random Forest and XGB were the best algorithms.

We first tried regression and the results were not very good. So we went ahead and tried classification models. To divide up the classes we tried different techniques. They were:

1. Use pandas qcut
2. Do a Box Cox transformation and then do a Z score
3. Do a log transformation and then a Z score
4. Used Quartiles(Divided citations based on quartile ranges)

We tried all combinations of dividing the classes and ML algorithms and the highest accuracy we received was 74%. This was when we used Log Transformation with XGB.

Tasks assigned to each Team Member:

Luke:
Created the following columns based on the NLP processing:

a) ‘Paragraph’ # of topics
b) ‘Paragraph’ top topic
c) ‘Paragraph’ top topic probability
d) ‘Title’ # of topics
e) ‘Title’ top topic
f) ‘Title’ top topic probability
g) ‘Author Keywords’ # of topics
h) ‘Author Keywords’ top topic
i) ‘Author Keywords’ top topic probability

These columns were added to the master_file for the purposes of Machine Learning.

Mouna:
Applied Regression algorithms like Random Forest Regressor and XGBRegressor on the data set.
Used quartile ranges to group citations into 3 categories. class1 includes quartile 1(0-1 citations), class2 includes IQR(2-12 citations) and class3 quartile 3(greater than 12 citations).
Applied classification machine learning algorithms on this

Darpan:
Applied classification machine learning models (Random Forest and XGB) on the dataset while trying out different techniques to divide the rows into classes (qcut, log transformation and box cox transformation).

Steve:
Utilizing gensim for my topic modeling aspect of our project, I was able to get scalable statistics semantics. This aided m analysis for any plain text documents I used in the ipython notebook based on LSA and LDA. This was sufficient until I ran out of RAM, being that gensim is memory dependant.

Logan:
Using the NLP toolkit, nltk, I assisted in converting our non-numerical data to numerical and general “cleaning of the data”. Such tasks included:

- Filling all NaN values with empty string and removing certain stopwords.
- Performed stemming using Snowball Stemmer.
- Creating lists from columns ‘Para’ , ‘Title’ and ‘Author Keyword’