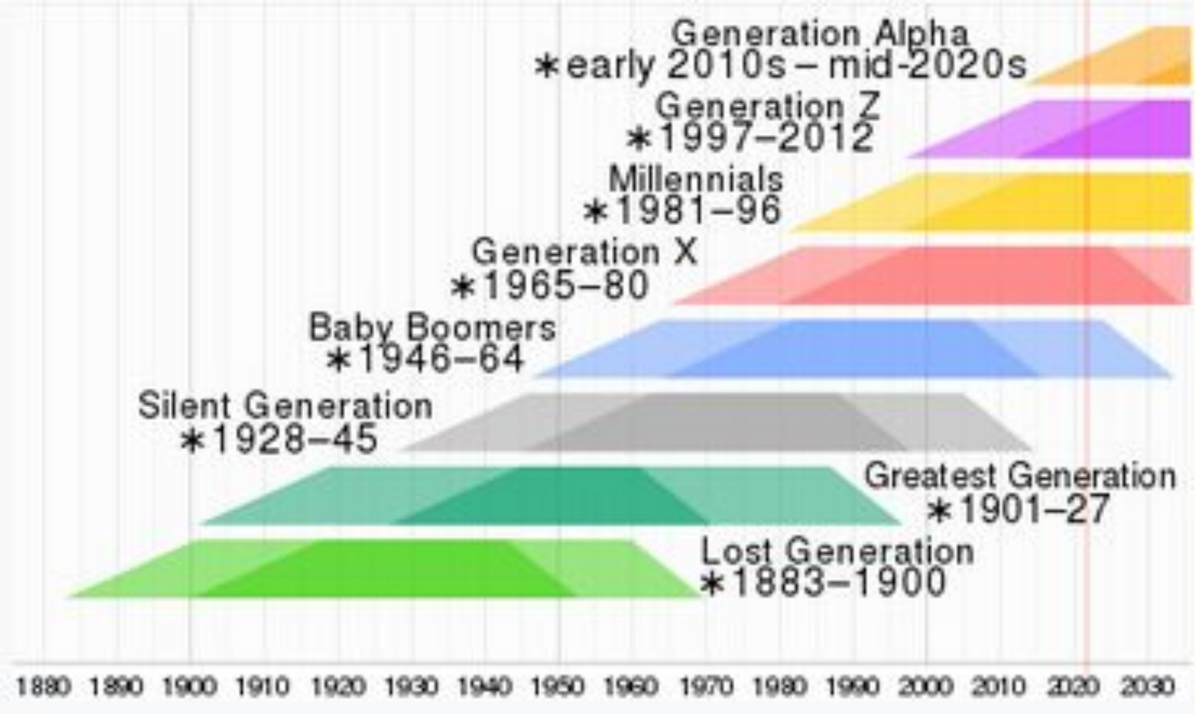
Project 3: Web APIs & NLP

By Matthew Edelmann

Project Description

- Scrape 2 different subreddits for their data with API’s
- Use NLP models to find an patterns and determine which subreddit a post belongs in just given the words.

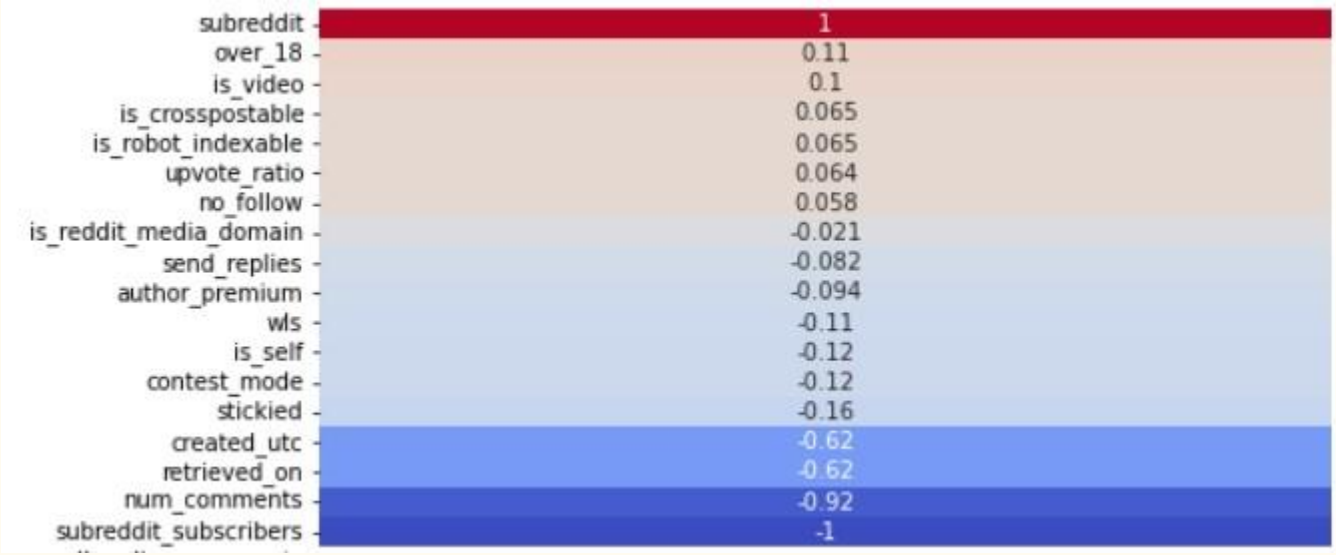
Millenial Vs GenZ



Scrapping

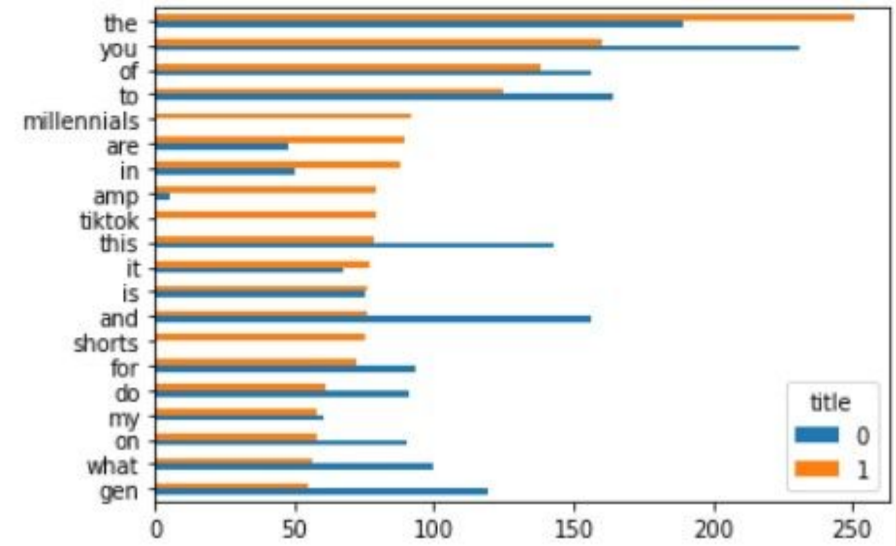
- I used API’s to scrape the 2 subreddits
- I used the request.get() API
- I took 1000 entries from each subreddit

Numeric Values Heatmap

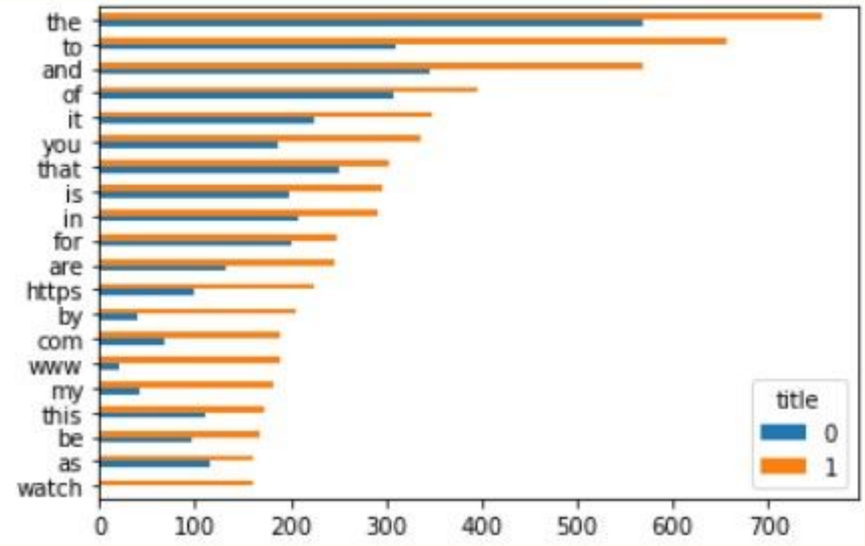


Bar Charts

0 is GenZ; 1 is Millennial

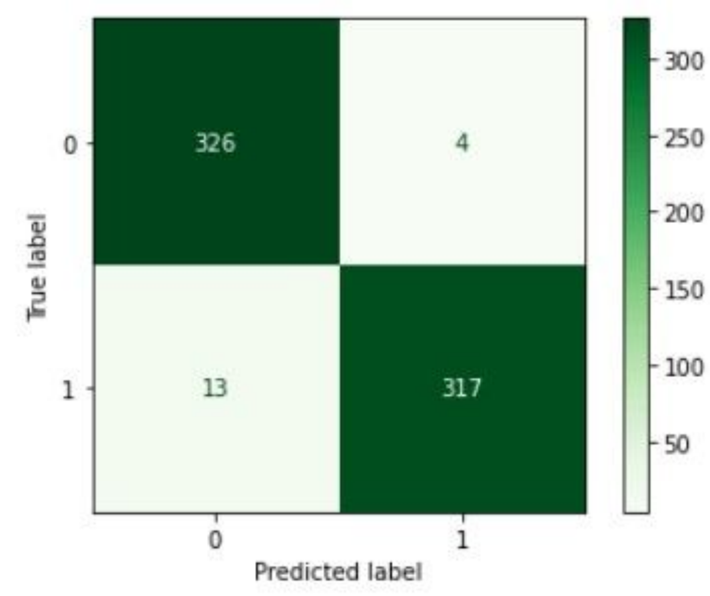


Title



Selftext

Confession Matrix; Title

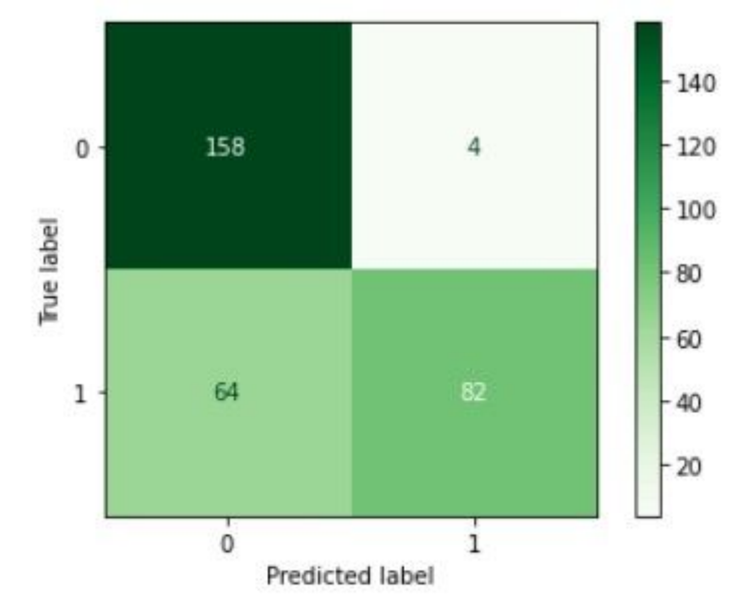


It seems that out of 660 testing data, the model only predicted 17 entries incorrectly. Here are it's train and r^2 score:

Train: 0.9828358208955223
Test: 0.9742424242424242

This indicates that out model is 97.4% accurate in predicting whether a title is in the millenial subreddit or the GenZ one.

Confession Matrix; Selftext



This shows most of the problem comes when the model predicts for the GenZ subreddit and it is really in the millenial subreddit. Here are it's train and r^2 score:

Train: 0.8057784911717496
Test: 0.7792207792207793

This shows that the model is 77.9% accurate in predicting the subreddit given subtext. It is slightly overfit.

Random Forest and Extra Trees Classification

Finally we look at the Random Forest and Extra Trees Classification models. The cross-validation score was the same for both Random Forest and Extra Trees but differ in title and selftext:

Title: 0.9858208955223879

Selftext: 0.8089548387096773

This shows that they give a 98.6% accuracy when predicting the title and a 81% accuracy when predicting the selftext.

Conclusion

Given a title, it is easy to create a model that can determine if the title is from the millenial subreddit or the GenZ subreddit. If given selftext, I recommend either the RandomForestClassifier or the ExtraTreesClassifier.

Questions?