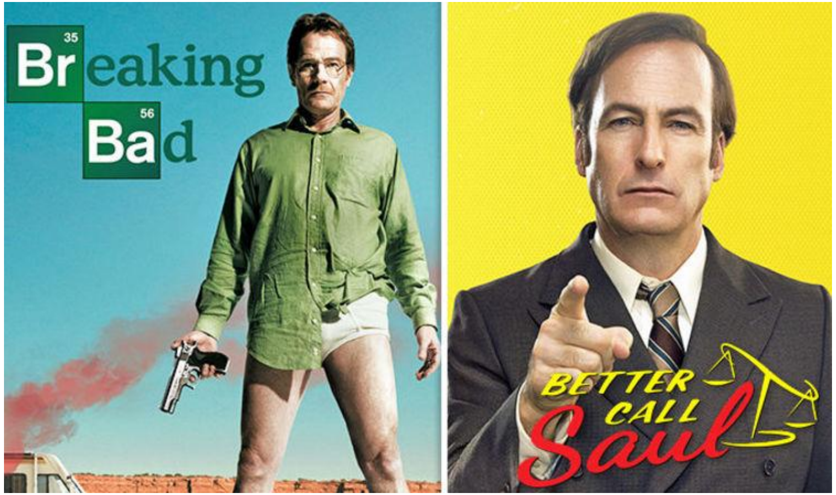
Breaking Bad or Better Call Saul:

Can text from Reddit be used to guess which show a user is talking about?

DSI Project 3

By: Mario Sanchez, Jr.



Objectives:

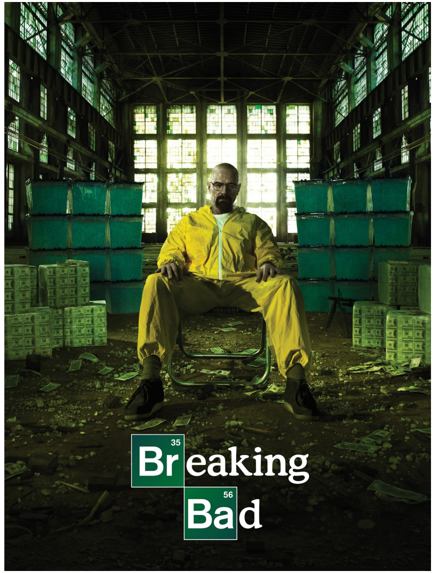
- Obtain text data from reddit posts
- Decide on a model that classifies best
- Which words mattered most?
- How can we use this information to improve our marketing?

Background

From IMDB:

Breaking Bad:

- “A high school chemistry teacher diagnosed with inoperable lung cancer turns to manufacturing and selling methamphetamine in order to secure his family's future.”
- Starring: Bryan Cranston, Aaron Paul, Anna Gunn
- Ran from 2008-2013



Better Call Saul:

- “The trials and tribulations of criminal lawyer, Jimmy McGill, in the time leading up to establishing his strip-mall law office in Albuquerque, New Mexico.”
- Starring : Bob Odenkirk, Rhea Seehorn, Jonathan Banks
- Started in 2015, still running

The Breaking Bad subreddit has 840k members
The Better Call Saul subreddit has 159k members

What Did We Do?

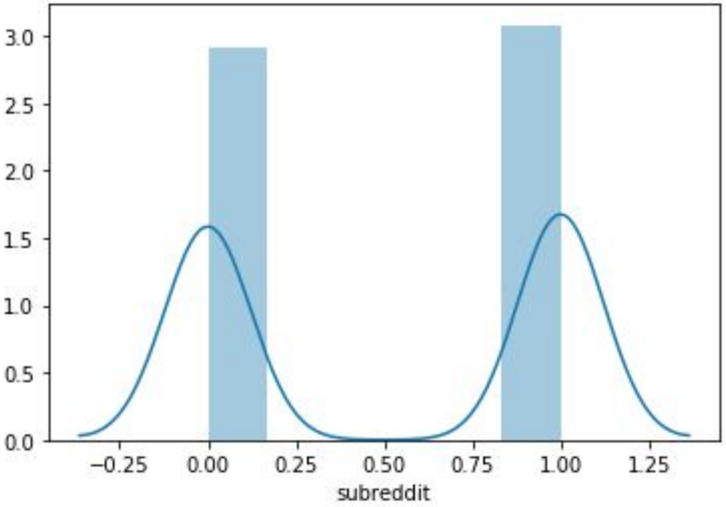
1. Obtain data from two subreddits by scraping posts over several days
2. Using text from these posts, create a classification model that can predict which show an individual is talking about
3. Do certain words matter most?
4. Can this model be used on other platforms in order to figure out if people are talking about either of these shows?

Modeling

Models:

- Logistic Regression
- K-nearest neighbors
- Multinomial Naive Bayes

Distribution of Subreddit Posts



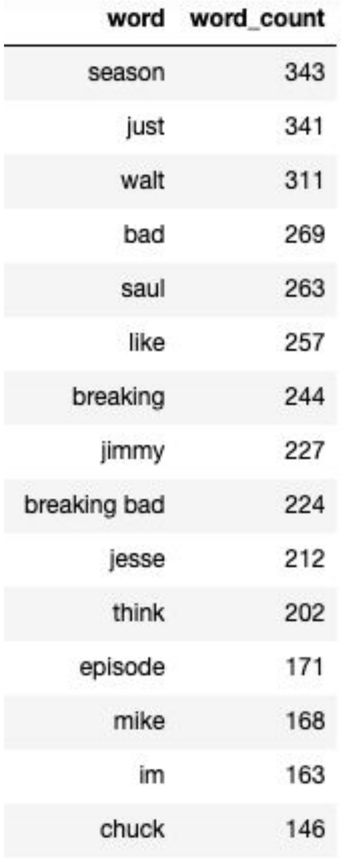
Post Distribution:

- Breaking Bad: 51%
- Better Call Saul: 49%

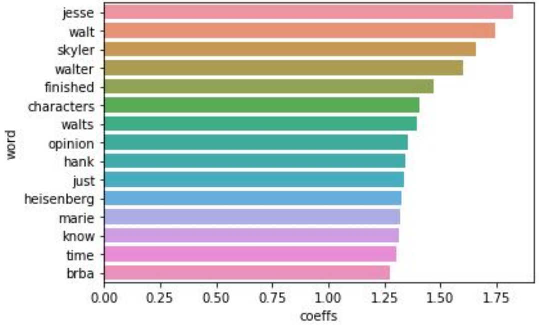
Text Manipulation:

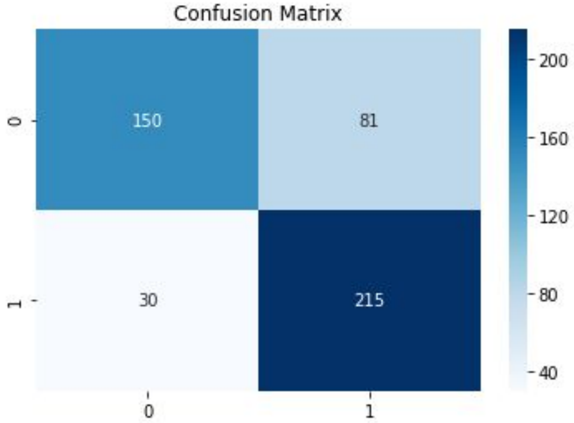
- Tokenize
- Lemmetize
- Remove common words

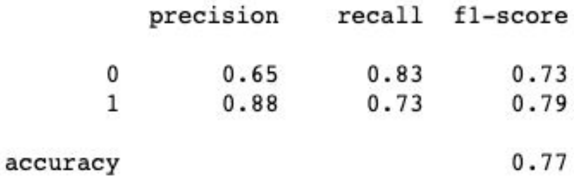
Model Results



word	word_count
season	343
just	341
walt	311
bad	269
saul	263
like	257
breaking	244
jimmy	227
breaking bad	224
jesse	212
think	202
episode	171
mike	168
im	163
chuck	146







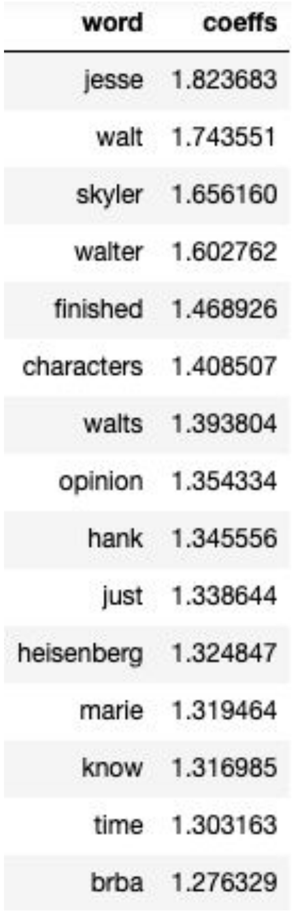
```
              precision    recall  f1-score

           0       0.65      0.83      0.73
           1       0.88      0.73      0.79

    accuracy                           0.77
```

Words per post:

- Breaking Bad: 417
- Better Call Saul: 532



word	coeffs
jesse	1.823683
walt	1.743551
skyler	1.656160
walter	1.602762
finished	1.468926
characters	1.408507
walts	1.393804
opinion	1.354334
hank	1.345556
just	1.338644
heisenberg	1.324847
marie	1.319464
know	1.316985
time	1.303163
brba	1.276329

Summary

Using a logistic regression model we were able to predict which post the text was from 77% of the time. While this is not perfect, I believe that it can effectively be used to classify if a person is talking about Breaking Bad or Better Call Saul using only text.

This model can then be used on other reddit, twitter, tvguide, etc. posts to determine if people are talking about a certain show.

This information can then be leveraged to direct your marketing efforts for new shows or projects related to these two great shows.

Limitations and Assumptions:

- More data can improve models
- Using frequency of words is best for prediction

Future Analysis:

- Utilize different text manipulation
- Test models on other TV show subreddits that lie in the say category