THIS IS WHERE THE FUN BEGINS...

Finding the differences between r/SequelMemes and r/PrequelMemes subreddit post titles

By: Rajan Davis

Presentation Agenda

- Potential Issues
- Process
  - Data Gathering
  - Metrics
  - Exploratory Data Analysis (EDA)
  - Attempt at using Singular Value Decomposition (SVD)
- Conclusions
- Further Analysis

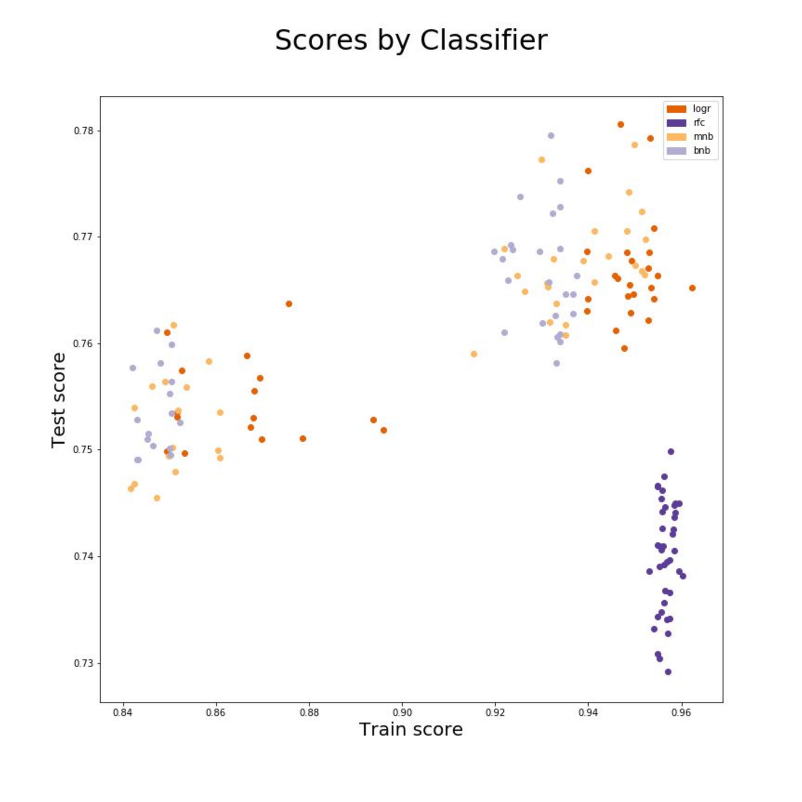
Potential Issues

- r/SequelMemes and r/PrequelMemes are primarily image-driven subreddits.
    - Context can be lost without images
    - Titles may not reveal subreddit as easily
- Class imbalance
    - Larger sample size of r/SequelMemes post titles (14600) in comparison to r/PrequelMemes posts (13250)
    - However, the date range of the created posts is shorter for r/PrequelMemes (December 26th, 2016 to March 14th, 2017 - a little less than 3 months) than for r/SequelMemes (January 14th, 2017 to May 4th, 2018 - almost a year and 4 months)

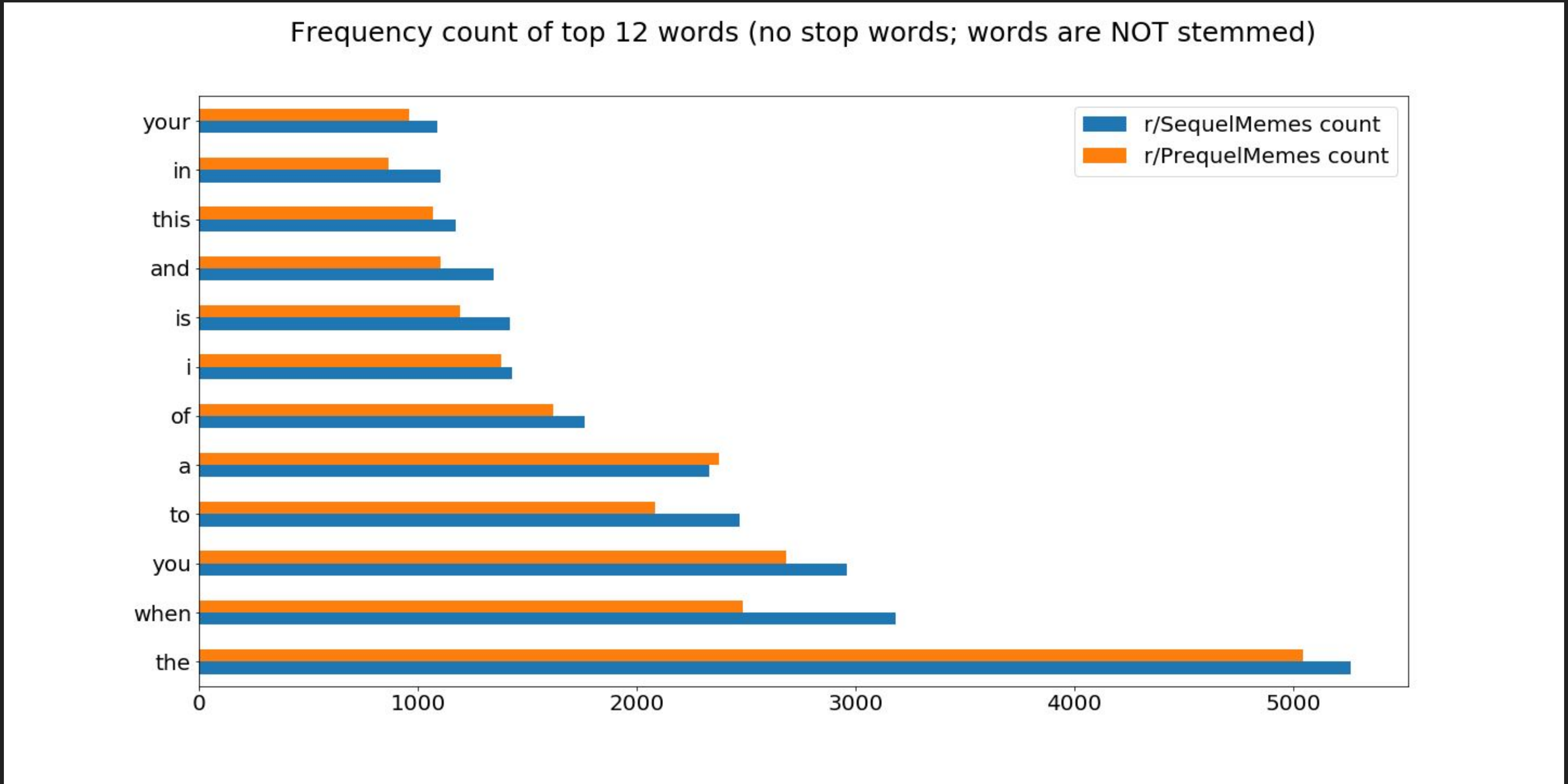
Process - Data Gathering

- Data gathering was initially performed using the Reddit API; however, I ended up using the Pushshift.io Reddit API for gathering my data.
- Main difference between the two:
  - Reddit API is live and post information can change
  - Pushshift is historical - has data from first posts of a subreddit, but may not have the most up-to-date data for a post

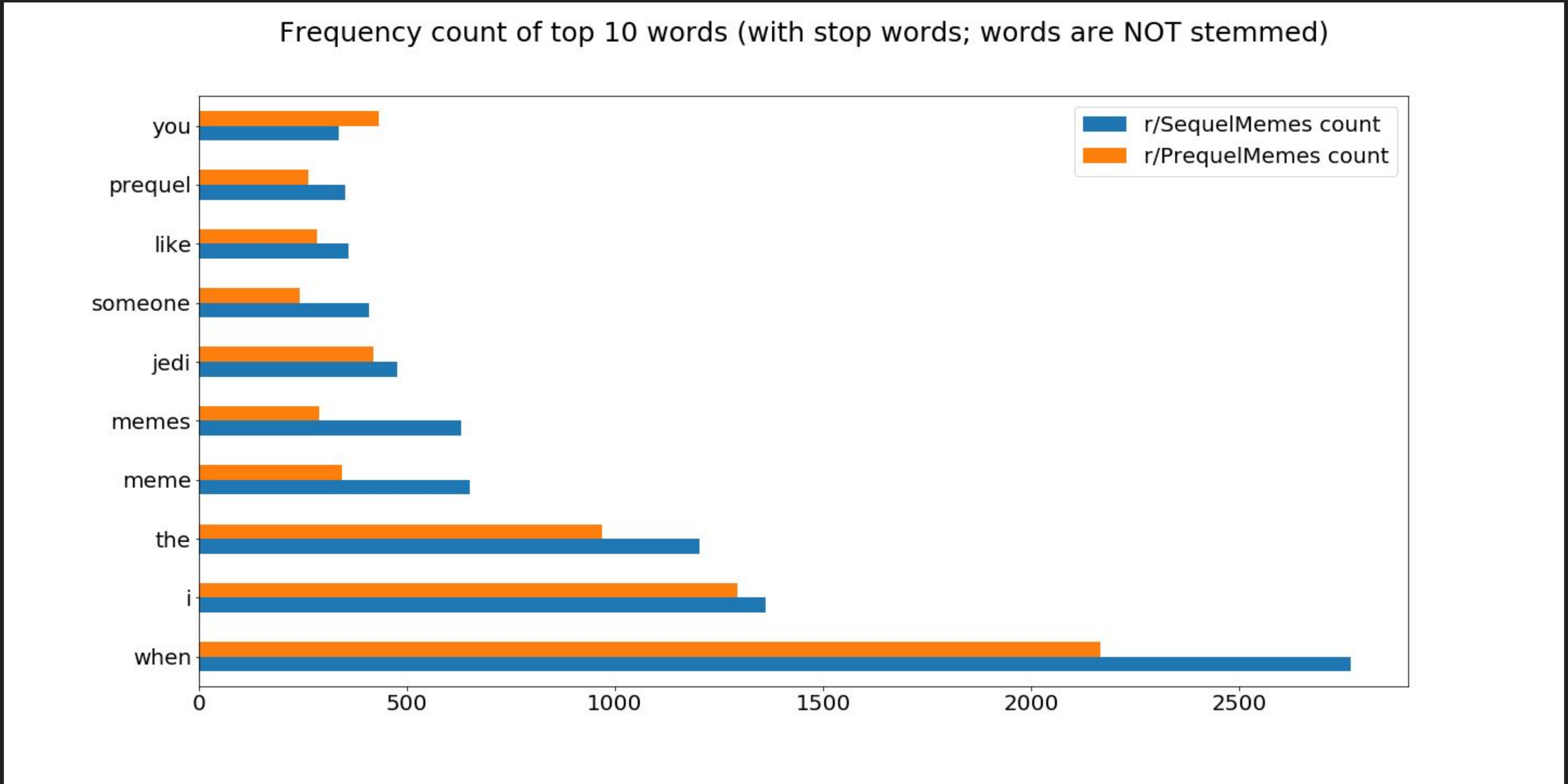
Process - Metrics



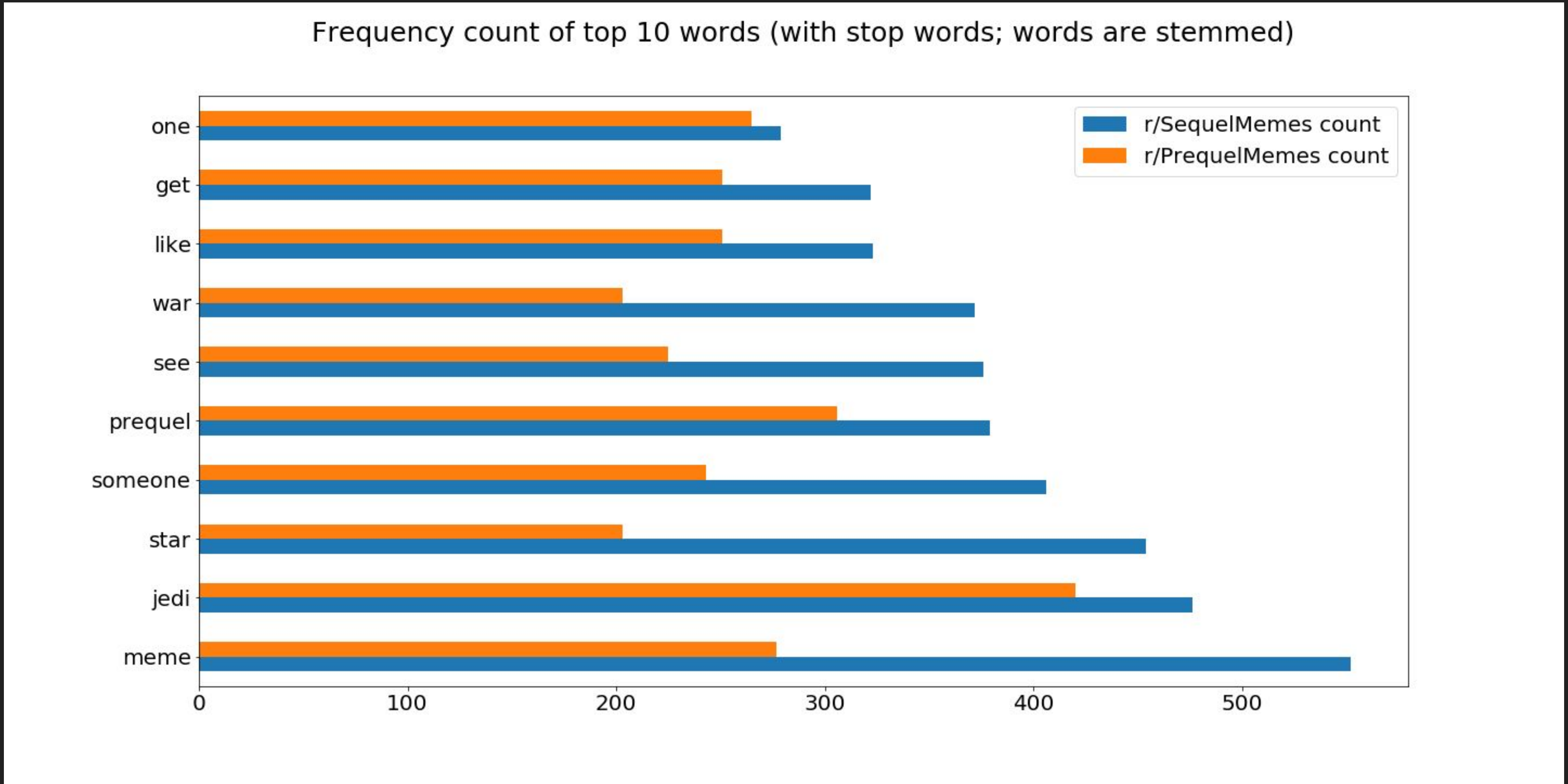
Process - EDA


Frequency count of top 12 words (no stop words; words are NOT stemmed)
r/SequelMemes count
r/PrequelMemes count
your
in
this
and
is
i
of
a
to
you
when
the
0
1000
2000
3000
4000
5000

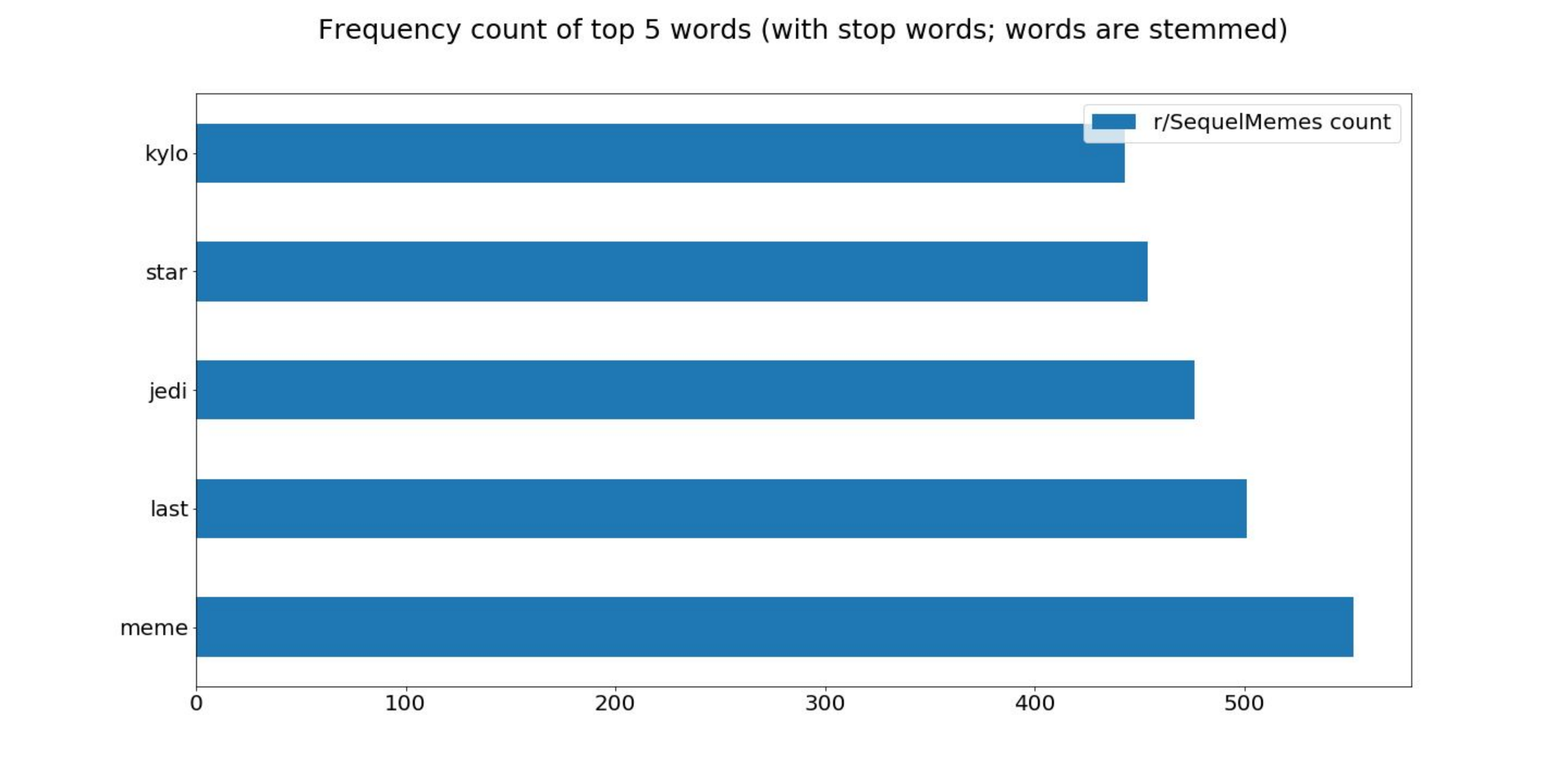
Process – EDA (Cont’d)



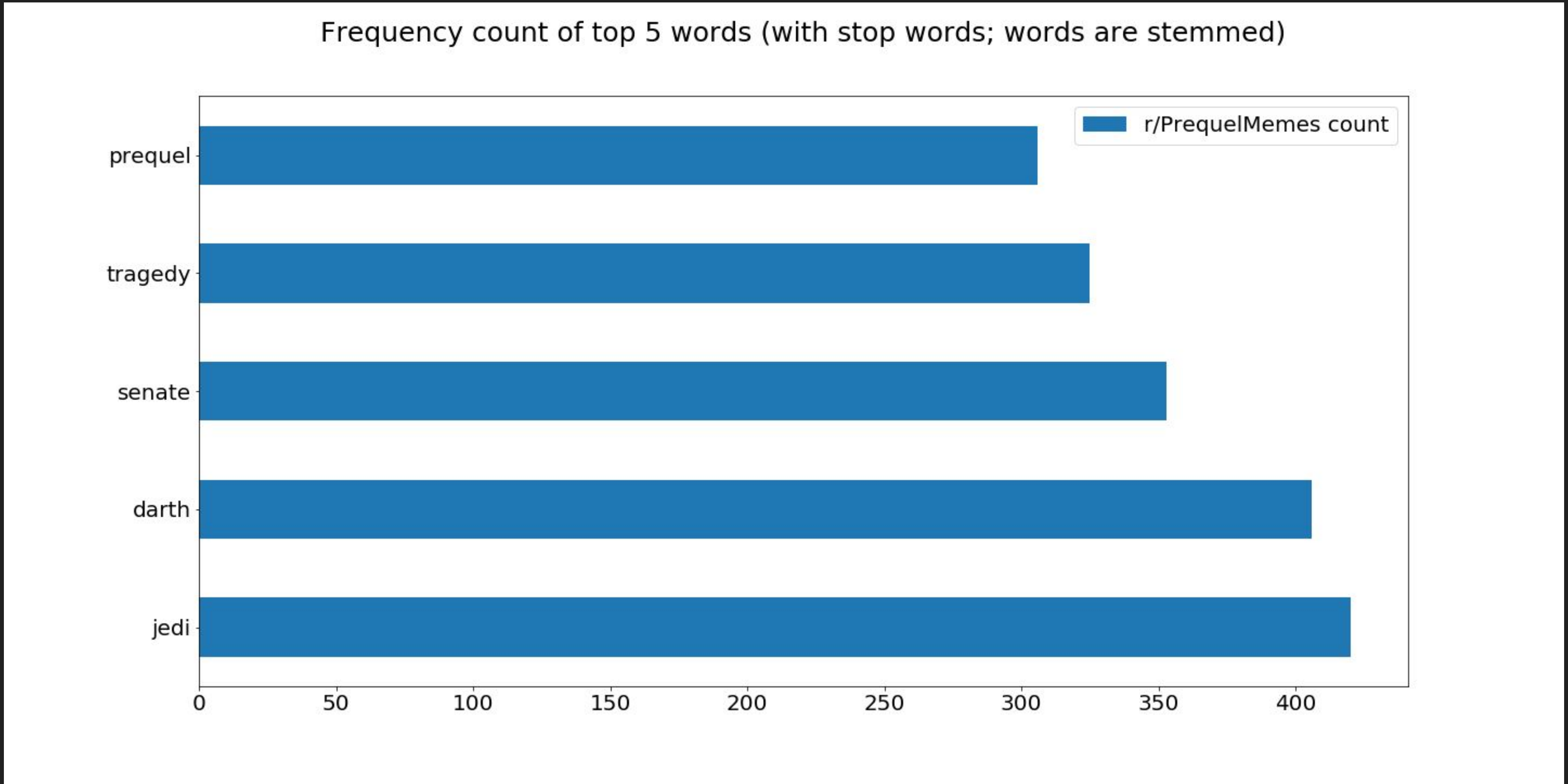
Process – EDA (Cont’d)



Process - EDA (Cont’d)



Process – EDA (Cont’d)



Process - SVD

- Attempted to utilize Singular Value Decomposition (SVD), a tool for scaling the meaning from text
- However, accuracy scores became *worse…*
- I suspect that similarities between subreddit post titles may impact how *meaning* can be extracted

Conclusions

- Between r/SequelMemes and r/PrequelMemes, we can classify a Reddit post with an accuracy as high as 78% using only the post title
- More analysis should be done using image classification
  - See how similar image macros are between subreddits
  - See how much signal images may add or detract from classification
- Tweaking hyperparameters did not help too much
  - Most accuracy scores were centered around 70 -75%
  - More review with lemmatizing and stemming words

Questions?