/r/Project 3: reddit and NLP

/u/Umair Evans

02.04.2021

/r/Data Mining & Cleaning

- Function to scrape for parameters and save to CSV:
  - Epoch
  - Subreddit
  - Number of posts

- For all analysis, only “selftext” and “title” were used

/r/LifeProTips

or

/r/ShittyLifeProTips

/r/Data Issues

/r/LifeProTips

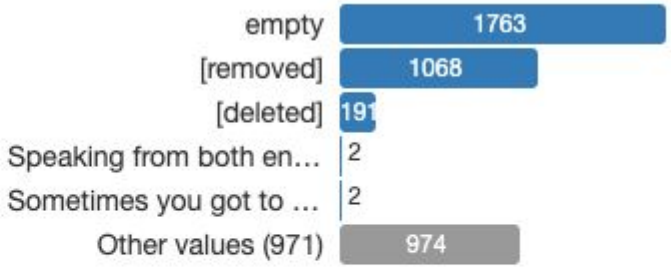
or

/r/ShittyLifeProTips

selftext
Categorical

HIGH CARDINALITY

Distinct	976
Distinct (%)	24.4%
Missing	0
Missing (%)	0.0%
Memory size	31.4 KiB



Toggle details

/r/Title Cheating

/r/LifeProTips

or

/r/ShittyLifeProTips

LPT: For people starting a new job. If a task typically takes someone ~3.5 hours and you can get it done in 1 hour, don't turn your task in right away - wait about an hour. If your manager(s) discover how productive you really are, they will quickly overwork you without proper compensation.

/r/Models

	LifeProTips	ShittyLifeProTips	accuracy
precision	0.940659	0.823853	0.877
recall	0.816794	0.943277	0.877
f1-score	0.874362	0.879530	0.877
support	524.000000	476.000000	0.877

CountVectorizer & Logistic Regression

/r/LifeProTips

or

/r/ShittyLifeProTips

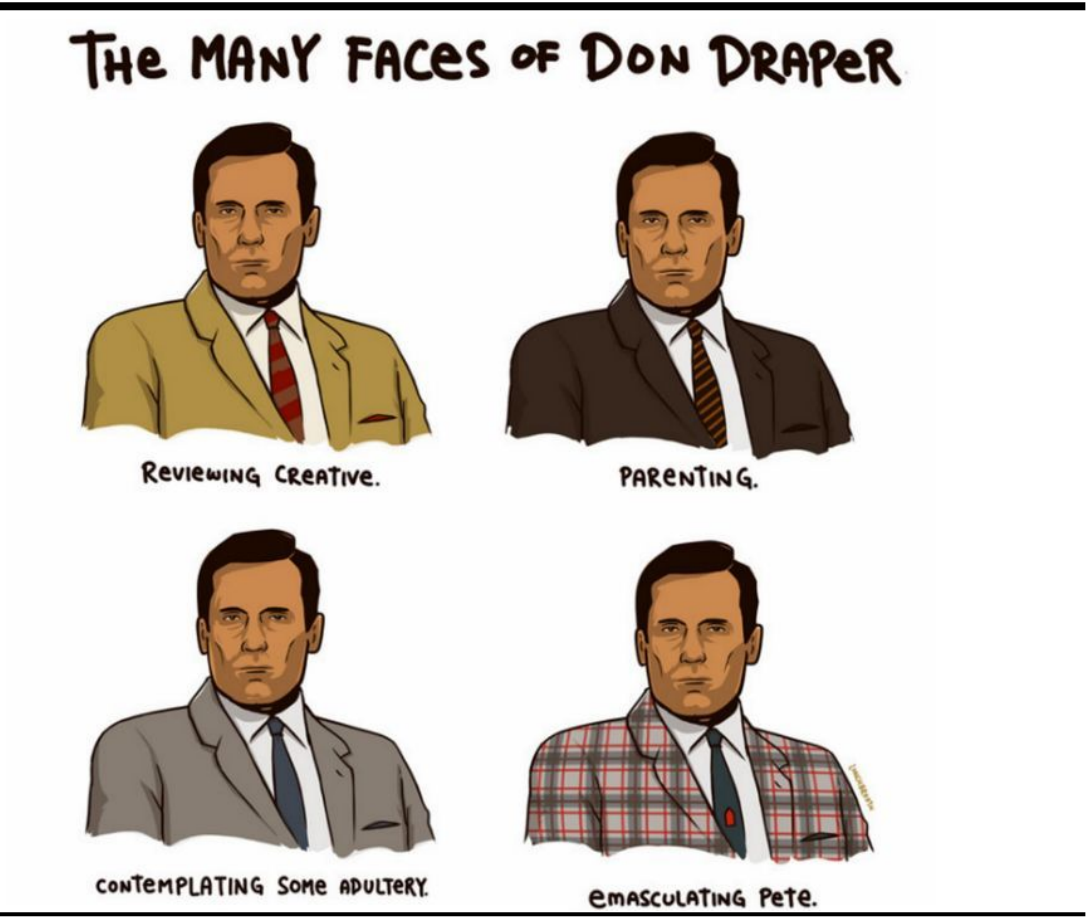
TfidfVectorizer and Bernoulli Naive Bayes

	LifeProTips	ShittyLifeProTips	accuracy
precision	0.972028	0.655462	0.746
recall	0.530534	0.983193	0.746
f1-score	0.686420	0.786555	0.746
support	524.000000	476.000000	0.746

/r/MadMen

or

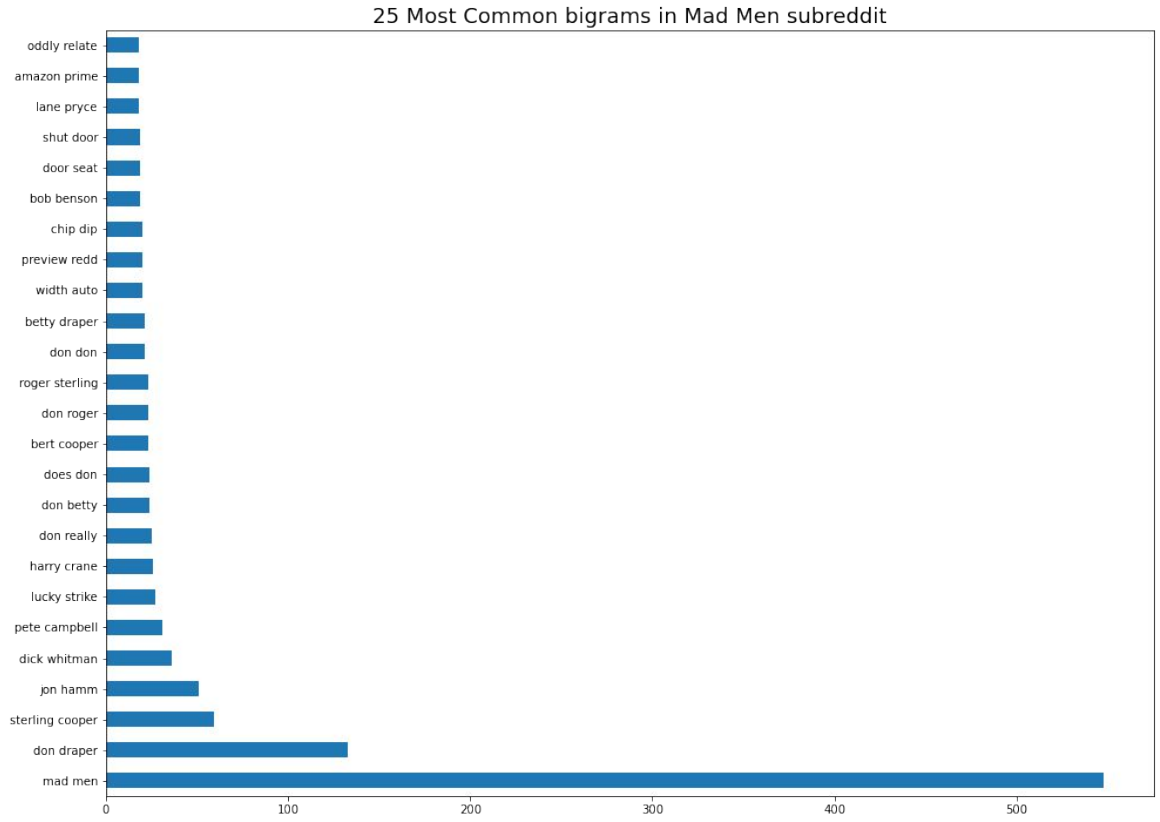
/r/TheSopranos



/r/MadMen

or

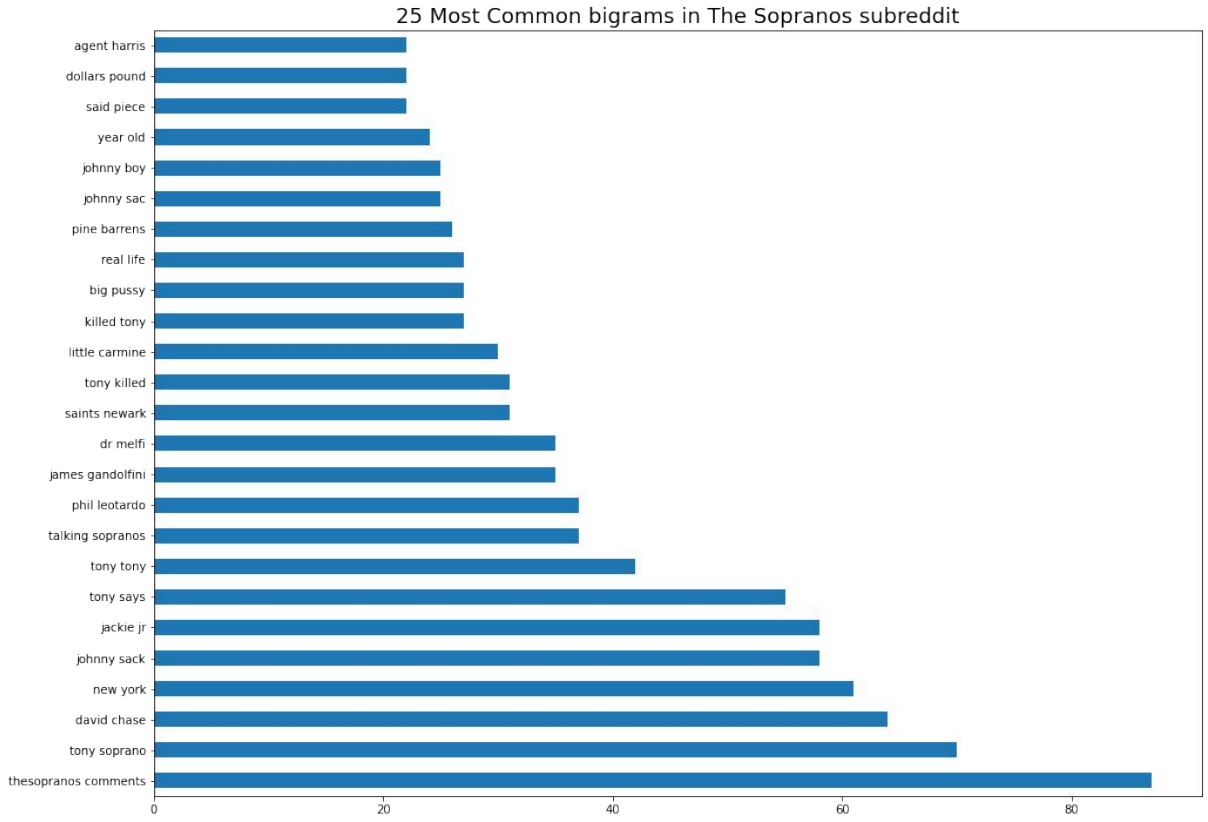
/r/TheSopranos



/r/MadMen

or

/r/TheSopranos



/r/LifeProTips

or

/r/ShittyLifeProTips

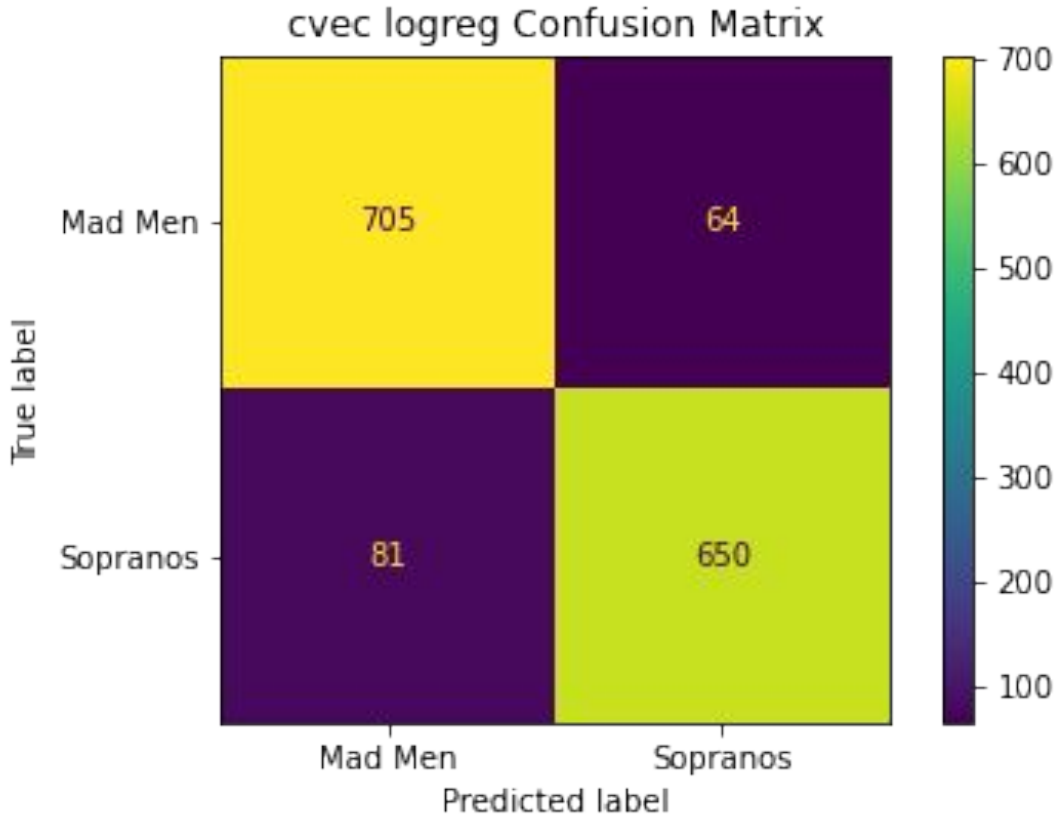
/r/Modeling Choices

- CountVectorizer & Logistic Regression
- TfidfVectorizer & Bernoulli Naive Bayes
- CountVectorizer & K-Nearest Neighbors
- A Sacagawea Dollar (baseline model)

/r/MadMen

or

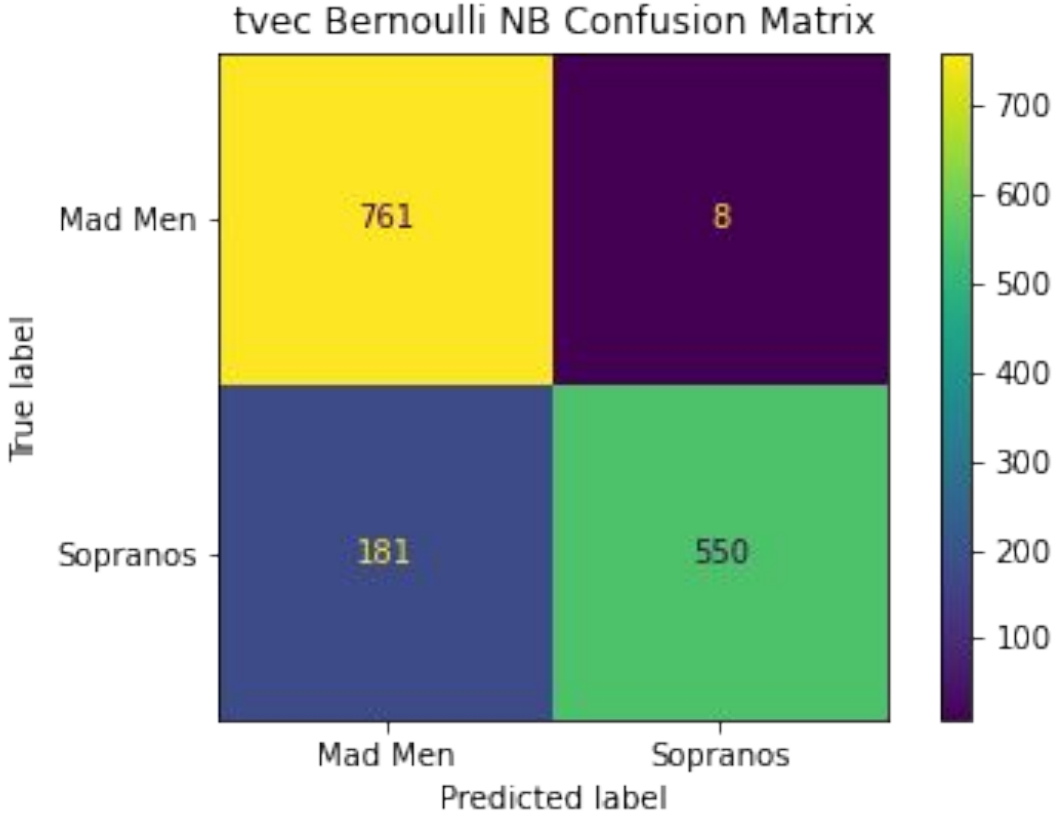
/r/TheSopranos



/r/MadMen

or

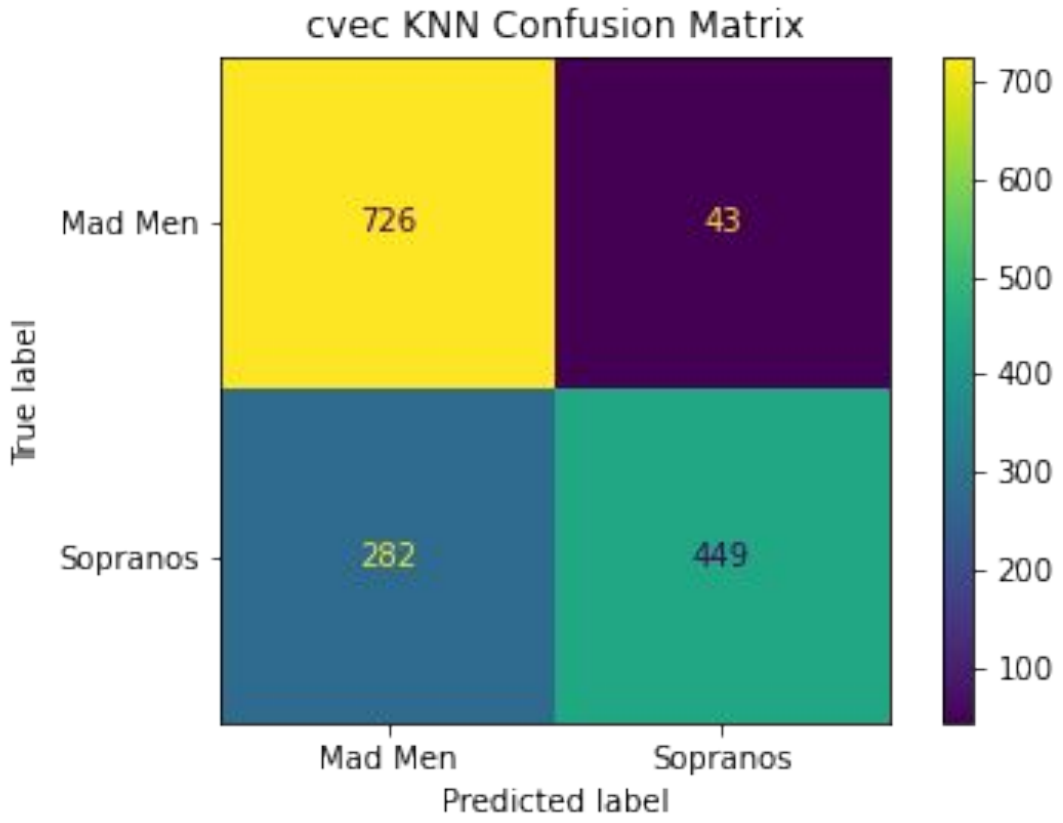
/r/TheSopranos



/r/MadMen

or

/r/TheSopranos



/r/MadMen

or

/r/TheSopranos

/r/Conclusions

- Use RegEx to clean your data
- Add common words to Stop Words
- Utilize BiGrams and TriGrams
- CountVectorizer and Logistic Regression: Winning Duo

/r/Project 3

or

/r/Capstone

/r/Future Research

- Increase amount of stop words, or…
- Explore using bag of words to create “definers”
- Explore grammar of each subreddit through SpaCy
- Find what is causing the poor Recall Scores for The Sopranos
  - Must be the gabagool

/r/Project 3

or

/r/Capstone