The West Wing vs. House of Cards

...

Web APIs, NLP, and Machine Learning

By Jonathan Benton

Overview

- Getting the Data
- Modelling Process
- Analysis
- Another Angle
- Next Steps
- Q & A

Why These Subreddits?

-Wholesome
-Idealistic
-Everyone is good!

Subreddit Members: 23,456

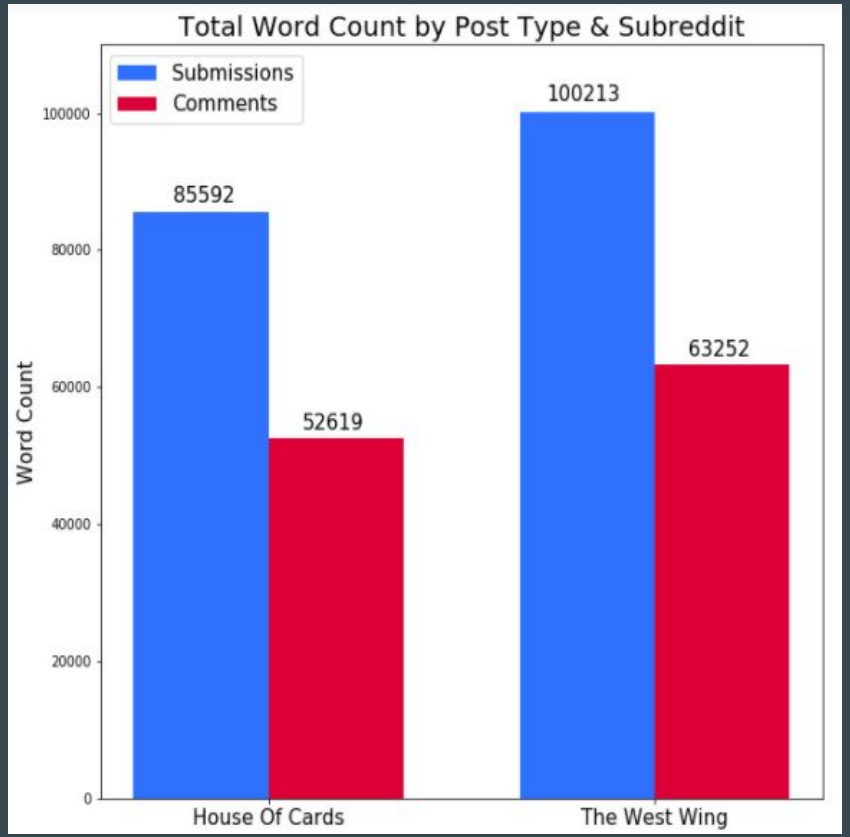
Scandalous-
Cynical-
Everyone is bad!-

Subreddit Members: 69,481

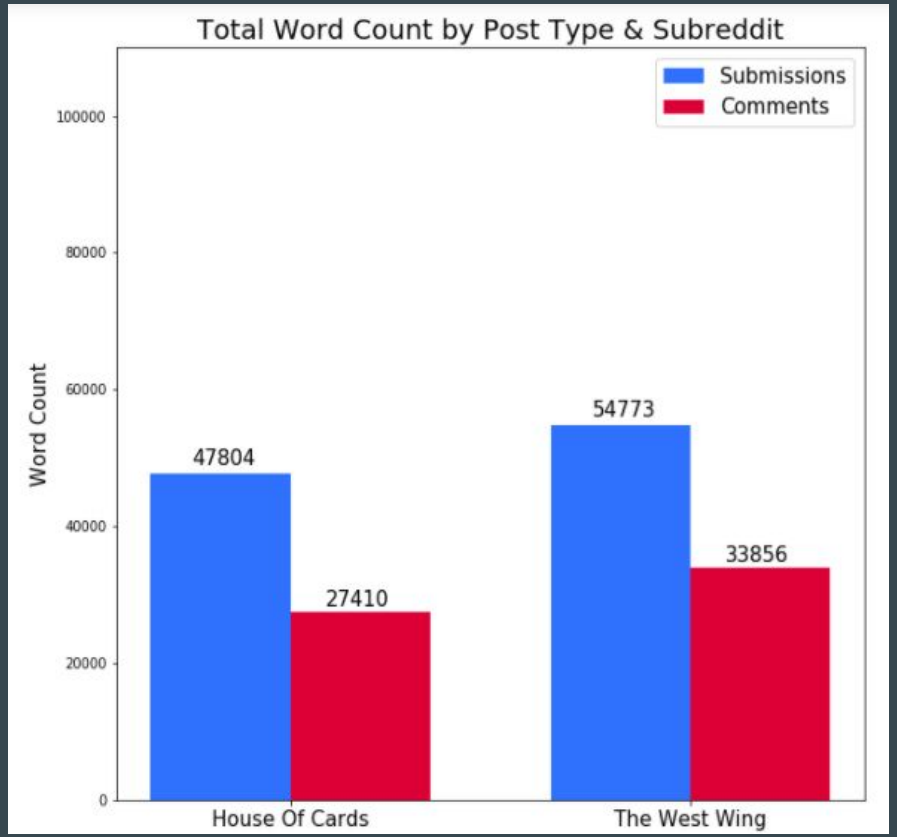
Preparing the Data

- Pushshift API
- Submission Posts? Comments? Why not both?
- 500-post limit (x16)
- Tokenize/Lemmatize/Stopwords

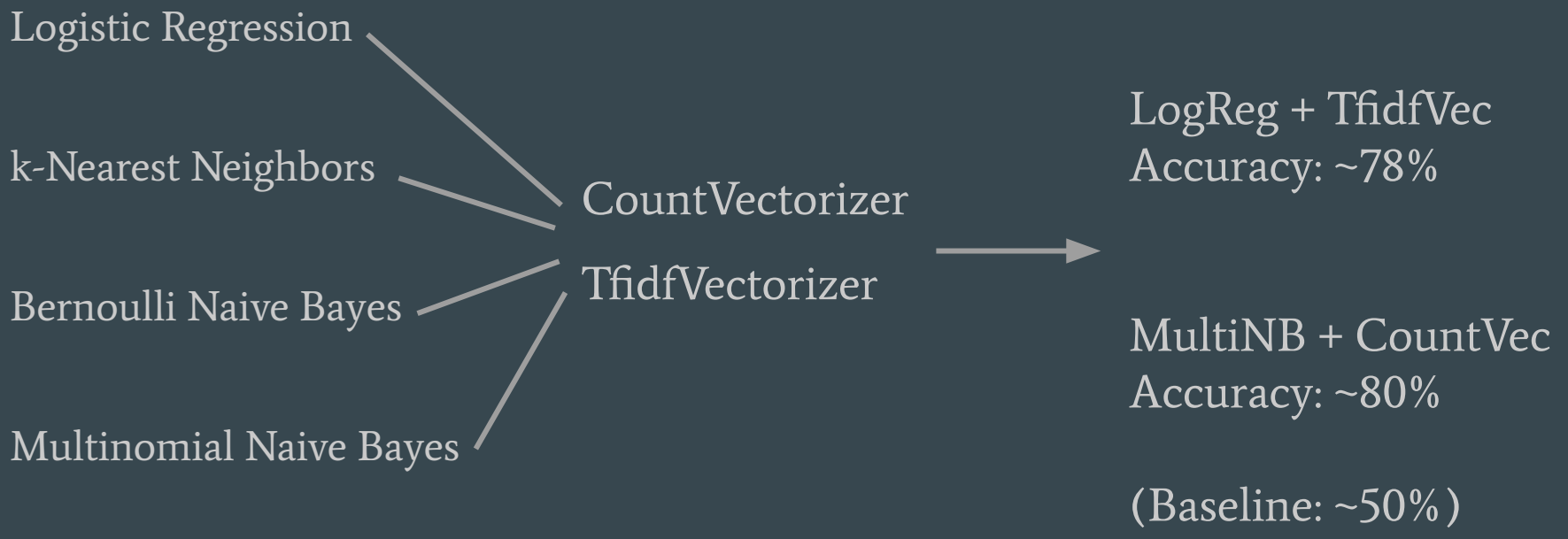
Total Word Count: 301,676

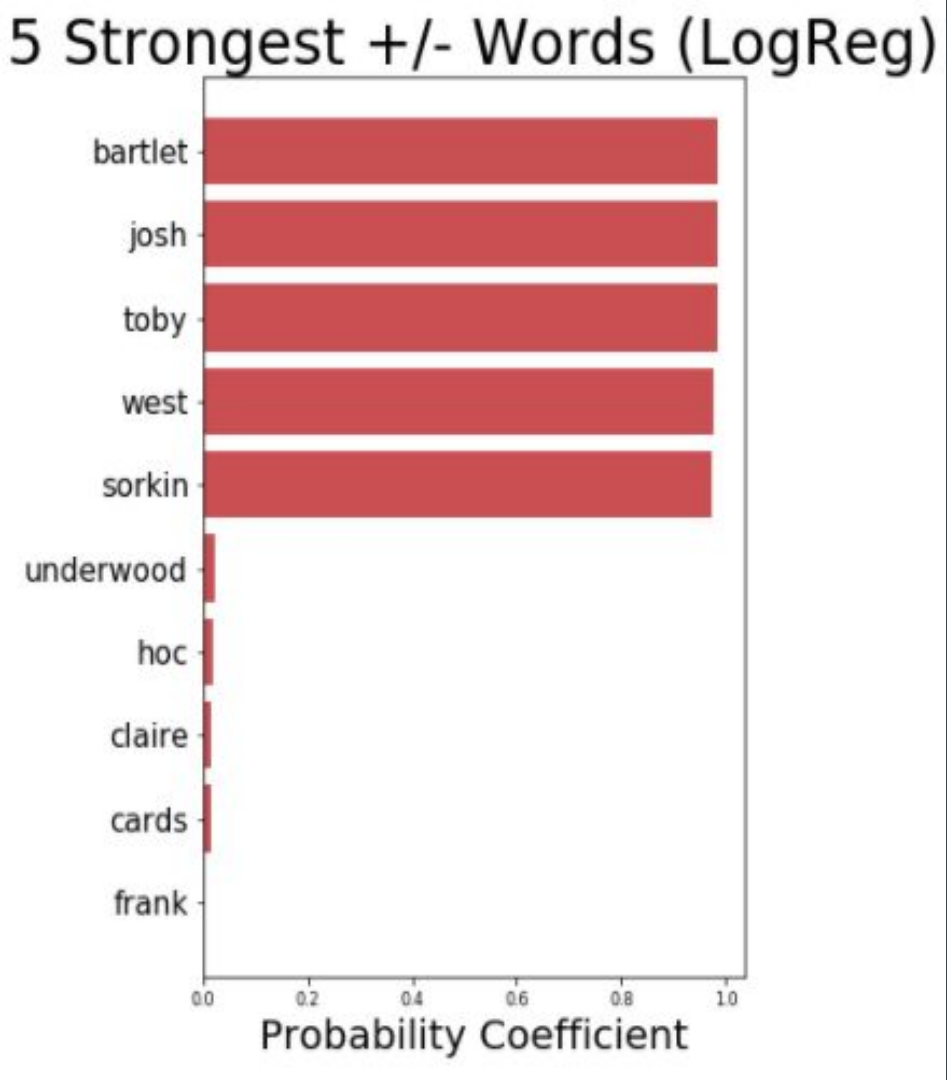


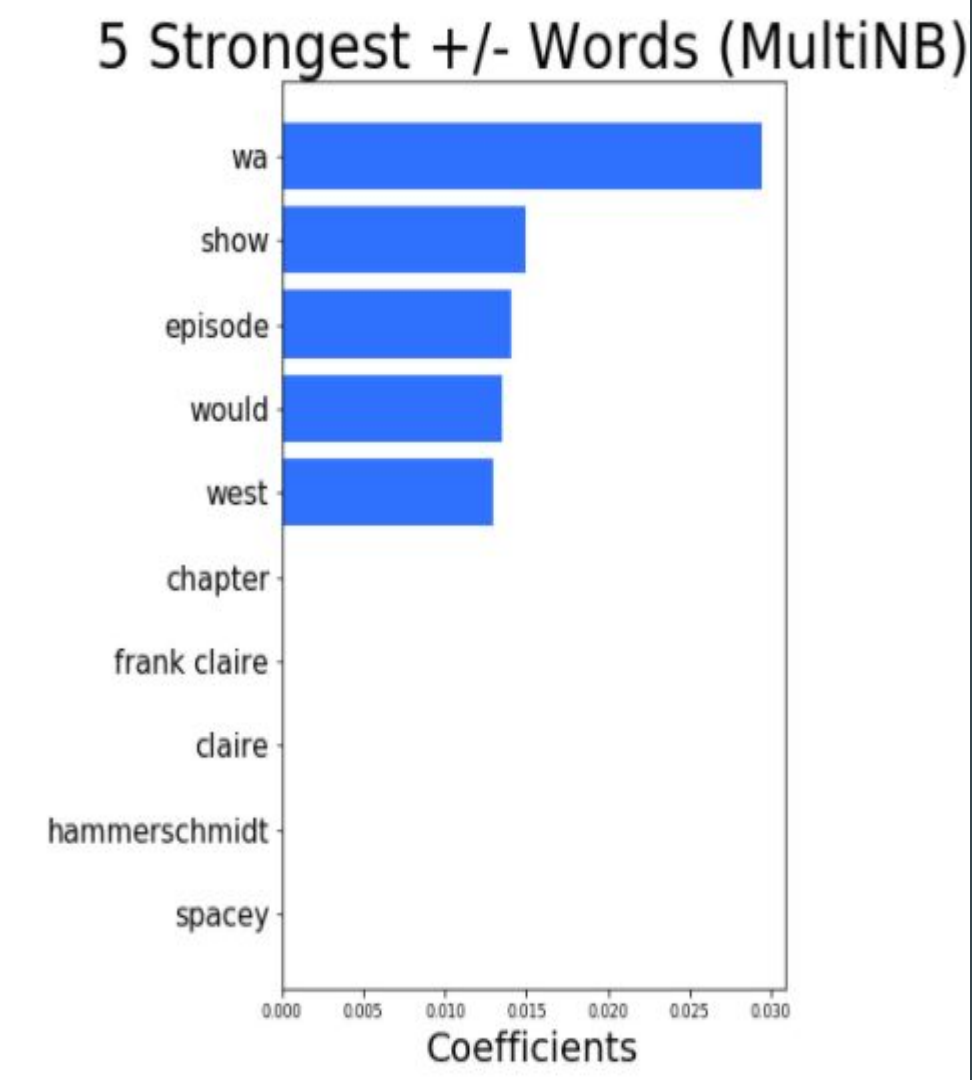
Total Word Count: 163,843



Which Models?




5 Strongest +/- Words (LogReg)
bartlet
josh
toby
west
sorkin
underwood
hoc
claire
cards
frank
0.0
0.2
0.4
0.6
0.8
1.0
Probability Coefficient


5 Strongest +/- Words (MultiNB)
wa
show
episode
would
west
chapter
frank claire
claire
hammerschmidt
spacey
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Coefficients

Drop in Accuracy Scores

With Show-Specific Words:

LogReg + TfidfVec
Accuracy: ~78%

MultiNB + CountVec
Accuracy: ~80%

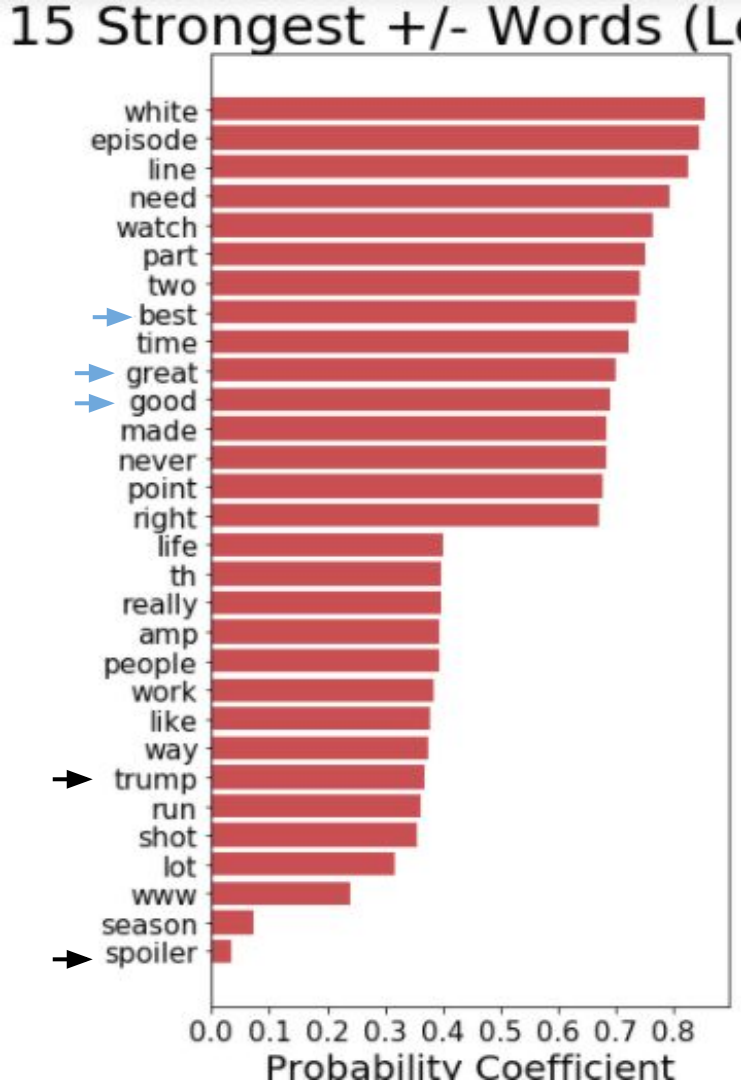
(Baseline: ~50%)

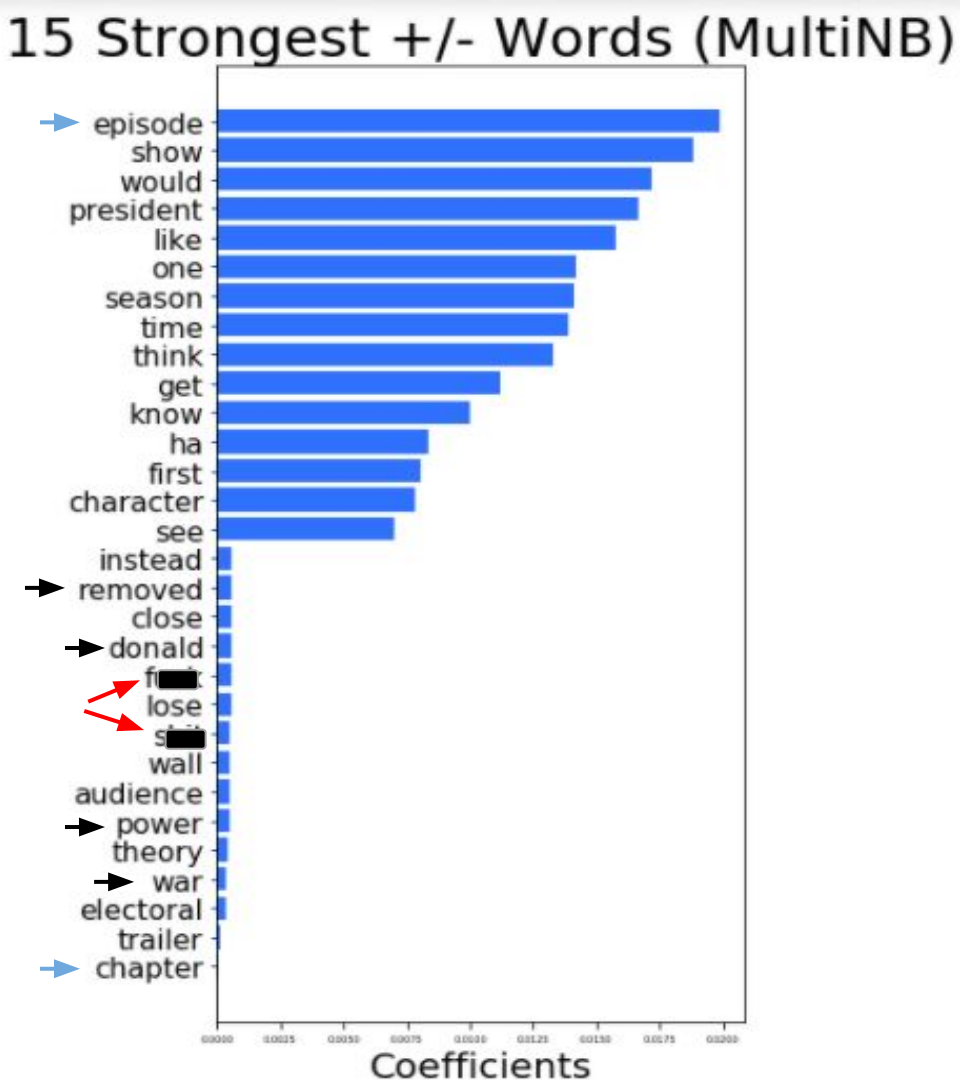
Without Show-Specific Words:

LogReg + TfidfVec
Accuracy: ~61%

MultiNB + CountVec
Accuracy: ~67%

(Baseline: ~50%)


15 Strongest +/- Words (LogReg)
white
episode
line
need
watch
part
two
best
time
great
good
made
never
point
right
life
th
really
amp
people
work
like
way
trump
run
shot
lot
www
season
spoiler
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
Probability Coefficient


15 Strongest +/- Words (MultiNB)
episode
show
would
president
like
one
season
time
think
get
know
ha
first
character
see
instead
removed
close
donald
lose
wall
audience
power
theory
war
electoral
trailer
chapter
Coefficients

Another Angle....

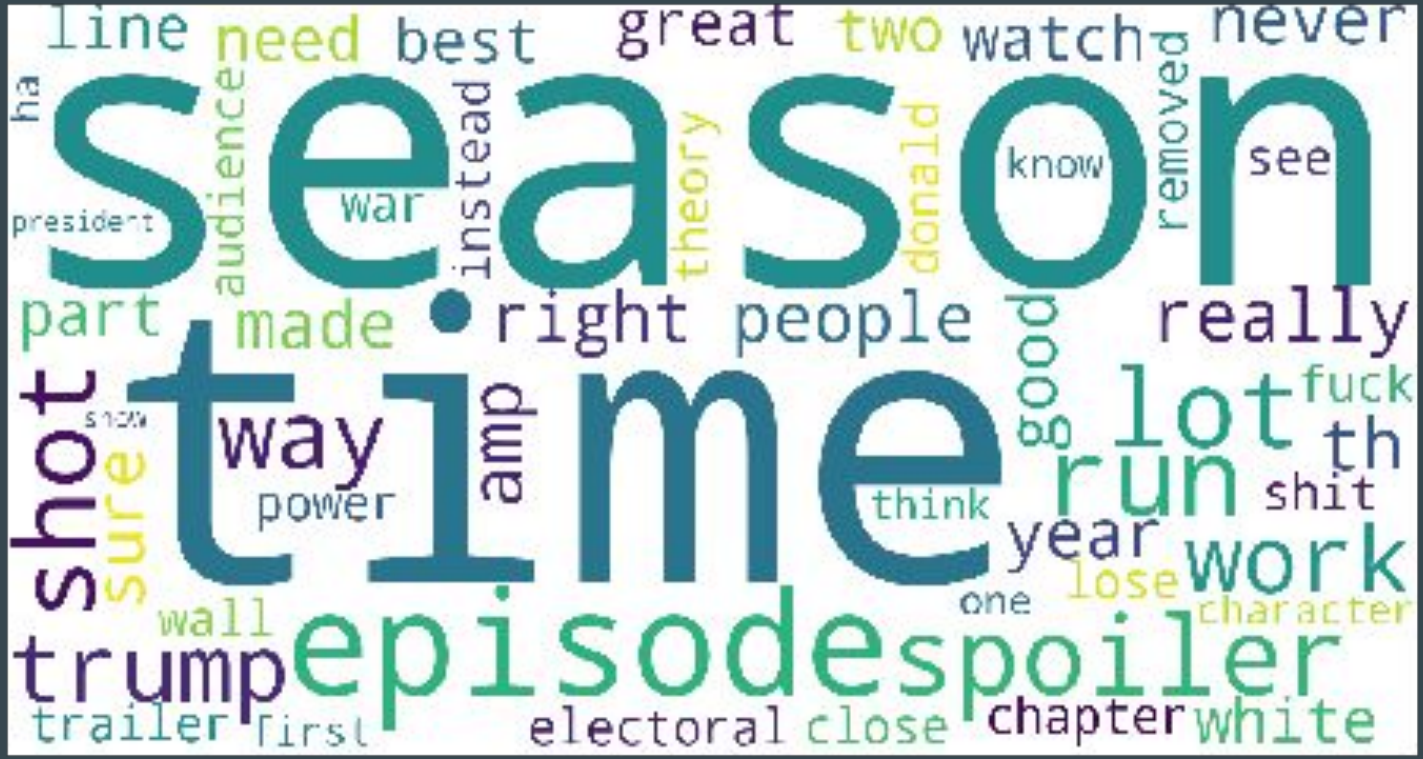
The Trump Factor

- Same Models
- Same target variable (broken up into pre-/post-Trump
- Pre-/post-Trump as target variable
- “Trump era” ill-defined

Next Steps

- Reexamine Trump factor in data
- Explore Other Current Events
- Submissions/Comments?
- Members/Activity gap
- Watch The West Wing!

THIS is Data Science



THIS is Data Science

Thank You

Q&A