What Sorkin Show Are You Talking About?

Classifying a Reddit post in to r/thewestwing
or r/thenewsroom

General Assembly - Data Science Immersive - Project 3

Marguerite Siboni

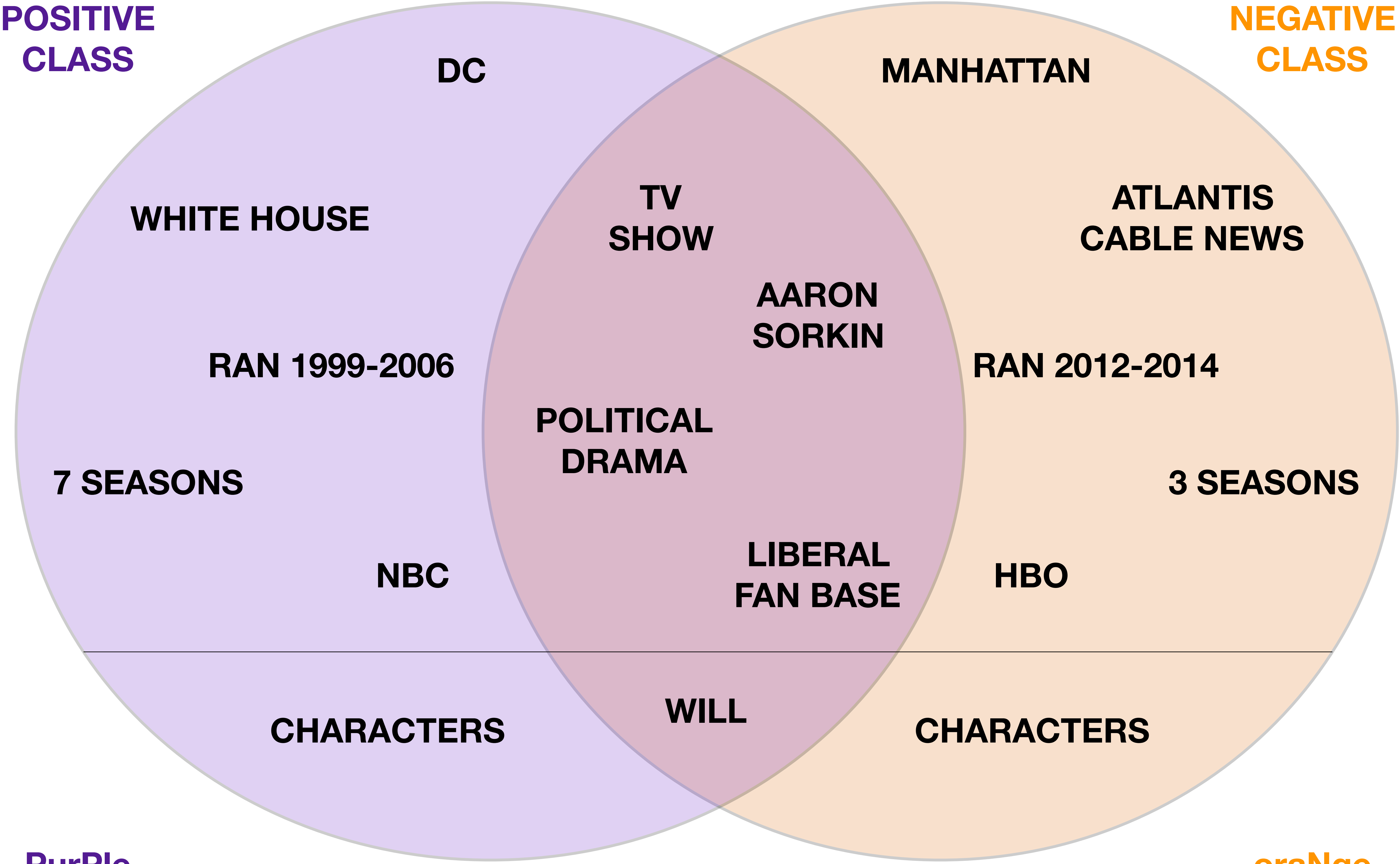
Overview

- Brief pitch of two of my favorite tv shows
- Scraping and Cleaning Data
- Logistic Regression Models
- Naive Bayes Models
- Next Steps

The West Wing vs *The Newsroom*

POSITIVE CLASS

NEGATIVE CLASS



DC

WHITE HOUSE

RAN 1999-2006

7 SEASONS

NBC

CHARACTERS

TV SHOW

AARON SORKIN

POLITICAL DRAMA

LIBERAL FAN BASE

WILL

MANHATTAN

ATLANTIS CABLE NEWS

RAN 2012-2014

3 SEASONS

HBO

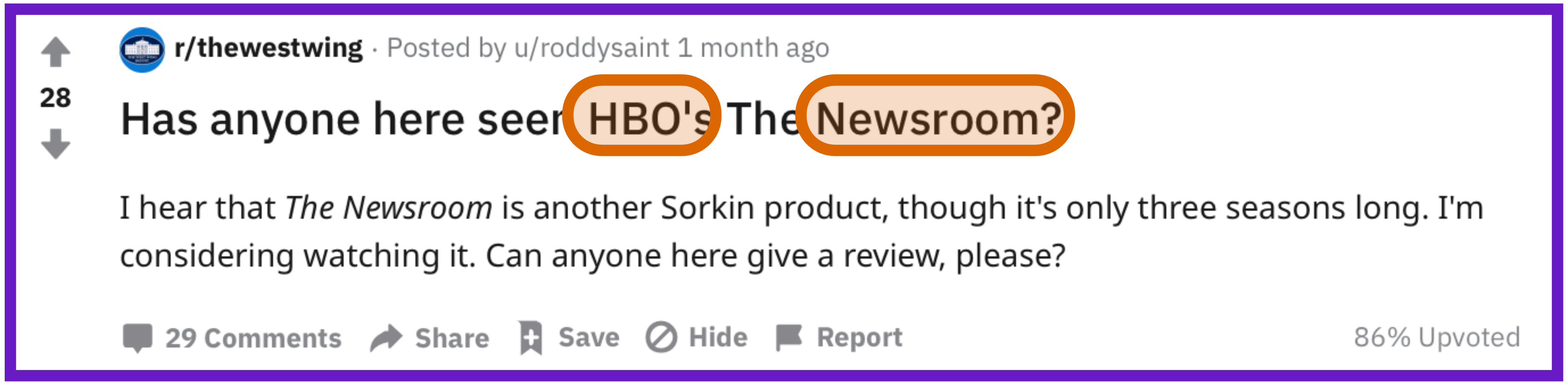
CHARACTERS

PurPle

oraNge

The West Wing vs The Newsroom

- Similar fan bases; parallel story lines

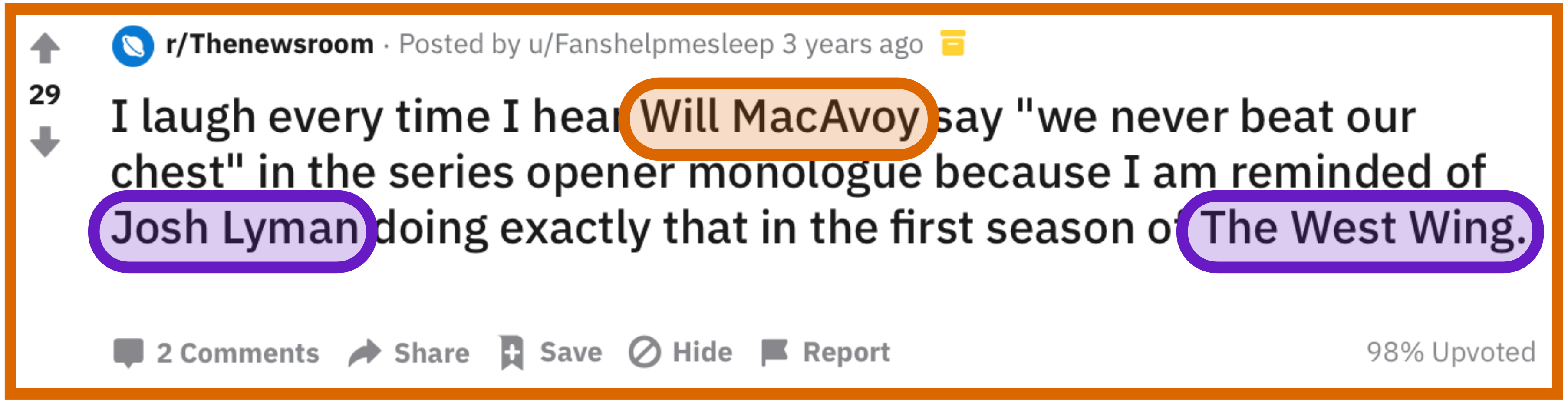


r/thewestwing · Posted by u/roddysaint 1 month ago

28

Has anyone here seen HBO's The Newsroom?

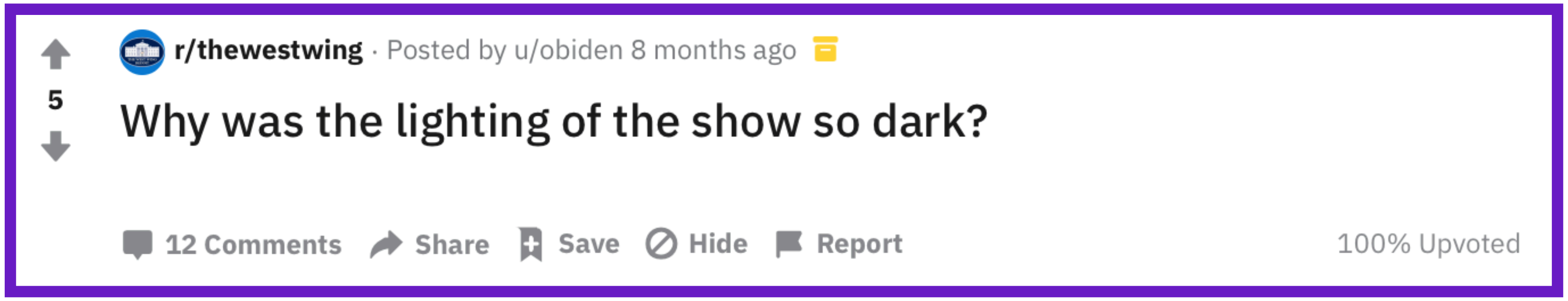
I hear that *The Newsroom* is another Sorkin product, though it's only three seasons long. I'm considering watching it. Can anyone here give a review, please?



r/Thenewsroom · Posted by u/Fanshelpmesleep 3 years ago

29

I laugh every time I hear Will MacAvoy say "we never beat our chest" in the series opener monologue because I am reminded of Josh Lyman doing exactly that in the first season of The West Wing.



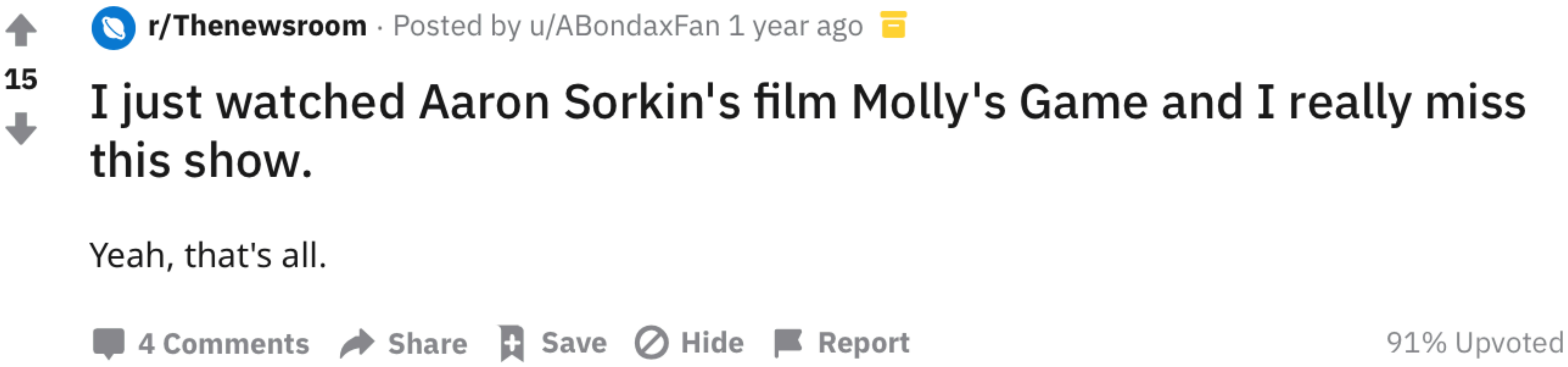
r/thewestwing · Posted by u/obiden 8 months ago

5

Why was the lighting of the show so dark?

Reddit Scraping: Step 0

- Post exists on Reddit



r/Thenewsroom · Posted by u/ABondaxFan 1 year ago

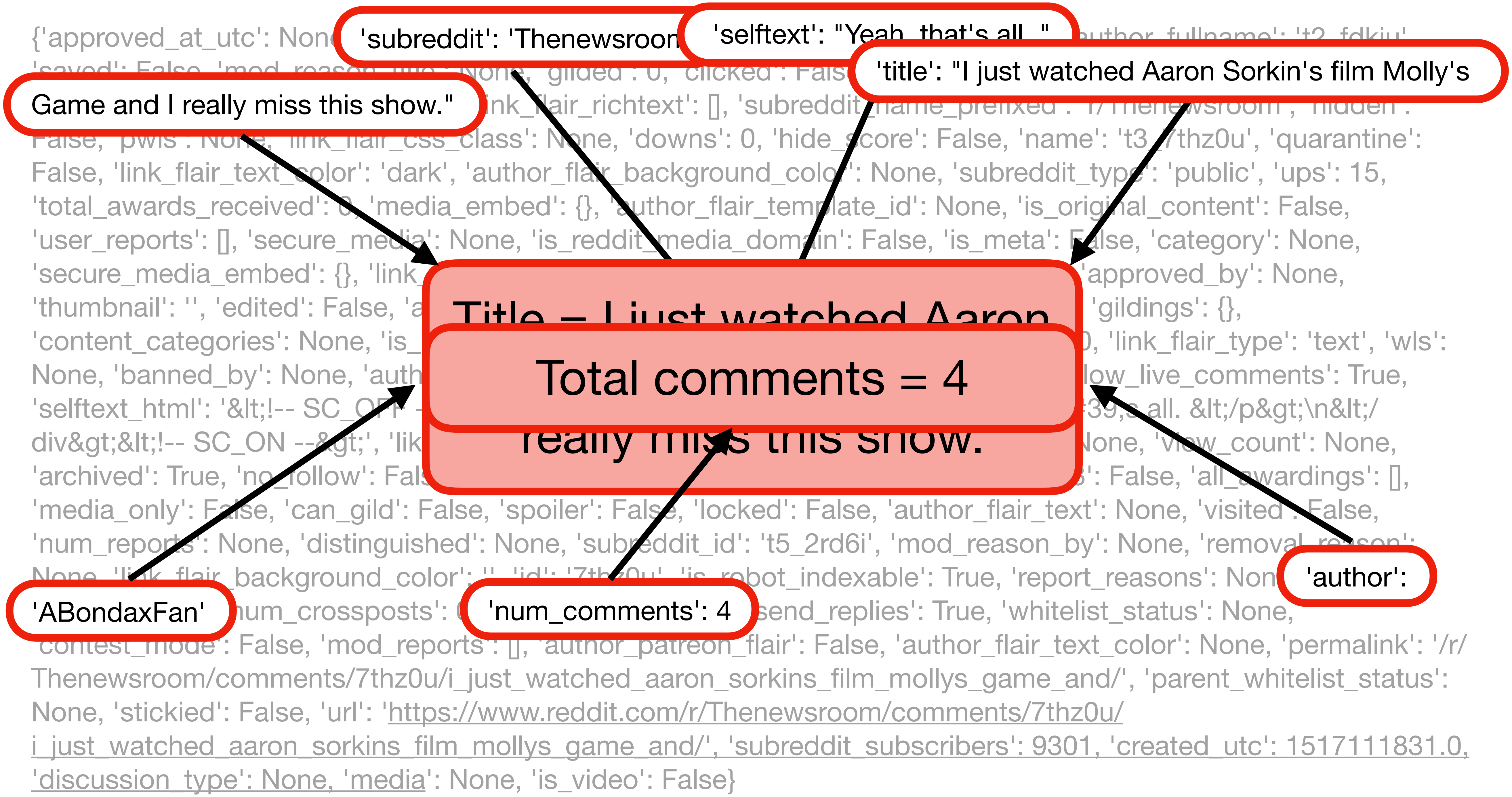
15

I just watched Aaron Sorkin's film Molly's Game and I really miss this show.

Yeah, that's all.

Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post



{'approved_at_utc': None author_fullname': 't2_fdkiu',
'saved': False, 'mod_reason_title': None, 'gilded': 0, 'clicked': Fals
ink_flair_richtext': [], 'subreddit_name_prefixed': 'r/Thenewsroom', 'hidden':
False, 'pwls': None, 'link_flair_css_class': None, 'downs': 0, 'hide_score': False, 'name': 't3_7thz0u', 'quarantine':
False, 'link_flair_text_color': 'dark', 'author_flair_background_color': None, 'subreddit_type': 'public', 'ups': 15,
'total_awards_received': 0, 'media_embed': {}, 'author_flair_template_id': None, 'is_original_content': False,
'user_reports': [], 'secure_media': None, 'is_reddit_media_domain': False, 'is_meta': False, 'category': None,
'secure_media_embed': {}, 'link_ 'approved_by': None,
'thumbnail': '', 'edited': False, 'a 'gildings': {},
'content_categories': None, 'is_ 0, 'link_flair_type': 'text', 'wls':
None, 'banned_by': None, 'auth low_live_comments': True,
'selftext_html': '&lt;!-- SC_OFF - #39;s all. &lt;/p&gt;\n&lt;/
div&gt;&lt;!-- SC_ON --&gt;', 'lik None, 'view_count': None,
'archived': True, 'no_follow': Fals ': False, 'all_awardings': [],
'media_only': False, 'can_gild': False, 'spoiler': False, 'locked': False, 'author_flair_text': None, 'visited': False,
'num_reports': None, 'distinguished': None, 'subreddit_id': 't5_2rd6i', 'mod_reason_by': None, 'removal_reason':
None, 'link_flair_background_color': '', 'id': '7thz0u', 'is_robot_indexable': True, 'report_reasons': Non
num_crossposts': 0 send_replies': True, 'whitelist_status': None,
'contest_mode': False, 'mod_reports': [], 'author_patreon_flair': False, 'author_flair_text_color': None, 'permalink': '/r/
Thenewsroom/comments/7thz0u/i_just_watched_aaron_sorkins_film_mollys_game_and/', 'parent_whitelist_status':
None, 'stickied': False, 'url': 'https://www.reddit.com/r/Thenewsroom/comments/7thz0u/
i_just_watched_aaron_sorkins_film_mollys_game_and/', 'subreddit_subscribers': 9301, 'created_utc': 1517111831.0,
'discussion_type': None, 'media': None, 'is_video': False}

'subreddit': 'Thenewsroom

'selftext': "Yeah, that's all. "

'title': "I just watched Aaron Sorkin's film Molly's

Game and I really miss this show."

Title = I just watched Aaron

really miss this show.

Total comments = 4

'author':

'ABondaxFan'

'num_comments': 4

Reddit Scraping: Step 2

- Extract title + text from posts

```
{'approved_at_utc': None, 'subreddit': 'Thenewsroom', 'selftext': "Yeah, that's all. ",
'author_fullname': 't2_fdkiu', 'saved': False, 'mod_reason_title': None, 'gilded': 0,
'clicked': False, 'title': "I just watched Aaron Sorkin's film Molly's Game and I really miss
this show.", 'link_flair_richtext': [], 'subreddit_name_prefixed': 'r/Thenewsroom',
'hidden': False, 'pwls': None, 'link_flair_css_class': None, 'downs': 0, 'hide_score':
False, 'name': 't3_7thz0u', 'quarantine': False, 'link_flair_text_color': 'dark',
'author_flair_background_color': None, 'subreddit_type': 'public', 'ups': 15,
'total_awards_received': 0, 'media_embed': {}, 'author_flair_template_id': None,
'is_original_content': False, 'user_reports': [], 'secure_media': None,
'is_reddit_media_domain': False, 'is_meta': False, 'category': None,
'secure_media_embed': {}, 'link_flair_text': None, 'can_mod_post': False, 'score': 15,
'approved_by': None, 'thumbnail': '', 'edited': False, 'author_flair_css_class': None,
'author_flair_richtext': [], 'gildings': {}, 'content_categories': None, 'is_self': True,
'mod_note': None, 'created': 1517140631.0, 'link_flair_type': 'text', 'wls': None,
'banned_by': None, 'author_flair_type': 'text', 'domain': 'self.Thenewsroom',
'allow_live_comments': True, 'selftext_html': '&lt;!-- SC_OFF --&gt;&lt;div
class="md"&gt;&lt;p&gt;Yeah, that&amp;#39;s all. &lt;/p&gt;\n&lt;/div&gt;&lt;!-- SC_ON
--&gt;', 'likes': None, 'suggested_sort': None, 'banned_at_utc': None, 'view_count':
None, 'archived': True, 'no_follow': False, 'is_crosspostable': False, 'pinned': False,
'over_18': False, 'all_awardings': [], 'media_only': False, 'can_gild': False, 'spoiler':
False, 'locked': False, 'author_flair_text': None, 'visited': False, 'num_reports': None,
'distinguished': None, 'subreddit_id': 't5_2rd6i', 'mod_reason_by': None,
'removal_reason': None, 'link_flair_background_color': '', 'id': '7thz0u',
'is_robot_indexable': True, 'report_reasons': None, 'author': 'ABondaxFan',
'num_crossposts': 0, 'num_comments': 4, 'send_replies': True, 'whitelist_status': None,
'contest_mode': False, 'mod_reports': [], 'author_patreon_flair': False,
'author_flair_text_color': None, 'permalink': '/r/Thenewsroom/comments/7thz0u/
i_just_watched_aaron_sorkins_film_mollys_game_and/', 'parent_whitelist_status': None,
'stickied': False, 'url': 'https://www.reddit.com/r/Thenewsroom/comments/7thz0u/
i_just_watched_aaron_sorkins_film_mollys_game_and/', 'subreddit_subscribers': 9301,
'created_utc': 1517111831.0, 'discussion_type': None, 'media': None, 'is_video': False}
```

```
function clean_posts(list_of_posts):

  created empty list clean_posts

  for post in list_of_posts:

    pull title
    pull text
    add title + text to clean_posts

  return clean_posts
```

I just watched Aaron Sorkin's film Molly's Game and I really miss this show. Yeah, that's all.

Reddit Scraping: Step 2

```
{'approved_at_utc': None, 'subreddit': 'Thenewsroom', 'selftext': "Yeah, that's all. ",
'author_fullname': 't2_fdkiu', 'saved': False, 'mod_reason_title': None, 'gilded': 0,
'clicked': False, 'title': "I just watched Aaron Sorkin's film Molly's Game and I really miss
this show.", 'link_flair_richtext': [], 'subreddit_name_prefixed': 'r/Thenewsroom',
'hidden': False, 'pwls': None, 'link_flair_css_class': None, 'downs': 0, 'hide_score':
False, 'name': 't3_7thz0u', 'quarantine': False, 'link_flair_text_color': 'dark',
'author_flair_background_color': None, 'subreddit_type': 'public', 'ups': 15,
'total_awards_received': 0, 'media_embed': {}, 'author_flair_template_id': None,
'is_original_content': False, 'user_reports': [], 'secure_media': None,
'is_reddit_media_domain': False, 'is_meta': False, 'category': None,
'secure_media_embed': {}, 'link_flair_text': None, 'can_mod_post': False, 'score': 15,
'approved_by': None, 'thumbnail': '', 'edited': False, 'author_flair_css_class': None,
'author_flair_richtext': [], 'gildings': {}, 'content_categories': None, 'is_self': True,
'mod_note': None, 'created': 1517140631.0, 'link_flair_type': 'text', 'wls': None,
'banned_by': None, 'author_flair_type': 'text', 'domain': 'self.Thenewsroom',
'allow_live_comments': True, 'selftext_html': '&lt;!-- SC_OFF --&gt;&lt;div
class="md"&gt;&lt;p&gt;Yeah, that&amp;#39;s all. &lt;/p&gt;\n&lt;/div&gt;&lt;!-- SC_ON
--&gt;', 'likes': None, 'suggested_sort': None, 'banned_at_utc': None, 'view_count':
None, 'archived': True, 'no_follow': False, 'is_crosspostable': False, 'pinned': False,
'over_18': False, 'all_awardings': [], 'media_only': False, 'can_gild': False, 'spoiler':
False, 'locked': False, 'author_flair_text': None, 'visited': False, 'num_reports': None,
'distinguished': None, 'subreddit_id': 't5_2rd6i', 'mod_reason_by': None,
'removal_reason': None, 'link_flair_background_color': '', 'id': '7thz0u',
'is_robot_indexable': True, 'report_reasons': None, 'author': 'ABondaxFan',
'num_crossposts': 0, 'num_comments': 4, 'send_replies': True, 'whitelist_status': None,
'contest_mode': False, 'mod_reports': [], 'author_patreon_flair': False,
'author_flair_text_color': None, 'permalink': '/r/Thenewsroom/comments/7thz0u/
i_just_watched_aaron_sorkins_film_mollys_game_and/', 'parent_whitelist_status': None,
'stickied': False, 'url': 'https://www.reddit.com/r/Thenewsroom/comments/7thz0u/
i_just_watched_aaron_sorkins_film_mollys_game_and/', 'subreddit_subscribers': 9301,
'created_utc': 1517111831.0, 'discussion_type': None, 'media': None, 'is_video': False}
```

```
function clean_posts(list_of_posts):

  created empty list clean_posts

  for post in list_of_posts:

    pull title
    pull text
    add title + text to clean_posts

  return clean_posts
```

I just watched Aaron Sorkin's film Molly's Game and I really miss this show. Yeah, that's all.

Tokenizing, Lemmatizing and Stemming

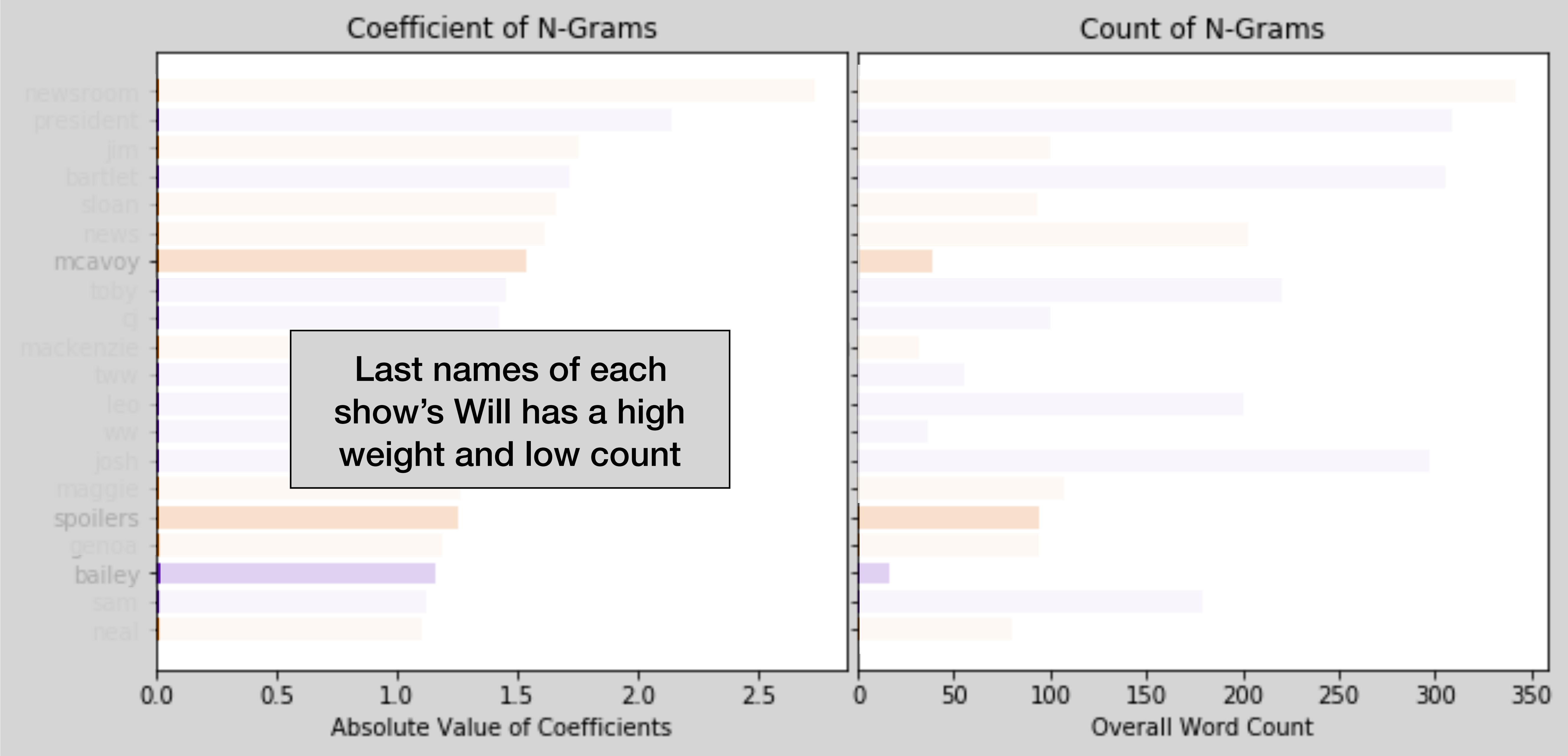
- Created function to tokenize, lemmatize and stem text.
- Kai informed me that CountVectorizer has a built in lemmatizer

Logistic Regression Models - Hyperparameters

		AS IS, STOP WORDS	LEMMATIZED, STOP WORDS	STEMMED. STOP WORDS
COUNT VECTORIZED	MAX DF	0.87	0.15	0.23
	MIN DF	3	2	2
	MAX FEATURES	2900	3000	4100
	TEST ACCURACY	86.5%	85.5%	87.5%
TFIDF VECTORIZED	MAX DF	0.3	0.25	0.29
	MIN DF	2	2	3
	MAX FEATURES	3000	3100	2950
	TEST ACCURACY	87.5%	87.5%	87.5%

Logistic Regression’s Most Predictive N-Grams

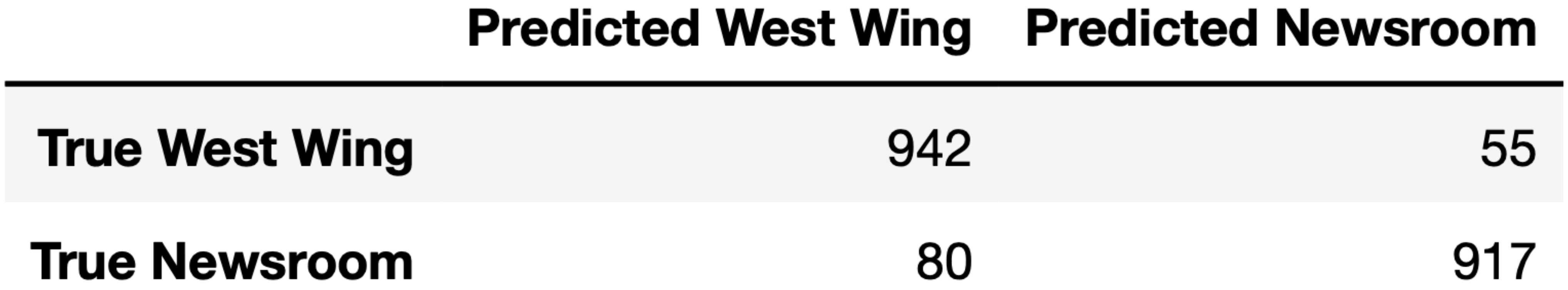
- Most predictive N-grams are all words (1, 1) on a (1, 2) model
- Mostly show names and characters
- “Spoilers” is the only generic tv term, indicates newer show



Last names of each show’s Will has a high weight and low count

Naive Bayes Model

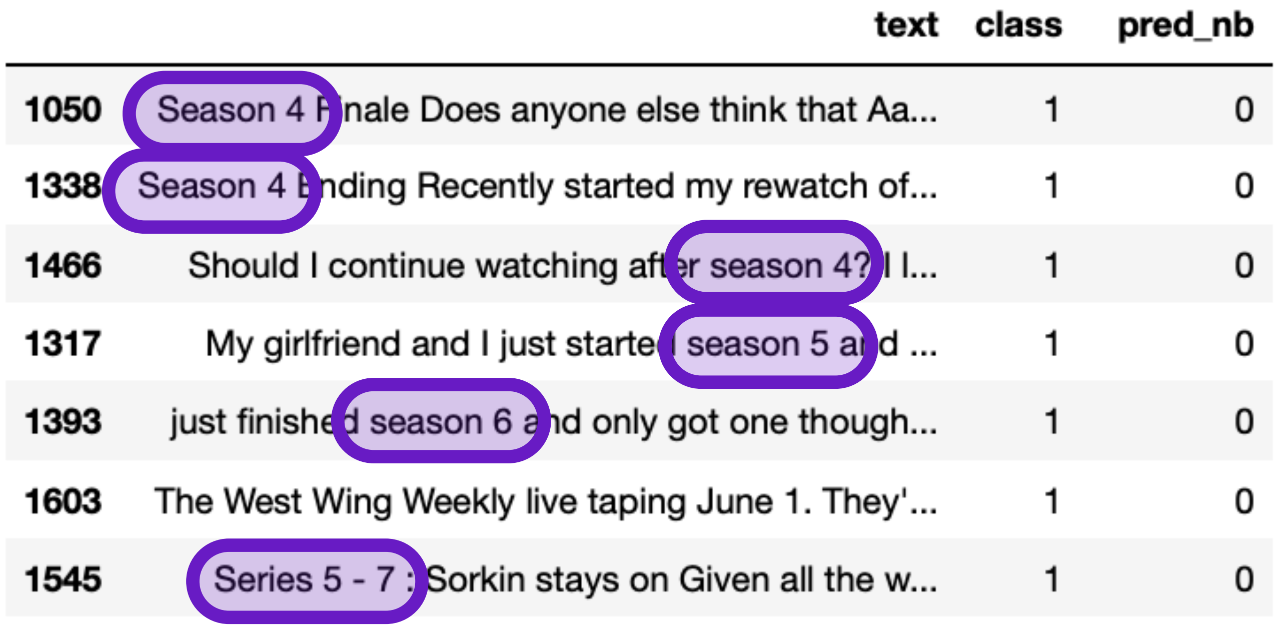
- Best Train Score lowered from 98.5% accuracy to 93.7%
- Best Test Score increased from 87.5% accuracy to 88.5%
- Appears to have made my model less overfit



	Predicted West Wing	Predicted Newsroom
True West Wing	942	55
True Newsroom	80	917

Improving the Model

- “Season 4” and onward appear in several posts.



	text	class	pred_nb
1050	Season 4 Finale Does anyone else think that Aa...	1	0
1338	Season 4 Ending Recently started my rewatch of...	1	0
1466	Should I continue watching after season 4? I l...	1	0
1317	My girlfriend and I just started season 5 and ...	1	0
1393	just finished season 6 and only got one though...	1	0
1603	The West Wing Weekly live taping June 1. They'...	1	0
1545	Series 5 - 7 : Sorkin stays on Given all the w...	1	0

- Feed in all character names
- Feed in popular quotes
- Date of post
- User

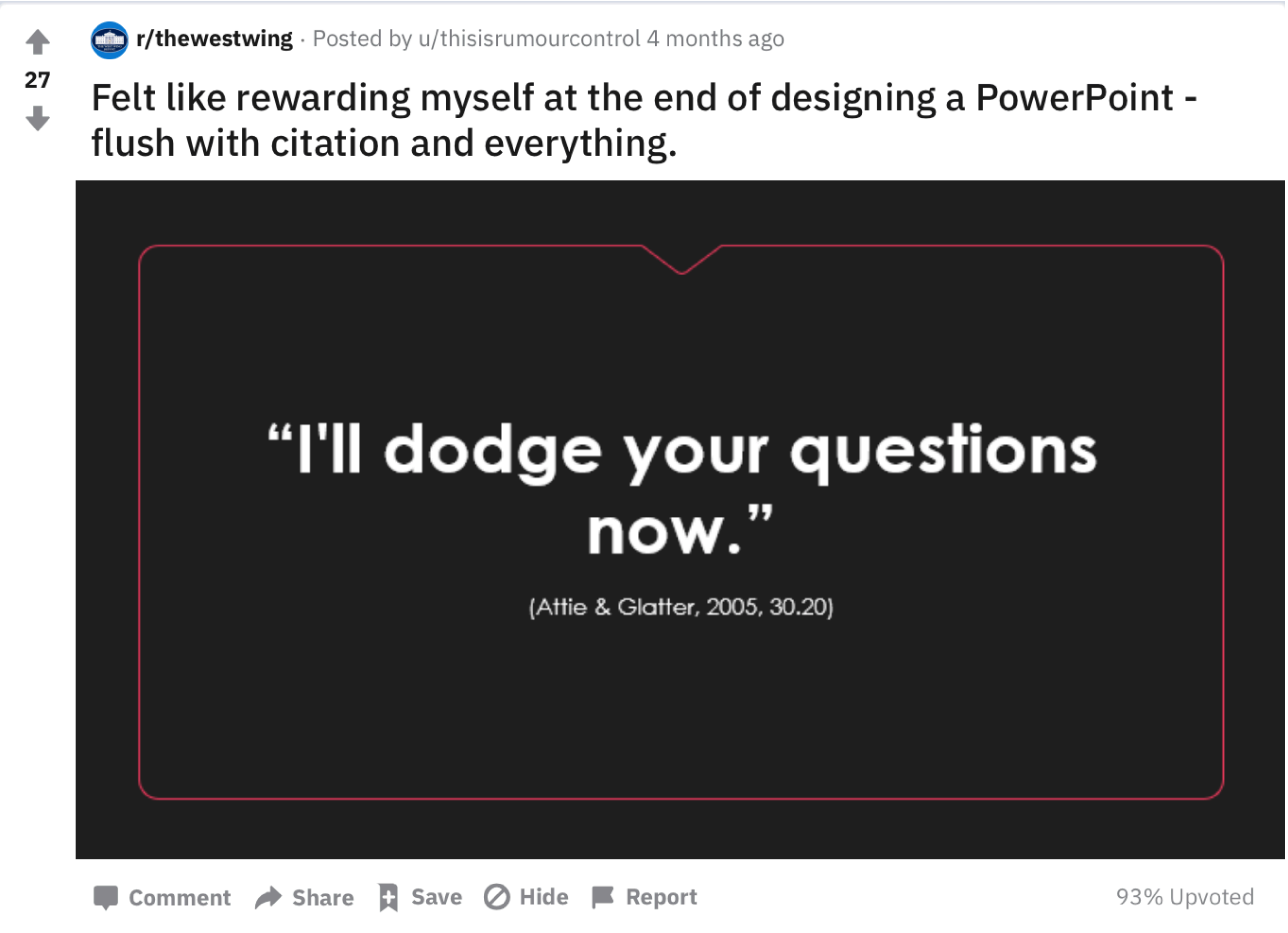
Expanding the Model to More Subreddits

- r/SportsNight - A Sorkin show that is NOT political
- r/Scandal - A political drama that is NOT written by Sorkin
- r/politics - A discussion of real life politics NOT tv politics

- A political movie by Sorkin would also be interesting. However, none of his movies have unique subreddits

Conclusion

- The West Wing and The Newsroom are both great tv shows. You should watch them.
- Use built in lemmatizer within Vectorize functions
- Most predictive words when classifying were characters and show names
- Naive Bayes Classification Model of Count Vectorized Data most accurately predicted test data
- Potential to add additional model rules and/or additional data to improve model



r/thewestwing · Posted by u/thisisrumourcontrol 4 months ago

Felt like rewarding myself at the end of designing a PowerPoint - flush with citation and everything.

"I'll dodge your questions now."

(Attie & Glatter, 2005, 30.20)