

# What Sorkin Show Are You Talking About?

---

Classifying a Reddit post in to [r/thewestwing](#)  
or [r/thenewsroom](#)

# Overview

---

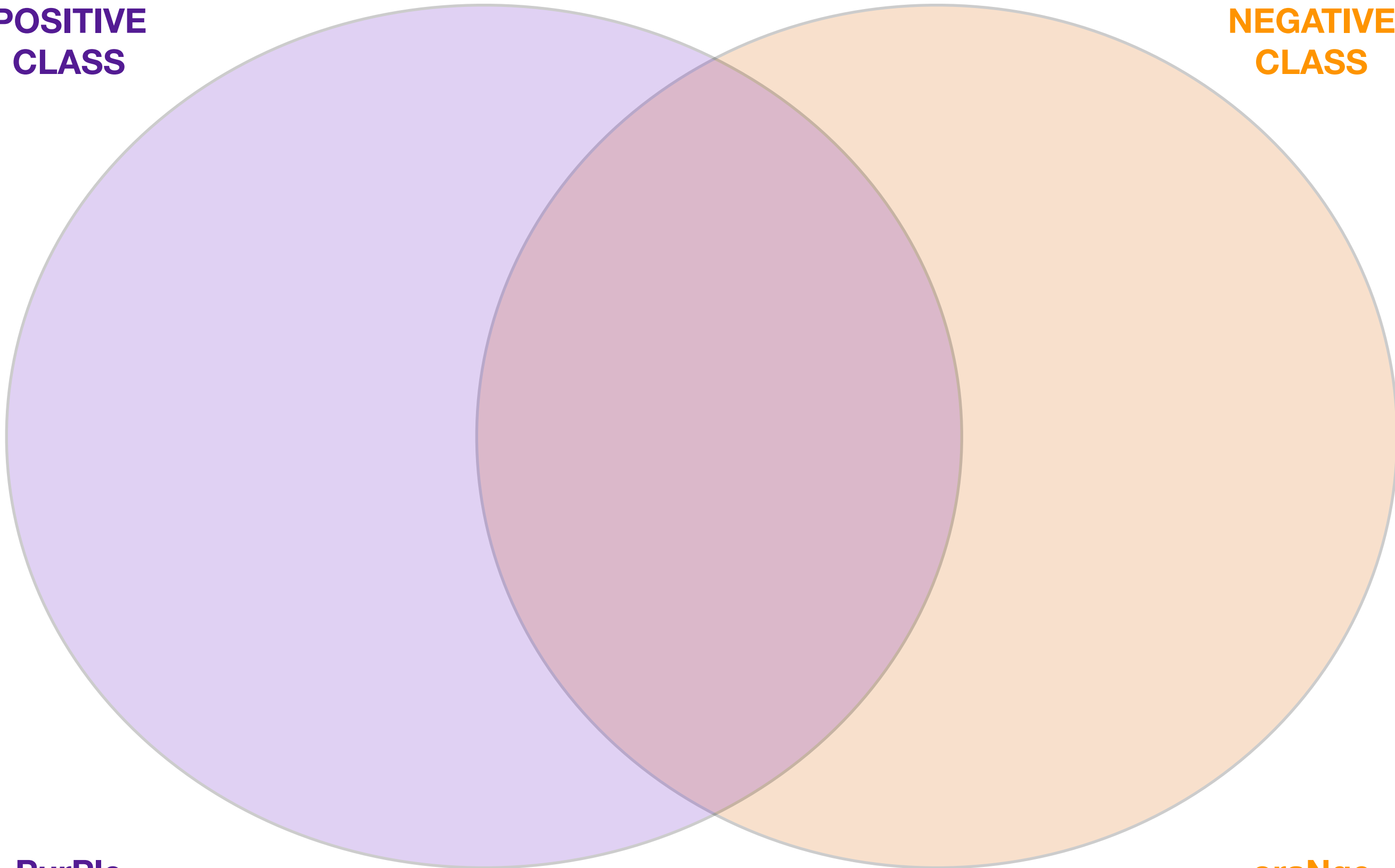
- Brief pitch of two of my favorite tv shows
- Scraping and Cleaning Data
- Logistic Regression Models
- Naive Bayes Models
- Next Steps

# ***The West Wing*** vs ***The Newsroom***

---

**POSITIVE  
CLASS**

**NEGATIVE  
CLASS**



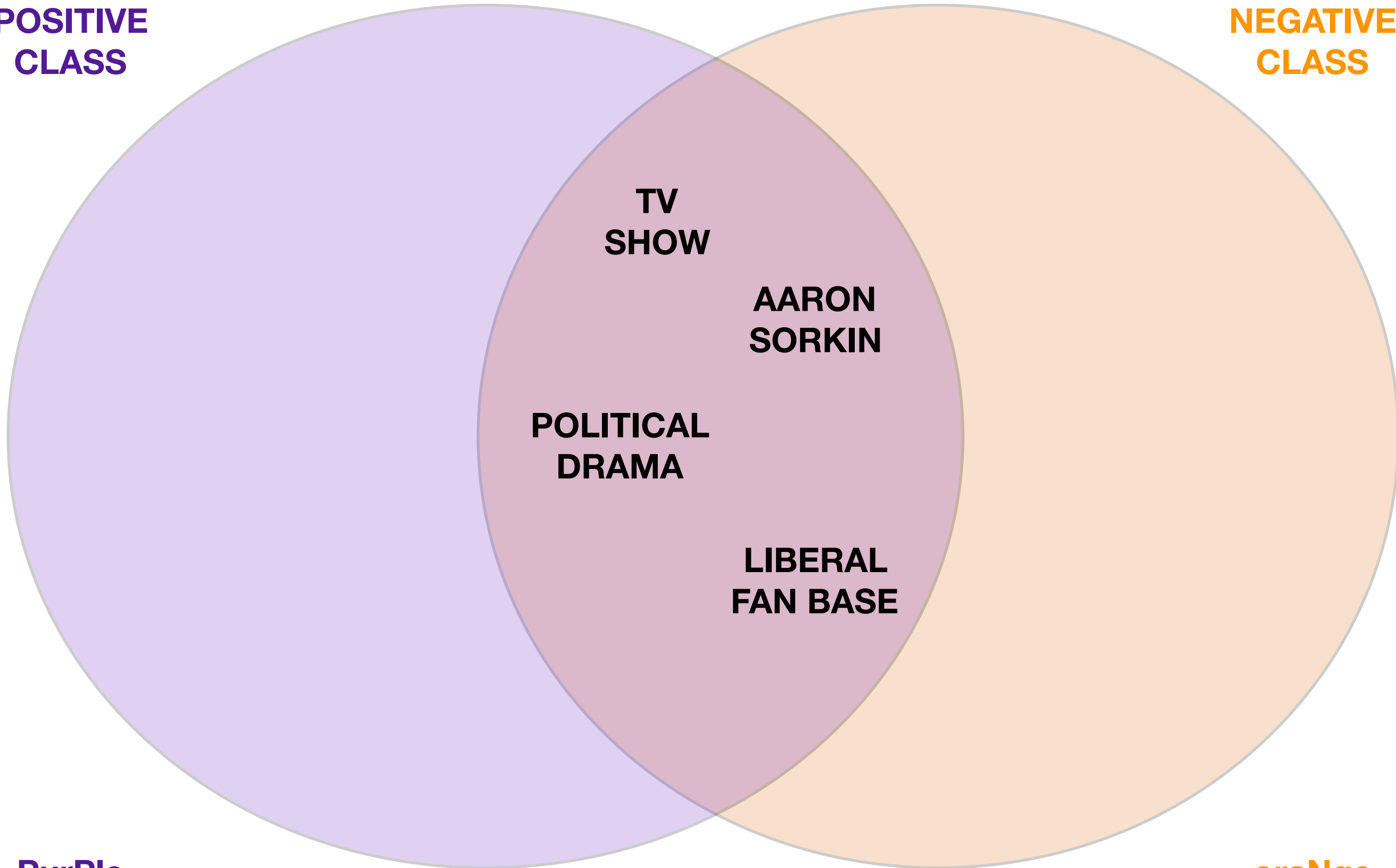
**PurPle**

**oraNge**

# ***The West Wing*** vs ***The Newsroom***

**POSITIVE  
CLASS**

**NEGATIVE  
CLASS**



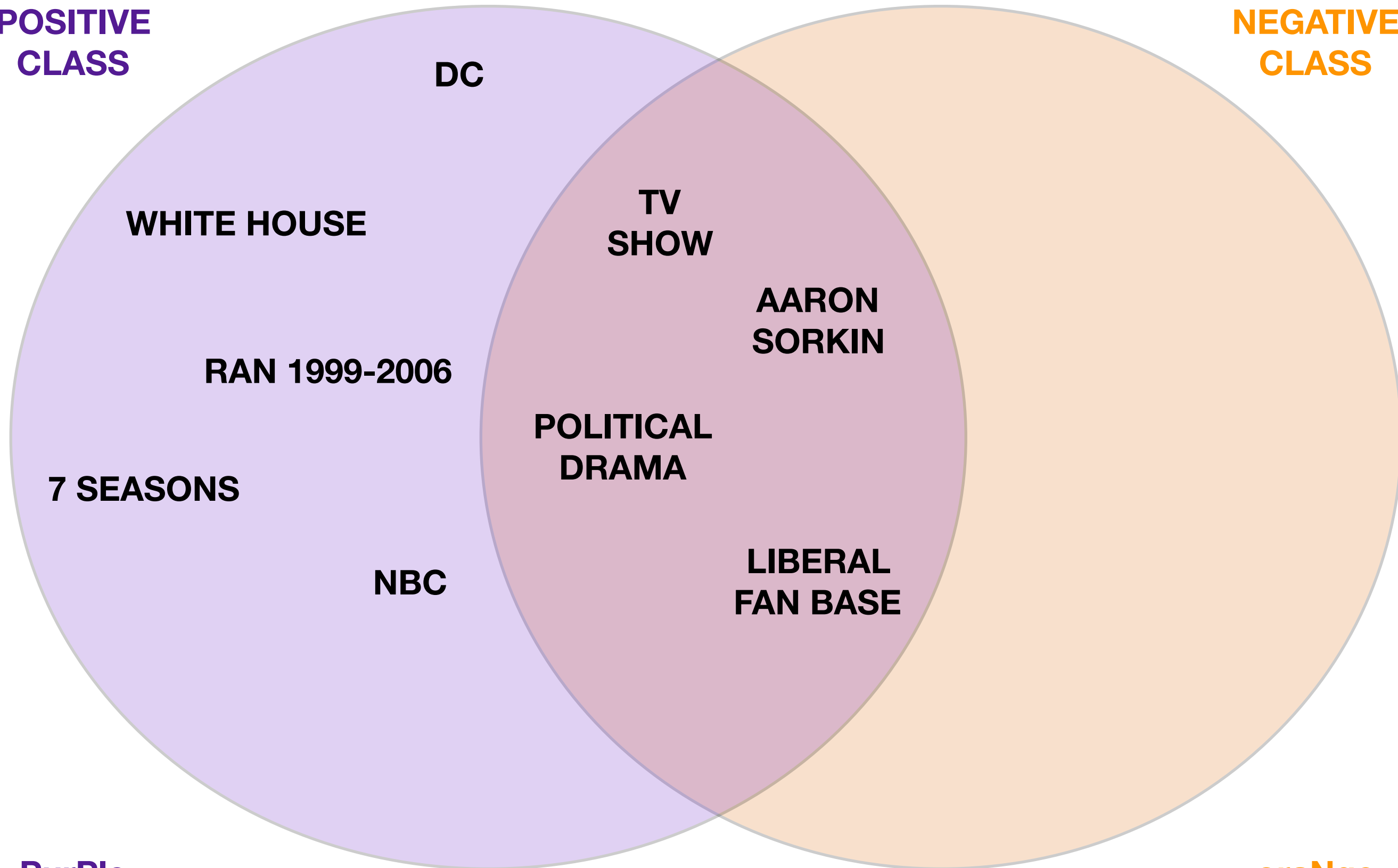
**PurPle**

**oraNge**

# ***The West Wing*** vs ***The Newsroom***

**POSITIVE  
CLASS**

**NEGATIVE  
CLASS**



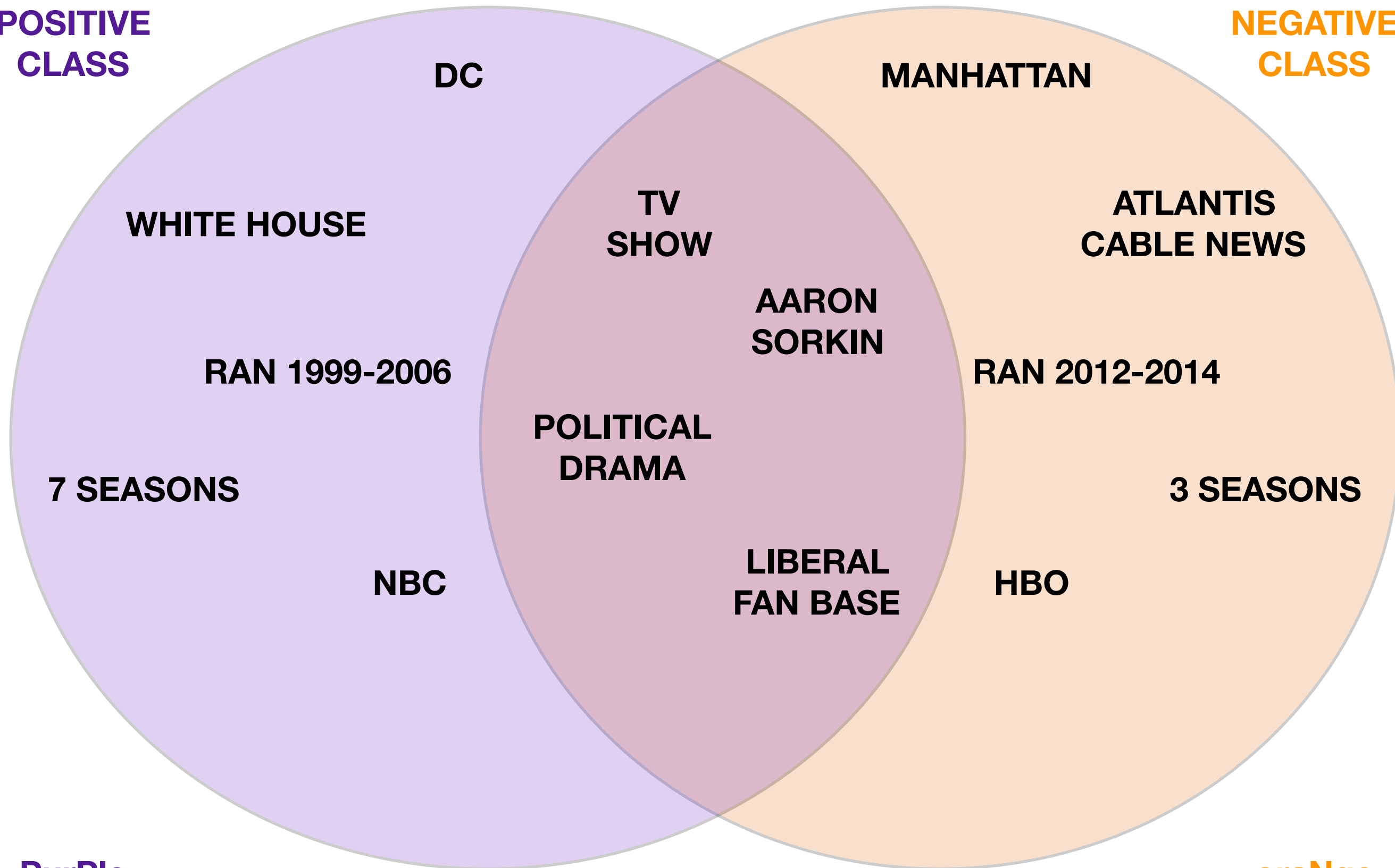
**PurPle**

**oraNge**

# ***The West Wing*** vs ***The Newsroom***

**POSITIVE  
CLASS**

**NEGATIVE  
CLASS**



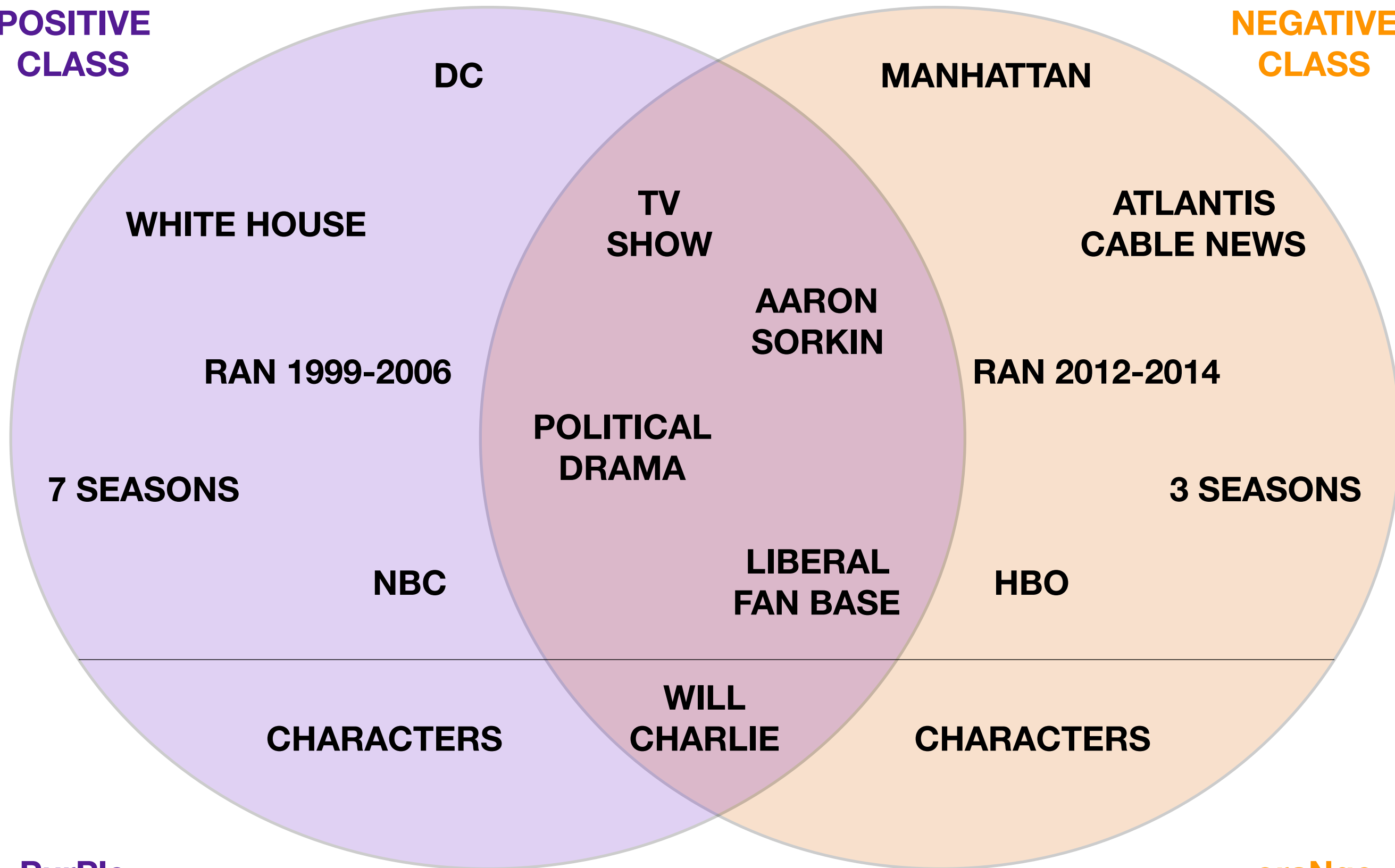
**PurPle**

**oraNge**

# ***The West Wing*** vs ***The Newsroom***

**POSITIVE  
CLASS**

**NEGATIVE  
CLASS**



**PurPle**

**oraNge**

# ***The West Wing*** vs ***The Newsroom***

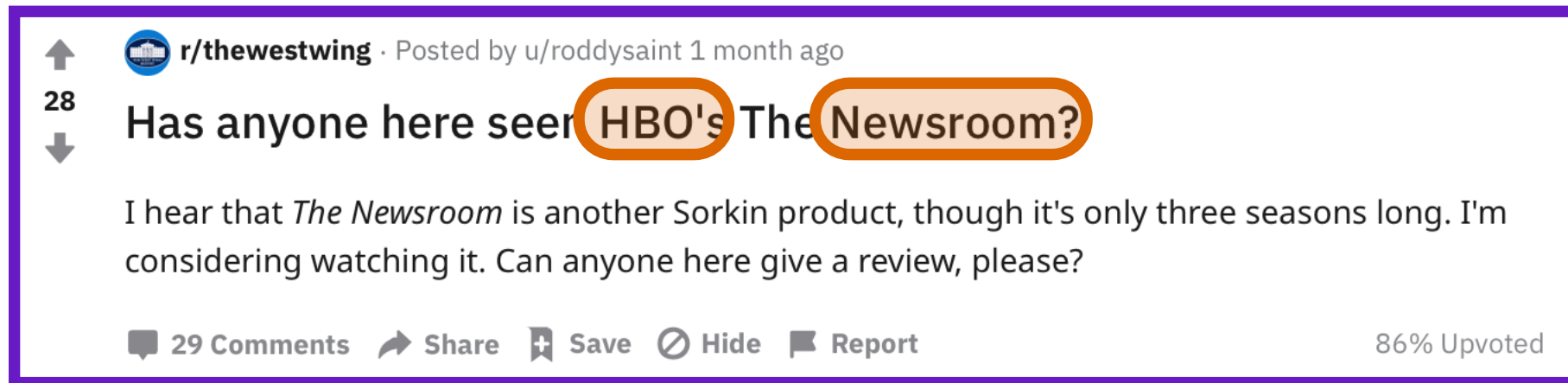
---

- Similar fan bases; parallel story lines





# *The West Wing* vs *The Newsroom*


- Similar fan bases; parallel story lines



# *The West Wing* vs *The Newsroom*





- Similar fan bases; parallel story lines



 28 



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

**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?

 29 Comments  Share  Save  Hide  Report 86% Upvoted

 29 

 **r/TheNewsroom** · Posted by u/Fanshelpmesleep 3 years ago 


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***.

 2 Comments  Share  Save  Hide  Report 98% Upvoted

# *The West Wing* vs *The Newsroom*


- Similar fan bases; parallel story lines


↑  
28  
↓


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


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?

 29 Comments

 Share


 Save

 Hide


 Report


86% Upvoted


↑  
29  
↓


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


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.**

 2 Comments

 Share


 Save

 Hide


 Report


98% Upvoted

↑  
5  
↓

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

Why was the lighting of the show so dark?

 12 Comments

 Share

 Save

 Hide

 Report

100% Upvoted

# Reddit Scraping: Step 0


---

- Post exists on Reddit

↑


15


↓


 **r/TheNewsroom** · Posted by u/ABondaxFan 1 year ago 


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

Yeah, that's all.

 4 Comments

 Share

 Save

 Hide

 Report

91% Upvoted

# Reddit Scraping: Step 1

---

- Utilized Reddit API to Pull Post

# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```

# Reddit Scraping: Step 1

---

- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```

# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

```
{'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  
Game and I really miss this show.", 'link_flair_richtext': [], 'subreddit_name_prefixed': 'r/Thenewsroom', 'hidden':  
False, 'pwis': 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': 'https://www.reddit.com/r/Thenewsroom/comments/7thz0u/i_just_watched_aaron_sorkins_film_mollys_game_and/',  
'thumbnail': '', 'edited': False, 'author': 'ABondaxFan', 'gildings': {}, 'content_categories': None, 'is_spoiler': False,  
None, 'banned_by': None, 'author_flair_text_color': None, 'allow_live_comments': True, 'selftext_html': '<!-- SC_OFF --><div><!-- SC_ON --></div><!-- SC_OFF -->', 'link_flair_text': None, 'view_count': None, 'archived': True, 'no_follow': False, 'is_crosspostable': 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}
```

**Title = I just watched Aaron Sorkin's Molly's Game and I really miss this show.**



# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

`{'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_spoiler': False, 'link_flair_type': 'text', 'wls': None, 'banned_by': None, 'author': 'ABondaxFan', 'allow_live_comments': True, 'selftext_html': '<!-- SC_OFF --><div><p>Yeah, that's all. </p></div><!-- SC_ON -->', '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}`

**Text = Yeah, that's all.**

# Reddit Scraping: Step 1

---

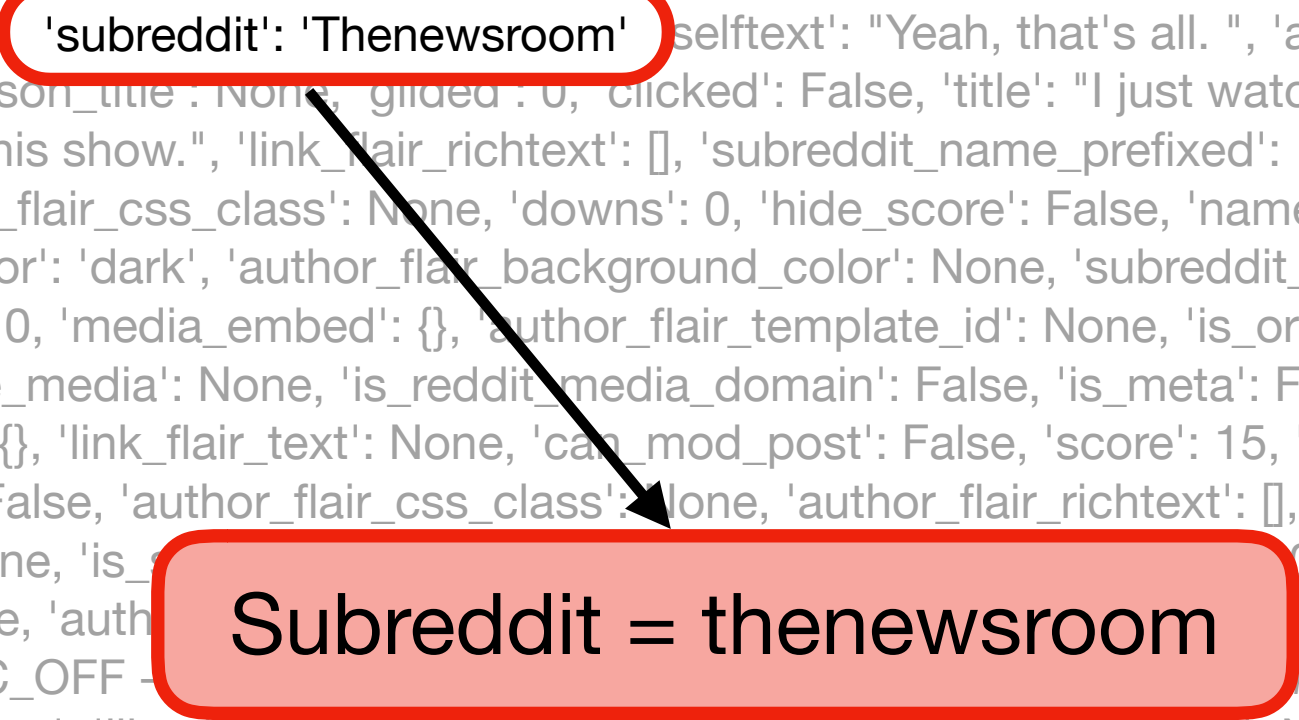
- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```

# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

```
{'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_spoiler': False, 'likes': None, 'link_flair_type': 'text', 'wls':  
None, 'banned_by': None, 'author': 'ABondaxFan', 'allow_live_comments': True,  
'selftext_html': '<!-- SC_OFF --><div class="md"><p>Yeah, that's all. </p></div><!-- SC_ON -->', '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}
```



A diagram consisting of a red rounded rectangle at the top containing the text `'subreddit': 'Thenewsroom'`. A black arrow points from this rectangle down to a larger red rounded rectangle at the bottom containing the text **Subreddit = thenewsroom**.

# Reddit Scraping: Step 1

---

- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```

# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

```
{'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_spoiler': False, 'link_flair_type': 'text', 'wls':  
None, 'banned_by': None, 'author': 'ABondaxFan', 'allow_live_comments': True,  
'selftext_html': '<!-- SC_OFF --><div class="md"><p>Yeah, that's all. </p></div><!-- SC_ON -->', '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}
```

# Reddit Scraping: Step 1

---

- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```



# Reddit Scraping: Step 1

- Utilized Reddit API to Pull Post

```
{'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_spoiler': False, 'link_flair_type': 'text', 'wls': None, 'banned_by': None, 'author': 'ABondaxFan', 'low_live_comments': True, 'selftext_html': '<!-- SC_OFF --><div class="md"><p>Yeah, that's all. </p></div><!-- SC_ON -->', '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': 'None', 'id': '7thz0u', 'is_robot_indexable': True, 'report_reasons': None, 'author': 'ABondaxFan', '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}
```

**Total comments = 4**

**'num\_comments': 4**

# Reddit Scraping: Step 1

---

- Utilized Reddit API to Pull Post

```
{'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': '<!-- SC_OFF --><div class="md"><p>Yeah, that&#39;s all. </p>\n</div><!-- SC_ON -->', '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}
```



# Reddit Scraping: Step 2

---

- Extract title + text from posts

# Reddit Scraping: Step 2

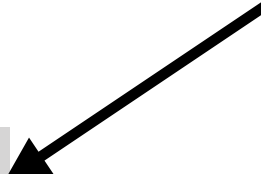
- Extract title + text from posts

```
{'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&#39;s all. &lt;/p&gt;&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}
```

# Reddit Scrapping: Step 2

- Extract title + text from posts

```
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
```



```
{'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&#39;s all. &lt;/p&gt;&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}
```

# Reddit Scrapping: Step 2

- Extract title + text from posts

```
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
```

```
{'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&#39;s all. &lt;/p&gt;&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}
```

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.

# Tokenizing, Lemmatizing and Stemming

---

- Created function to tokenize, lemmatize and stem text.

```
function lem_stem(many_posts):  
  
    create empty list of lemmatized_posts  
    create empty list of stemmed_posts  
  
    for post in many_posts:  
  
        break post up in to tokens  
        create empty string lemmatized_text  
        create empty string stemmed_text  
  
        for token in tokens:  
            lemmatize token  
            concatenate to lemmatized_text  
            stem token  
            concatenate to stemmed_text  
  
        add lemmatized_text to lemmatized_posts  
        add stemmed_text to stemmed_posts  
  
    add lemmatized_posts to many_posts  
    add stemmed_posts to many_posts  
  
    return many_posts
```

# Tokenizing, Lemmatizing and Stemming

- Created function to tokenize, lemmatize and stem text.

```
function lem_stem(many_posts):  
  
    create empty list of lemmatized_posts  
    create empty list of stemmed_posts  
  
    for post in many_posts:  
  
        break post up in to tokens  
        create empty string lemmatized_text  
        create empty string stemmed_text  
  
        for token in tokens:  
            lemmatize token  
            concatenate to lemmatized_text  
            stem token  
            concatenate to stemmed_text  
  
        add lemmatized_text to lemmatized_posts  
        add stemmed_text to stemmed_posts  
  
    add lemmatized_posts to many_posts  
    add stemmed_posts to many_posts  
  
    return many_posts
```

Function lem\_stem()  
requires

31  
minutes

to return a DataFrame  
with 1994 posts.

# Tokenizing, Lemmatizing and Stemming

---

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



# Tokenizing, Lemmatizing and Stemming

---

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

```
X = df['text']
y = df['class']

start = time.time()
vect = CountVectorizer(max_df=0.87,
                      max_features=2900, min_df=3,
                      ngram_range=(1,2),
                      stop_words='english',
                      tokenizer=LemmaTokenizer())
vect.fit(X)

X_transform = vect.transform(X)

end = time.time()
print(end - start)
```

# Tokenizing, Lemmatizing and Stemming

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

```
X = df['text']
y = df['class']

start = time.time()
vect = CountVectorizer(max_df=0.87,
                      max_features=2900, min_df=3,
                      ngram_range=(1,2),
                      stop_words='english',
                      tokenizer=LemmaTokenizer())
vect.fit(X)

X_transform = vect.transform(X)

end = time.time()
print(end - start)
```

Adding parameter  
tokenizer=LemmaTokenizer()  
to CountVectorizer added

4.59

seconds

to fitting data.

# Tokenizing, Lemmatizing and Stemming

---

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



# Tokenizing, Lemmatizing and Stemming

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



# Classification Models

- Baseline Accuracy = 50%

Model	K Nearest Neighbors	K Nearest Neighbors, TFIDF	Support Vector Machine	Random Forrest	Logistic Regression	Naive Bayes	All Model Average
Vectorizer Max Features	50	41	300	1800	4100	4100	-
Train Accuracy	74%	75%	89%	90%	95%	94%	92%
Test Accuracy	72%	75%	85%	85%	88%	89%	87%

# Classification Models

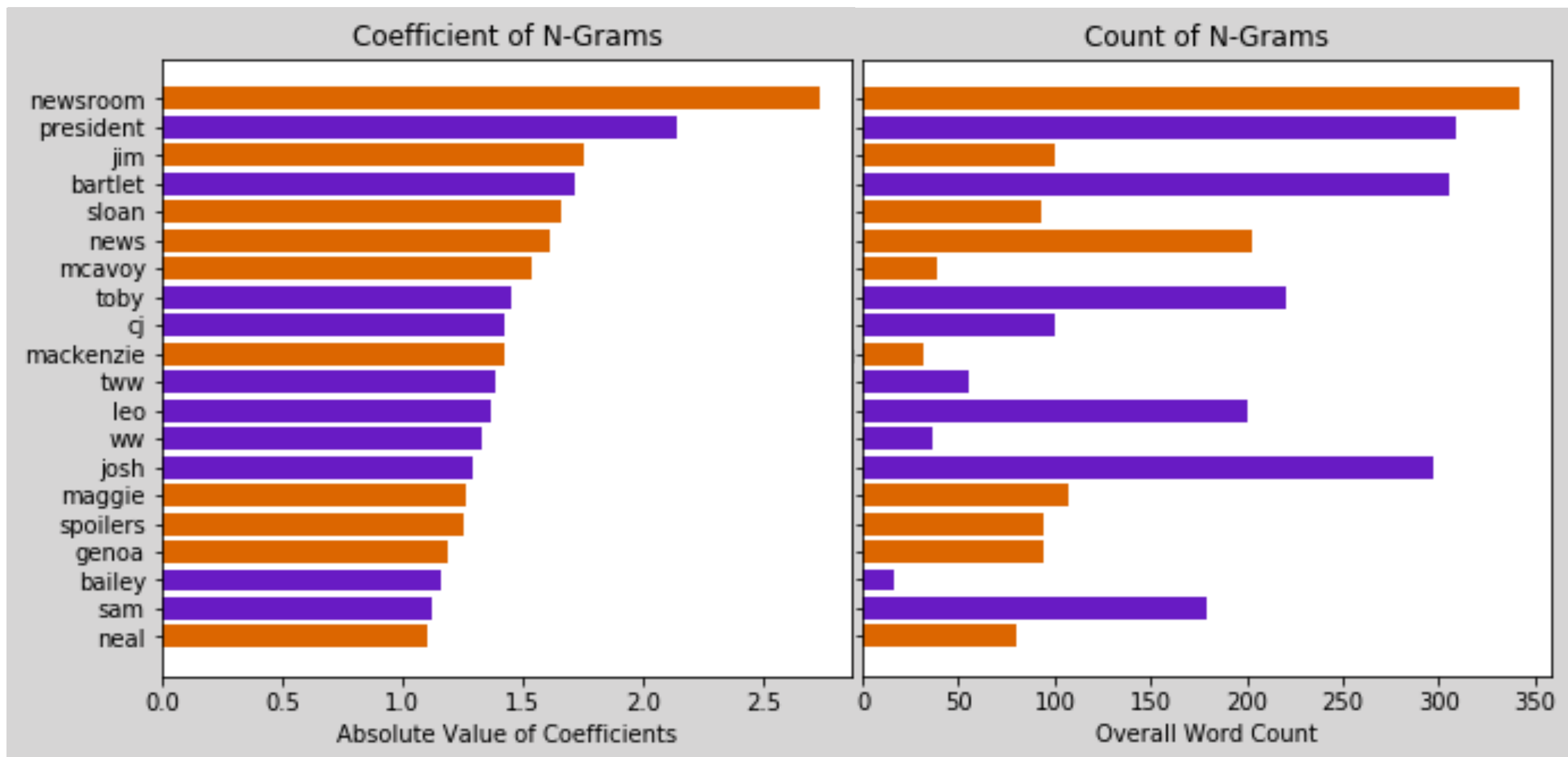
- Baseline Accuracy = 50%

Model	K Nearest Neighbors	K Nearest Neighbors, TFIDF	Support Vector Machine	Random Forrest	Logistic Regression	Naive Bayes	All Model Average
Vectorizer Max Features	50	41	300	1800	4100	4100	-
Train Accuracy	74%	75%	89%	90%	95%	94%	92%
Test Accuracy	72%	75%	85%	85%	88%	89%	87%

Best Model Confusion Matrix:	Predicted West Wing   Predicted Newsroom	
	True West Wing	55
	True Newsroom	80

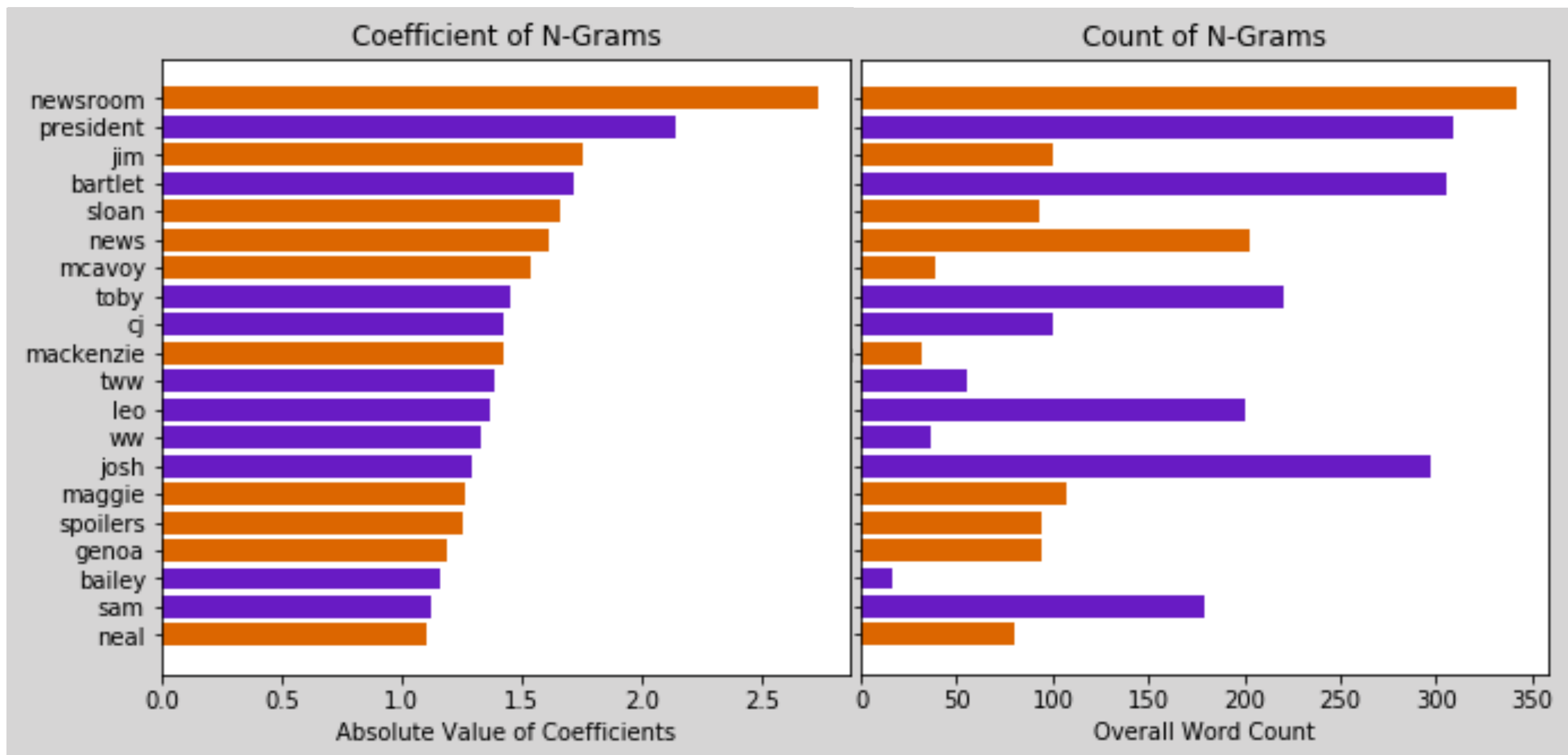
# Logistic Regression's Most Predictive N-Grams

- Most predictive N-grams are all words (1, 1) on a (1, 2) model



# Logistic Regression's Most Predictive N-Grams

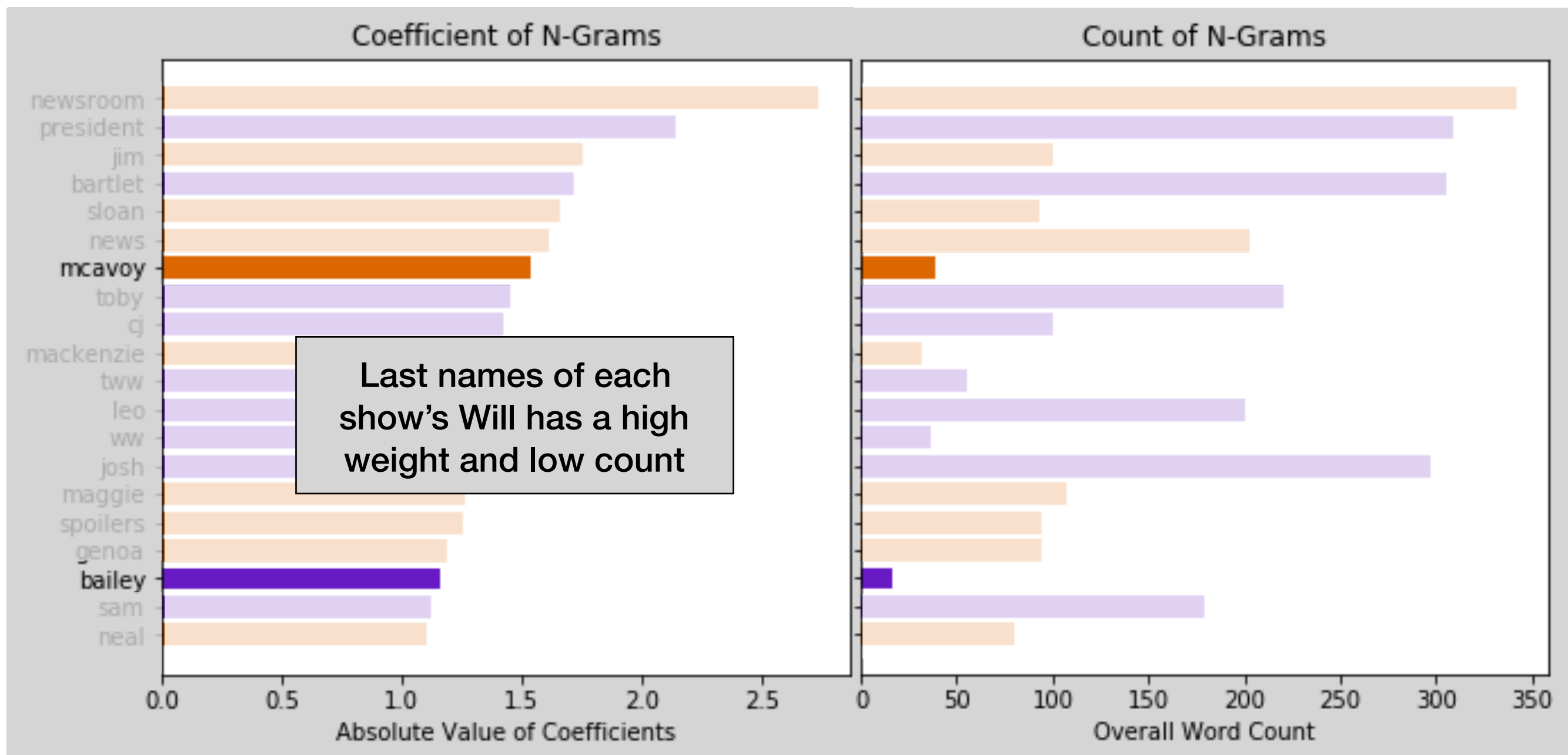
- Most predictive N-grams are all words (1, 1) on a (1, 2) model
- Mostly show names and characters





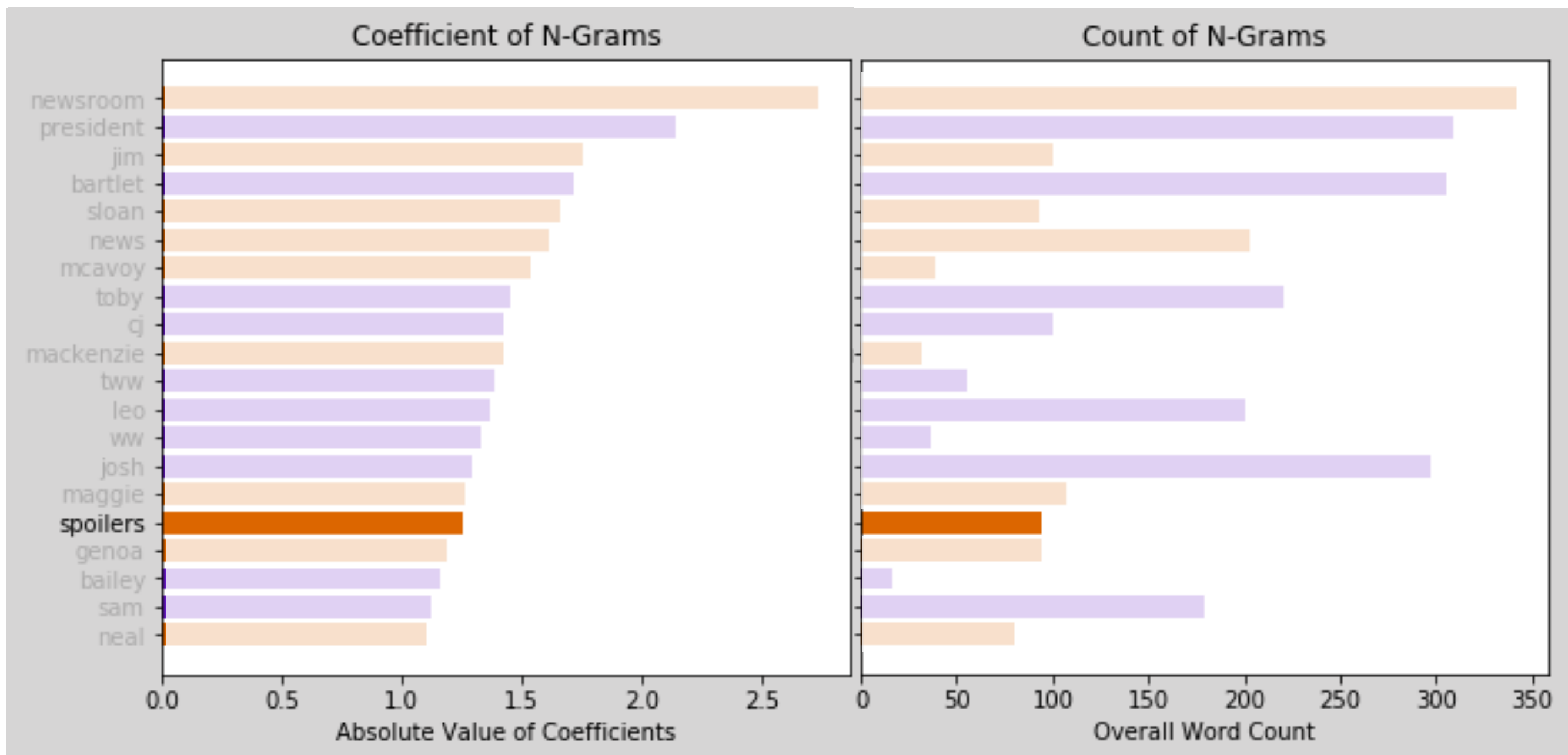
# Logistic Regression's Most Predictive N-Grams

- Most predictive N-grams are all words (1, 1) on a (1, 2) model
- Mostly show names and characters



# 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



# KNN's 50 Stemmed N-Grams

wa	1244	don	384	wing	281	anyon	226		
thi	1143	newsroom	343	love	281	peopl	221		
episod	774	sorkin	333	west	280	want	221	way	201
season	705	presid	331	whi	276	tobi	221	look	196
just	671	ha	329	west wing	275	onli	220	gt	193
like	545	charact	325	make	275	stori	216	someth	192
hi	536	say	314	did	274	final	216	ani	185
watch	519	charli	307	end	264	seri	215	doe	184
think	430	bartlet	305	scene	241	thing	207		
time	406	josh	297	becaus	240	news	203		
know	389	ve	282	realli	239	leo	201		

# 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
- Feed Logistic Regression’s predictive words in to KNN
- Seaborn plots!!!!!!

# 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)

Comment Share Save Hide Report

93% Upvoted