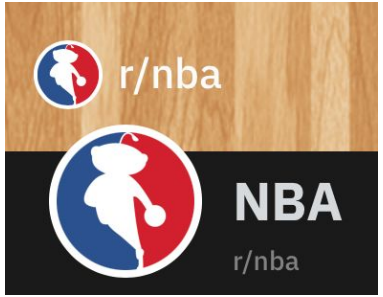
Project 3: Web Scraping & Classification

Shawn Seo, DSI 725

Reddit

- Website, collection of interest-based communities a.k.a subreddits
- Ex:
  - r/AskReddit
  - r/gaming
  - r/funny

R/nba



- Largest platform to discuss basketball with now over 5 million subscribers
- About 150 posts per day
- Over 5500 comments per day

Problem Statement

- Webscrape 10,000 threads from r/NBA with at least four pieces of information
  - Title
  - Subreddit
  - Time
  - Number of comments

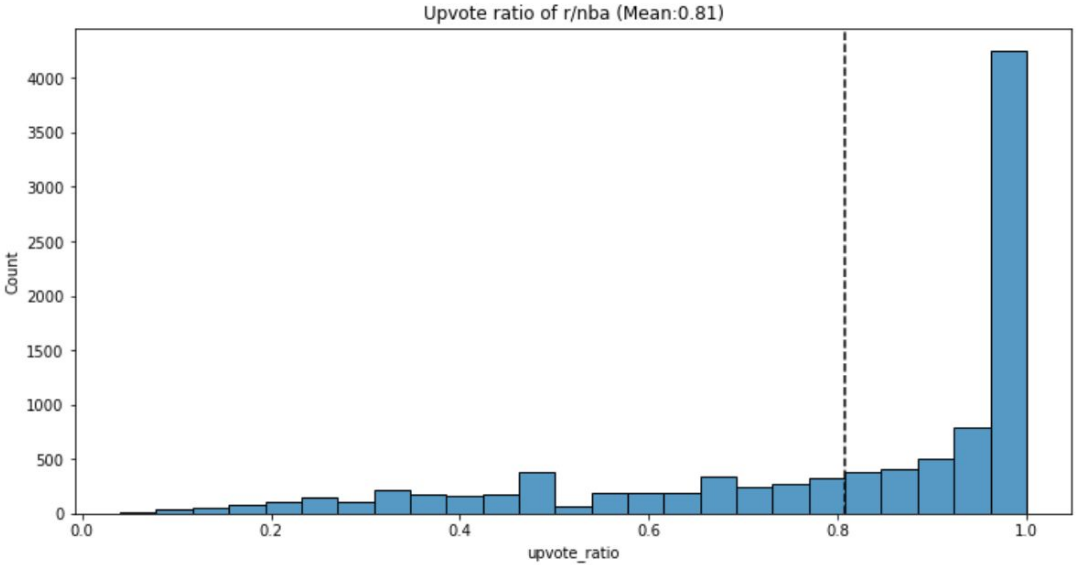
Problem Statement

- Build classification model that predicts whether or not a given Reddit post will have above or below the median number of comments

Data Cleaning & Preprocessing

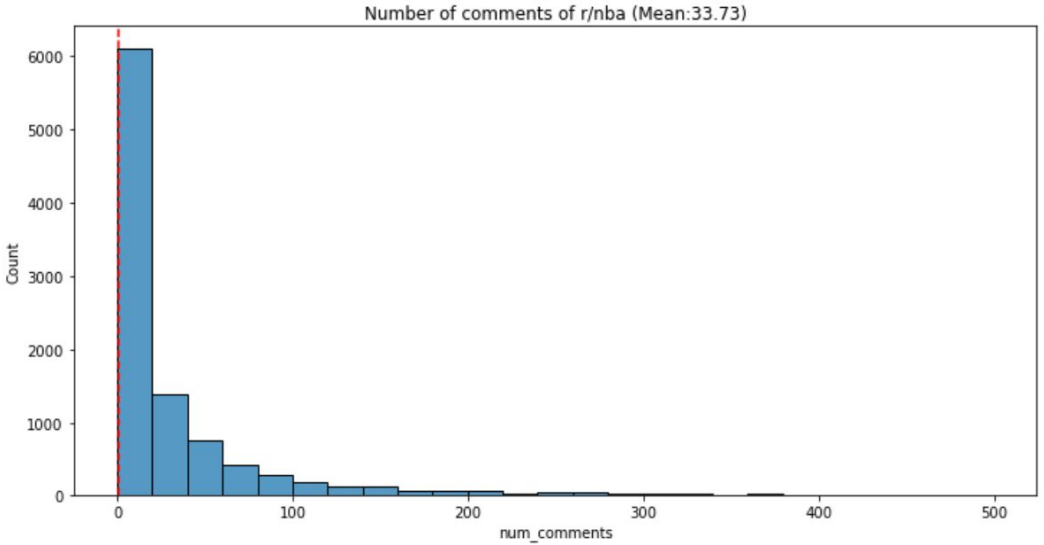
- Retrieved 10,000 posts from r/nba using PMAW (Pushshift Multithread API Wrapper)
- Dropped columns with more than 50% NaN’s
- Removed irrelevant characters from each post

EDA : Upvote Ratio



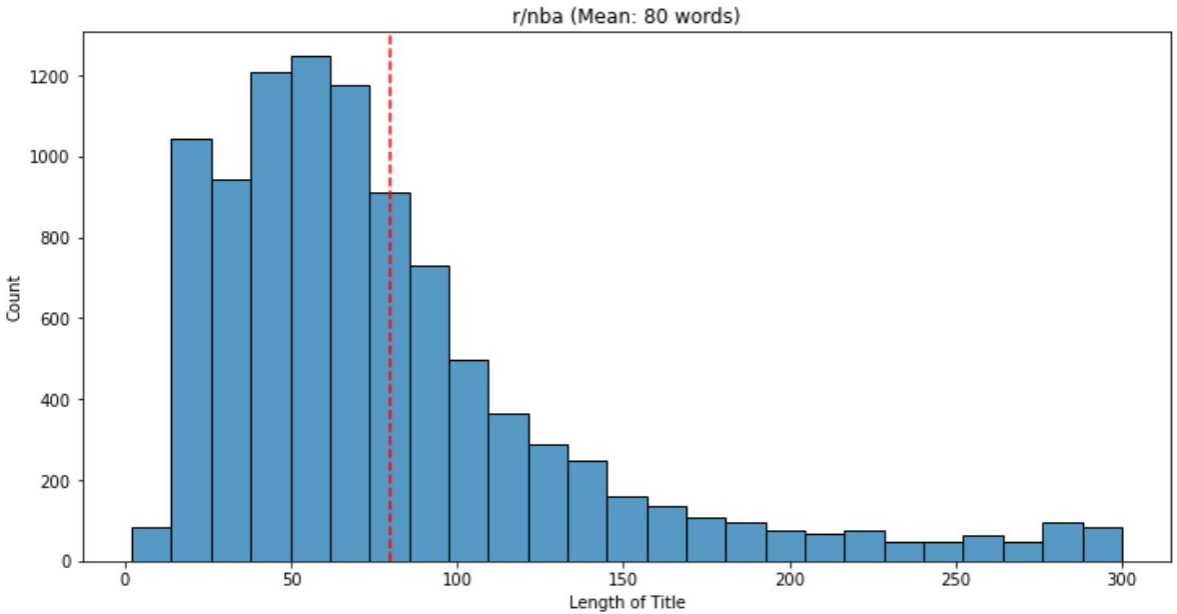
- Mean 0.81
- Left skewed
  - Most posts are ignored/buried

EDA : Number of Comments



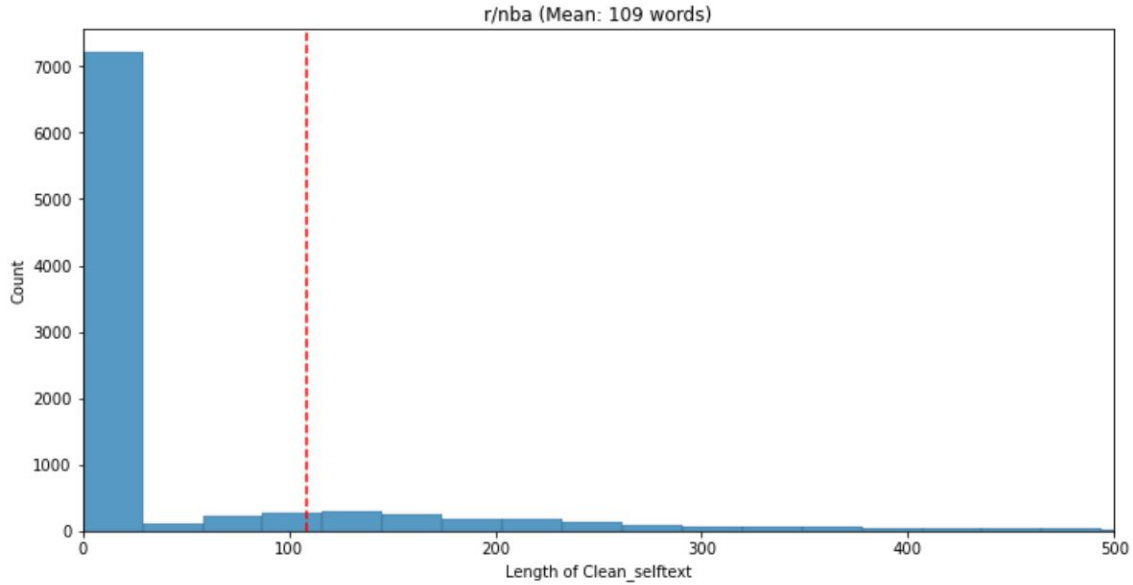
- Mean: 33.73
- Right Skewed

EDA: Average Title Length



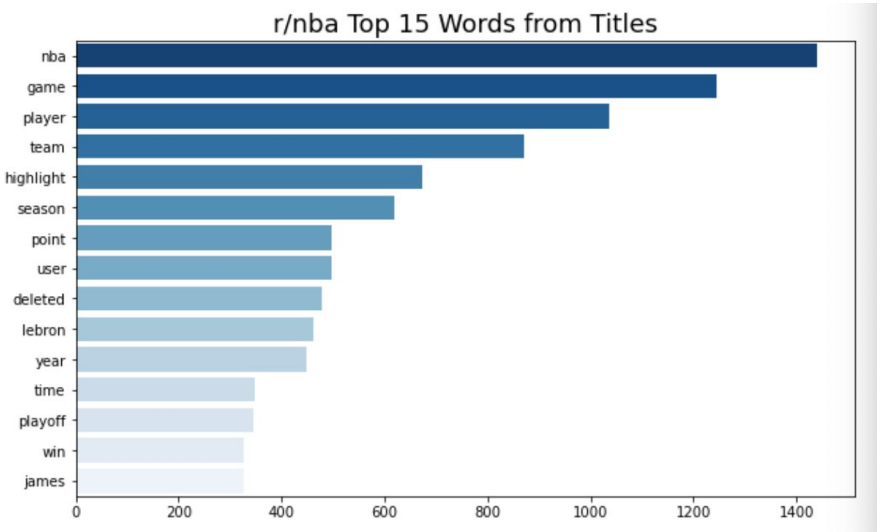
- Mean: 80 words
- Often uses quotes from interviews

EDA: Length of post



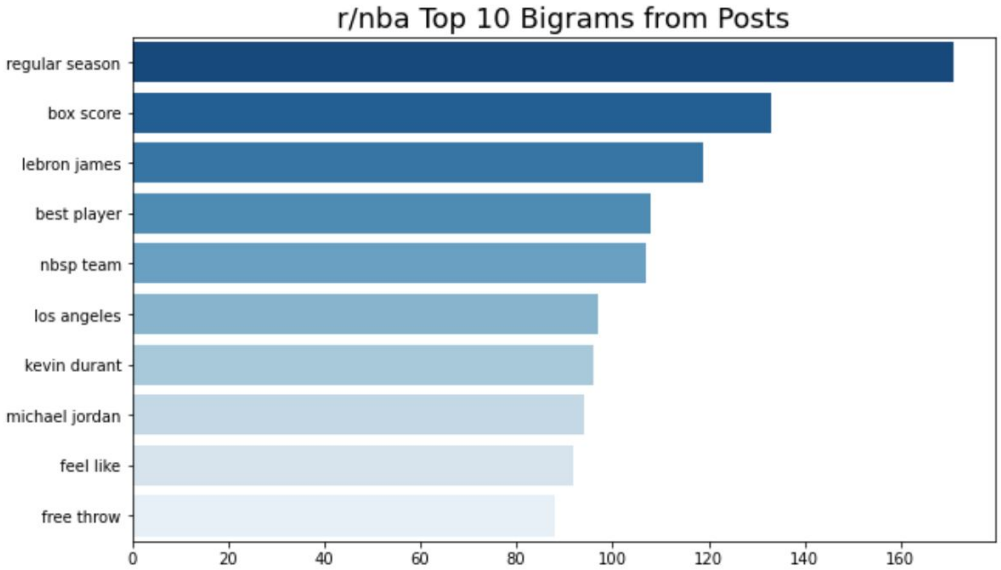
- Mean : 109 words
- Large amount of submissions without text

EDA: Top 15 words from Titles



- General basketball terms
- Lebron James

Top 10 Bigrams



- More names

Model Selection

```
{'model': 'rf',
 'vectorizer': 'cvec',
 'train': 0.9829222011385199,
 'test': 0.7483739837398374}

True Negatives: 1076
False Positives: 180
False Negatives: 439
True Positives: 765
```

- Test Accuracy: 0.75

Limitations

- User interaction differs greatly between season and out-of-season
- Less analysis out of season / more memes
- May be some bias in the title structure due to subreddit posting rules

Recommendations

- Use star player names in the title/post.
- Include highlights
-