r/Kanye or r/drizzy?:

Are kanye and drake fans alike?

An NLP Classification Analysis by Nati Marcus

Problem Statement

As two of the most popular hip-hop artists of the past decade (or longer), Kanye and Drake have sparked debate over which artist is better, and if either is the best ever. Although Kanye fans might laugh at the idea that Drake may be the better artist and vice versa, the two fan bases might not be that different. In this study, I will be analyzing reddit comments on the Kanye subreddit (r/kanye) and the Drake subreddit (r/drizzy) to see if fans can be properly classified as Kanye fans or Drake fans. Using classification accuracy score, I will evaluate how well I can properly classify the comments in comparison to the baseline accuracy of 51%.

Source: https://www.billboard.com/articles/columns/hip-hop/9604684/kanye-west-rolling-loud-miami-2021

Source: https://people.com/music/drake-makes-history-first-artist-to-debut-tracks-at-1-2-3-on-hot-100/

Data Collection

Pushshift API

Source: https://www.redditinc.com/brand

- Used function to pull subreddit comments from Pushshift Reddit API
- Drake Total: 21,300 observations
- Kanye Total: 22,300 observations
- Features for Both: body, author, created_utc, subreddit

EDA

- Dropping index columns from each df
- Merging two datasets, resetting index
- Removing ‘[removed]’ & ‘[deleted]’ values
- Removing ‘Yes’ & ‘No’ values

“New one reminds me of the MBDTF ballerina”

–Internet-Ivan, r/Kanye

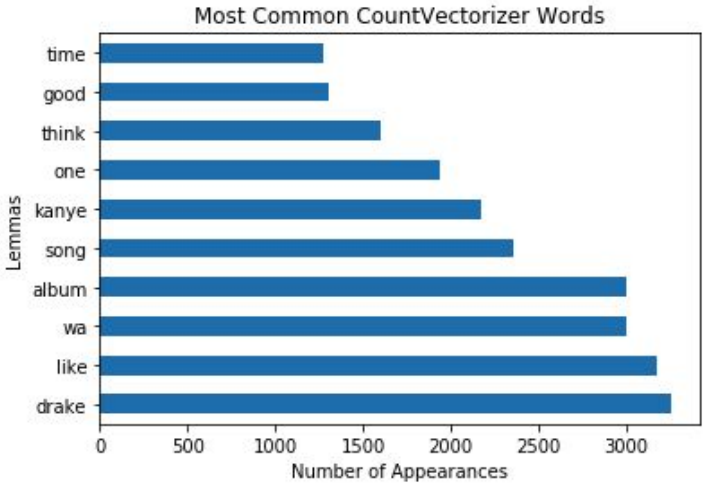
"Views is my favourite album and I actually agree with you. It would've actually been way better. But notice how it's like 90% love songs and Drake knows he has to satisfy the rap guys too on every album. So he probably still would've gotten lots of criticism"

–Charisma_Percept67, r/drizzy

Preprocessing

- Regex Tokens
- Created Lemmas with WordNetLemmatizer
- Removed bad lemmas such as ‘’ and emojis
- Cleaned tokens to remove numbers from a token
- Cleaned tokens to remove words with more than 2 consecutive letters (e.g. ‘helllo’ → ‘hello’)
- Updated lemmas based on updated tokens
- Assigned 1 to r/drizzy, 0 to r/kanye for classification

Modeling



- CountVectorizer & TfidfVectorizer Models
- Random Forest, SVC, AdaBoost, Naive Bayes
- RandomSearchCV for some test, cross val score for others

Model results: CountVectorizer

- ~71% accuracy on train, ~70% accuracy on test for Random Forest
- Best SVC Model was 70% accurate on train, 68% accurate on test
- AdaBoost Model was only 65% accurate on train, and 64% accurate on test
- Naive Bayes was 70% accurate on both train and test
- All improvements from baseline score of 51%

Model results: tfidfvectorizer

- 72% accuracy on train, 69% accurate on test for Random Forest Model
- 71% accurate on train, 70% accurate on test for SVC
- 65% accurate on train and test for AdaBoost Model
- 70% accurate on train and test for Naive Bayes
- All improvements from baseline score of 51%

Further exploration

- GridSearch and RandomSearch to tune hyperparameters
- Look deeper into stop_words

Conclusions

- All Models performed between 15% and 20% better than the baseline score
- For CountVectorized models, Random Forest Models were best at classifying the comments
- For TfidfVectorized models, SVC Models were best at classifying the comments

Thank You!