My Beautiful Dark Twisted Triadic Closure

Summary Report

NETS 150
Spring 2016
Daniel Moreno
Connor Swords
Serena Tsay
Connor Wen

PARENTAL
ADVISORY
EXPLICIT CONTENT

Part 1:
Kanye West's Vernacular in Album Lyrics and Tweets

Inspiration
Kanye West is one of the most popular artists in the world. He is popular for his candid personality and his fusion of old and new sounds to create timeless music. Last year, one of Kanye's collaborators, Rhymesmith (formerly known as Che Smith) released a documentary detailing his extensive work for Kanye's songs. Understanding how songwriting often comes from a variety of external sources spiked our interest in quantifying Kanye's own influence in writing his music. To explore Kanye's originality and creativity, we compared how similar Yeezy's tweets (which are as close to his own thoughts as we can get) to his released song lyrics. Kanye's twitter feed is at times a cryptic barrage of intriguing Kanye-isms, which leads to a very interesting data set to use in our analysis.

Hypotheses
In this part of the project, we decided to test the following 2 hypotheses:
1. There will be high cosine similarity between Kanye's Tweets and his album lyrics, assuming that his vocabulary is limited and/or he writes his own songs.
2. When comparing similarities between albums, there will be cosine similarities for the albums released further from each other will be smaller than the cosine similarities for albums released closer to each other.

Approach & Analysis
To test these hypotheses, we first needed to compile Kanye's album lyrics. We luckily found a RESTful API called *kanyeREST* that we could send queries to and retrieve the lyrics for album in a JSON response. We compiled each album's lyrics into a text file (i.e. "yeezus.txt"), which we then added to our corpus. We scraped the tweets from his Twitter profile (@kanyewest) using the Twitter4j library, used regex to retrieve the body of each tweet and compiled it in a text file "kanyeTweets.txt". We then ran Professor Swap's Vector Space Model cosine similarity methods to find the similarities between album lyrics and tweets.

These albums include:
•   The College Dropout (2004)

- Late Registration (2005)
- Graduation (2007)
- 808s and Heartbreak (2008)
- My Beautiful Dark Twisted Fantasy (2010)
- Watch the Throne (2011)
- Yeezus (2013)
- The Life of Pablo (2016)

Hypothesis 1: Accepted Hypothesis (with caveats)

In analyzing the cosine similarities between Kanye's tweets and album song lyrics (See Figure A), we found a trend where his early and recently released albums, with an exception of the recent album Yeezus (2013), are similar to how he writes. This suggests that lyrics in his early and later albums are written, for the most part, by Kanye himself. Therefore, we accept the hypothesis that there is high cosine similarity for albums Kanye wrote himself.
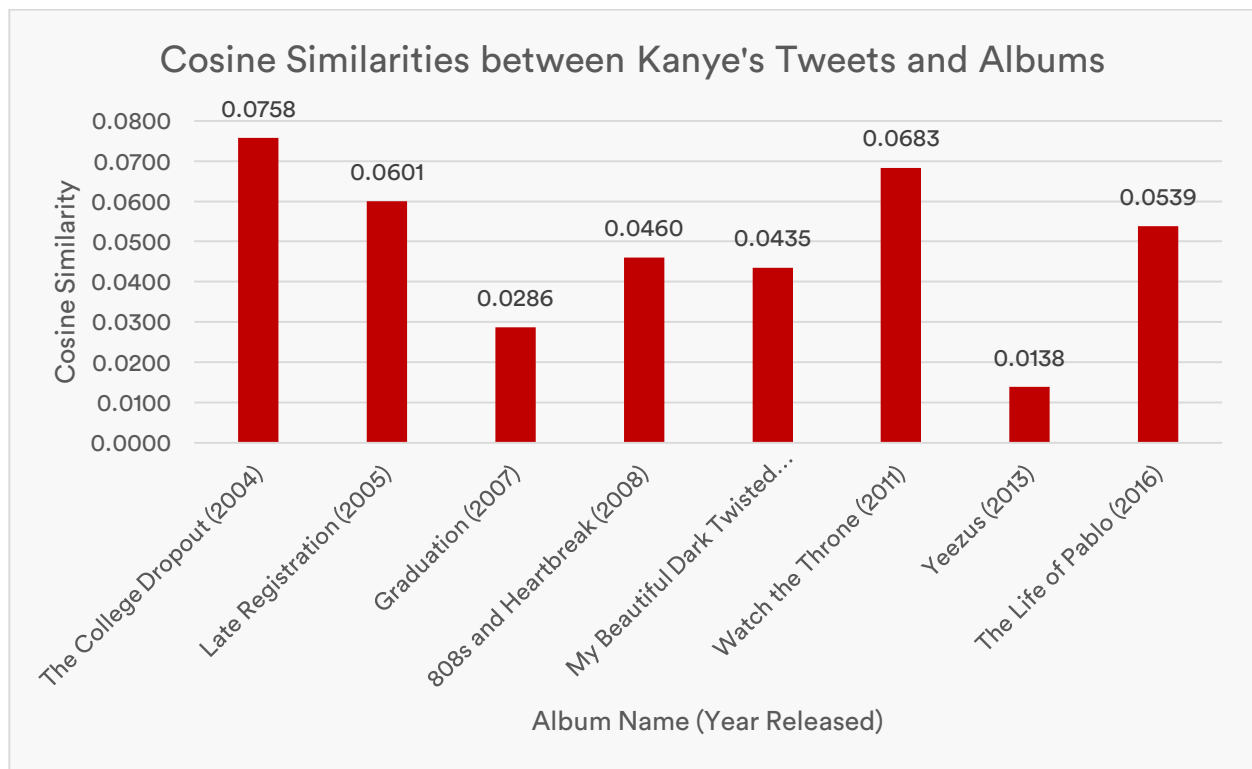

Figure A: Cosine similarities between Tweets and song lyrics in Kanye's albums

Hypothesis 2: Reject Hypothesis

After analyzing the cosine similarities between each pair of albums (See Figure B), we found that while there is a general trend that albums released closer to each other

have higher cosine similarities, there are a few cases where this is not the case. For example, for Kanye's recently released The Life of Pablo album, the highest cosine similarity value is with one of his earliest albums, The College Dropout.

Album (Year Released)	The College Dropout (2004)	Late Registration (2005)	Graduation (2007)	808s and Heartbreak (2008)
The College Dropout (2004)		0.150600799	0.061820272	0.017411635
Late Registration (2005)	0.150600799		0.056886495	0.029657665
Graduation (2007)	0.061820272	0.056886495		0.016012039
808s and Heartbreak (2008)	0.017411635	0.029657665	0.016012039	
My Beautiful Dark Twisted Fantasy (2010)	0.07994871	0.068026646	0.085532197	0.01527598
Watch the Throne (2011)	0.096146637	0.07651424	0.052113355	0.014188599
Yeezus (2013)	0.025772712	0.026942569	0.010655662	0.007158961
The Life of Pablo (2016)	0.050260533	0.040842155	0.032351494	0.008239215

continued	My Beautiful Dark Twisted Fantasy (2010)	Watch the Throne (2011)	Yeezus (2013)	The Life of Pablo (2016)
The College Dropout (2004)	0.07994871	0.096146637	0.025772712	0.050260533
Late Registration (2005)	0.068026646	0.07651424	0.026942569	0.040842155
Graduation (2007)	0.085532197	0.052113355	0.010655662	0.032351494
808s and Heartbreak (2008)	0.01527598	0.014188599	0.007158961	0.008239215
My Beautiful Dark Twisted Fantasy (2010)		0.064690742	0.019381618	0.035833393
Watch the Throne (2011)	0.064690742		0.047936011	0.043535593
Yeezus (2013)	0.019381618	0.047936011		0.011812464
The Life of Pablo (2016)	0.035833393	0.043535593	0.011812464	

Figure B: Cosine similarities between Kanye's albums. The darker the green color, the higher the relative cosine similarity value. Light green/white represents relatively low cosine similarity value

Code
Corpus.java - Generates an inverted index for documents
Document.java - Keeps track of term frequencies
TweetsMain.java - Scrapes album lyrics from kanyeREST, saves lyrics as files. Adds album lyrics files and tweets file into a corpus.
TwitterScrape.java - Scrapes Kanye's Twitter for all his tweets, saves tweets into a file.
VectorSpaceModel.java - Implementation of the vector space model

Dataset
- Tweets: scraped from @kanyewest
- Album lyrics: scraped by album from kanyeREST API (i.e. http://www.kanyerest.xyz/api/album/the_life_of_pablo)

Part 2:

Kanye's Grammy Network

Inspiration

Kanye is one of the most decorated artists of all-time, totaling 57 Grammy nominations and 21 Grammy wins throughout his career. However, he's had significant help. Kanye has depended on countless other artists and producers to create some of his most memorable albums and tracks. We wanted to explore this network in order to gain insight into the community of supporters behind Kanye's biggest hits. Using JSoup, we crawled through Kanye's award page on IMDb and pulled data on his collaborators. In order to avoid duplication, we decided to narrow our graph to Grammy awards and nominations, the most credible and expansive of Kanye's awards. We used the data from IMDb to construct an undirected graph of Person objects, where the nodes represent artists/producers and the edges represent an album or track for which they were both nominated. After the graph was constructed, we were able to apply graph analysis techniques from the course.

Hypothesis

In this part of the project, we decided to test the following hypothesis:

> We expect that Kanye's neighborhood overlap values will be higher with non-artists (producers & songwriters) than with featured artists on his songs.

Analysis

After calculating the neighborhood values between Kanye and each respective collaborator, we are able to visually portray the data in Figure C below:

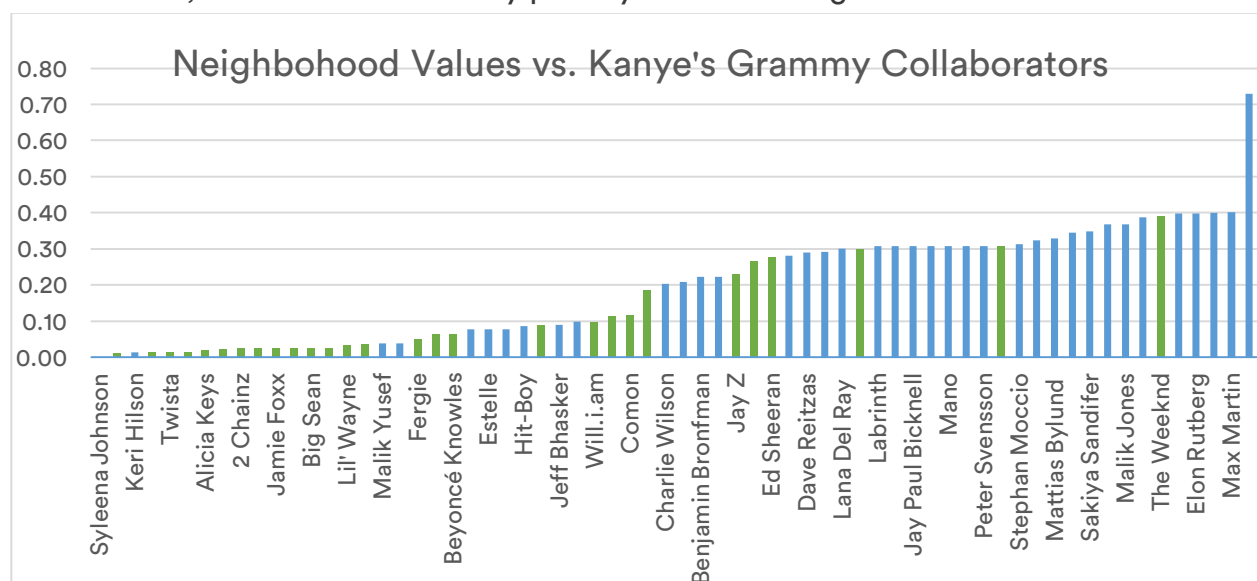Neighbohood Values vs. Kanye's Grammy Collaborators

Figure D below also shows how the average neighborhood overlap values for producers ended up being much higher than the average value for artists.
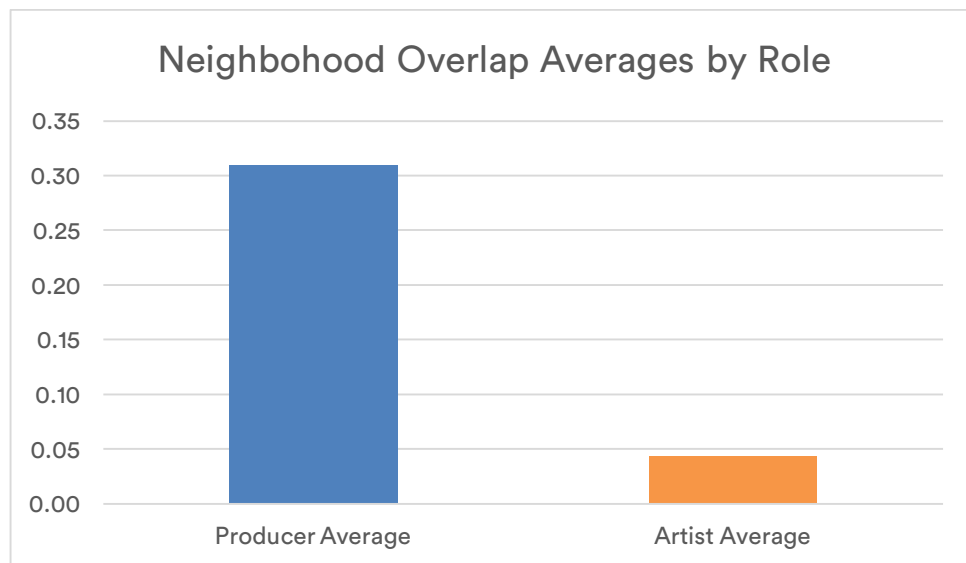

Figure D: Average Neighborhood Overlap Scores by Role

Conclusion

As you can see, it is evident that producers (especially Mike Dean) have much higher neighborhood overlap values than other artists. This may be because there are a lot less producers and songwriters so they're in higher demand. Or, this may simply be because producers are able to work on a lot more projects at a time. Or, Kanye's producers are a special case in that they are incredibly connected and significant within the music community. Regardless, it's cool to see that they are awarded/nominated for Grammy's despite not being that publicly recognized for their work.

Code

JSoupQuery.java - Scrapes data from IMDb
Person.java - Person object that stores the name and collaborator list (edges)
PersonURL.java – Maps a Kanye collaborator to his IMDb page.
GraphCreator.java - Creates the graph as a set of Person objects
GrammyMain.java - Parses and prints the IMDB data scraped from the web

Dataset

- Grammy Data from Kanye West and others' IMDb pages.