Song Clustering Analysis Report

By: Akash Jauhar & Brian Kreidberg

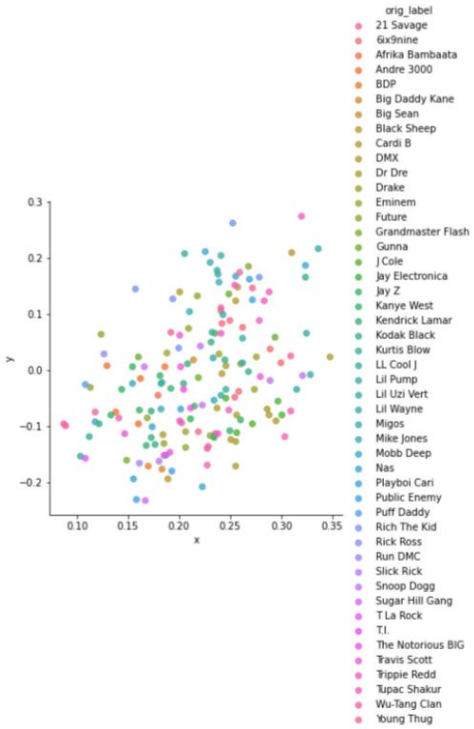
CSC149: Text Mining

Dr. Doboli

Preliminary test:

Initially we wanted to run the clustering with all the artists. The output was very hard to read though because there were 150 dots on the graph which is hard to distinguish. The different colors are difficult to distinguish.

```
In [24]: 1 sb.lmplot(data=data_x2, x='x', y='y', hue='orig_label',fit_reg=False, legend=True, l
Out[24]: <seaborn.axisgrid.FacetGrid at 0x7f832e923670>
```



The purity was very low, because with that many songs of the same genre it is likely they will use similar language. The purity for the TF-IDF was 9% and for the LDA the purity was 7%.

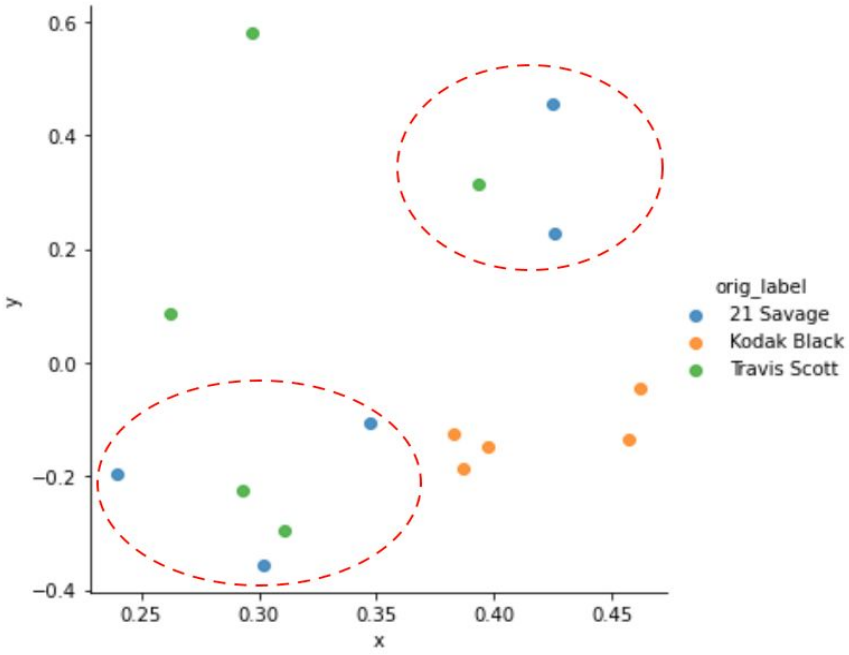
Goal:

We wanted to have 2 data sets, one with artists that were very similar to each other and one that had very different artists, however finding the amount of different artists to put in those two tests was challenging. Initially we tried finding 5 artists that were similar and 5 that were very different. However, finding 5 that were different turned out to be very challenging, so we decided to go with groups of 3 artists.

Algorithm used - K-means with tf-id

We selected a few artists in our dataset to find the common lyrics/words used by them. We were able to create clusters based on k-means clustering. This allowed us to see the different clusters of words in songs used by the several different artists.

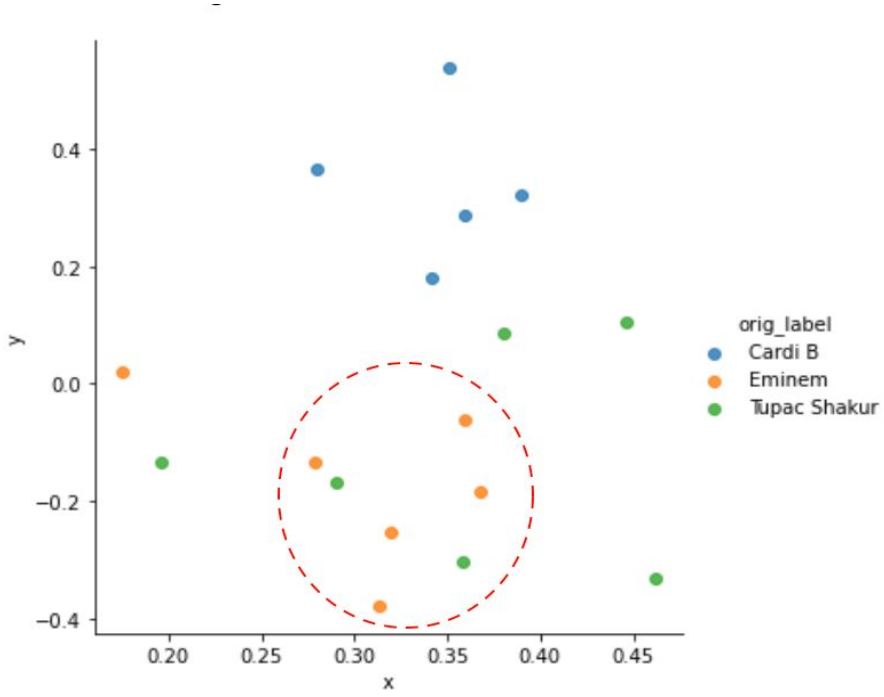
Test 1:



Initially we tried to pick 3 artists we thought were pretty similar. We were able to obtain a purity of 53% with 4 clusters using the following 3 artists: 21 Savage, Kodak Black, and Travis Scott.

21 Savage and Travis Scott seem to be pretty similar however Kodak Black seems to be a bit of an outlier. The words that we got back were expected for the most part. Initially I expected there to be a lot of curse words because cursing is a large part of rap music and the results show that there was.

Test 2:



For our second set of data we tried to pick artists that are less similar to each other, i.e. ones that had very different styles and lyrics from each other. The hope was to try and find a high purity between different artists. The artists that were chosen this time were Cardi B, Eminem, and Tupac Shakur. Again we used 4 clusters but this time the purity was 65%. As predicted the purity was higher because the artist used have very different styles/lyrics.

Overall Analysis For the Above 2 Tests/Analysis:

We spent time picking and choosing the artists that we thought would help us with the analysis above. Deciding how many clusters to look at was also something that was discussed and debated. For the CountVectorizer parameter we went back and forth regarding the max_df and min_df numbers to give us appreciated results. This helped us to filter out the data and reach the purity levels stated above. We also

used the SVD data reduction method to reduce the dimensionality of the input matrix. Using SVD helped us to visualize the clusters and make further observations.

Analysis using LDA:

For our second test we used LDA with K-means. We ran the test again with the same data as last time. We first tried clustering the artists we thought would be similar which were 21 Savage, Kodak Black, and Travis Scott. We found that with 4 clusters we had purity 46% which was similar to the 53% we got before with the tf-idf version. We then tested the second data set which contained Eminem, Tupac Skakur, and Cardi B and obtained a purity of 64% which was similar to the 65% purity that we had from the tf-idf version. We expected the output to be similar to the tf-idf k-means which it was. We tested with several different numbers of clusters, but the purity for data set two with Eminem, Tupac Shakur, Cardi B was higher than data set one with Kodak Black, 21 Savage, and Travis Scott.

We also played around with the "number of topics" parameter to see many topics made sense for our dataset. We came to a conclusion that looking at 3 topics with our dataset was the most ideal and gave us better results.

We continued to compare the artists and saw expected results: We saw that 21 Savage had similar songs/lyrics to Travis scott which valids the clusters analysed in the k-means tf-idf part of this report on the top.

```
Original Song:
21 Savage

         brand new mak nineti with the drum attach drum you a shit talker we got drum for that on god tryna f
ist fight boy you dumb for that stupid you gone get a bullet in your lung for that stupid draco get to kick l
ike liu kang twenty-on fn on me in the mulsann straight up straight up glock nineteen in the blue flame strai
ght up straight up i wa strap when i slid insid your boo thang on god twenty-on immort we'll never die twenty
-on loyal to my brother yeah i'll never lie on god call me bird dog 'caus i cheat and i'm fli straight up and
i love win i'm upset if it' a tie on god get you off the ground just to knock you to the floor yeah let' go b
low for blow yeah let' go toe to toe straight up when it' time to battl they don't never ever show twenty-on
when it' time to battl i'm the first one at the door twenty-on at the door with a draco don't nobodi move i d
on't wanna have to blow rap ass nigga get spin at they show that' whi i need some help i got a fetish for the
smoke it get fatal in the bottom use to rumbl in the den twenty-on

tf-idf most similar
Travis Scott

         she got hip i gotta grip for yeah a lot of ass don't need to have more i know it' sweet i like that
mmmmmm straight up i got word that it' wet well let' drown toot it up back it up slap it down don't say a wor
d of what you heard from when i came around it' lit you get it first you get thi work right when you come in
town yeah need you right here yeah know you the queen of give idea no more new friend don't bring the hype he
re ooh know you got problem but it' not fair

lsi most similar
Travis Scott

         she got hip i gotta grip for yeah a lot of ass don't need to have more i know it' sweet i like that
mmmmmm straight up i got word that it' wet well let' drown toot it up back it up slap it down don't say a wor
d of what you heard from when i came around it' lit you get it first you get thi work right when you come in
town yeah need you right here yeah know you the queen of give idea no more new friend don't bring the hype he
re ooh know you got problem but it' not fair

lda most similar
Travis Scott

          made thi here with all the ice on in the booth at the gate outsid when they pull up they get me loo
s yeah jump out boy that' nike boy hop out coup thi shit way too big when we pull up give me the loot gimm th
e loot wa off the remi had a papoos had to hit my old town to duck the news twofour hour lockdown we made no
move now it' four am and i'm back up pop with the crew i just land in chase b mix pop like jamba juic differ
color chain think my jewel realli sell fruit and they choke man know the cracker wish it wa a noos somesomeso
mesomeon said
```

Furthermore, we saw that Eminem had similar songs/lyrics to Tupac Shakur in this analysis (screenshot shown below). This also valides our analysis for the k-means tf-idf parts of this report and makes sense why Eminem and Tupac Shakur have a higher similarity. This also saws that Cardi B uses different lyrics and has a distinct style when representing herself since she didn't appear in the analysis we did using the similar_artists function towards the end of our project.

```
Original Song:
Eminem

        y'all act like you never seen a white person befor jaw all on the floor like
pam like tommi just burst in the door and start whoop her ass wors than befor they fir
st were divorc su her over furnitur agh it' the return of the ah wait no way you'r kid
he didn't just say what i think he did did he and dr dre said noth you idiot dr dre' d
ead he' lock in my basement ha ha feminist women love eminem chicka chicka chicka slim
shadi i'm sick of him look at him walk around grab hi youknowwhat flip the youknowwho
yeah but he' so cute though yeah i probabl got a coupl of screw up in my head loos but
no wors than what' go on in your parents' bedroom sometim i wanna get on tv and just l
et loos but can't but it' cool for tom green to hump a dead moos my bum is on your lip
my bum is on your lip and if i'm lucki you might just give it a littl kiss and that' t
he messag that we deliv to littl kid and expect them not to know what a woman' clitori
is of cours they'r gonna know what intercours is by the time they hit fourth grade the
y'v got the discoveri channel don't they we aingt noth but mammalswel some of us canni
b who cut other peopl open like cantaloup but if we can hump dead anim and antelop the
n there' no reason that a man and

tf-idf most similar
Tupac Shakur

        say the blacker the berri the sweeter the juic i say the darker the flesh the
n the deeper the root oh i give a holla to my sister on welfar two pac care if don't n
obodi els care oh and uh i know they like to beat you down a lot when you come around
the block brother clown a lot but pleas don't cri dri your eye never let up forgiv but
don't forget girl keep ya head up and when he tell you you aingt noth don't believ him
and if he can't learn to love you you should leav him 'caus sister you don't need him
and i aingt tri to ga you up i just call 'em how i see 'em you don't need him you know
what make me unhappi what' that when brother make babi and leav a young mother to be a
pappi oh yeah yeah yeah and sinc we all came from a woman got our name from a woman an
d our game from a woman yeah yeah i wonder whi we take from our women whi we rape our
women do we hate our women whi whi whi whi whi whi i think it' time to kill for our wo
men time to heal our women be real to our women and if we don't we'll have a race

lsi most similar
Tupac Shakur

        say the blacker the berri the sweeter the juic i say the darker the flesh the
n the deeper the root oh i give a holla to my sister on welfar two pac care if don't n
obodi els care oh and uh i know they like to beat you down a lot when you come around
the block brother clown a lot but pleas don't cri dri your eye never let up forgiv but
don't forget girl keep ya head up and when he tell you you aingt noth don't believ him
and if he can't learn to love you you should leav him 'caus sister you don't need him
and i aingt tri to ga you up i just call 'em how i see 'em you don't need him you know
what make me unhappi what' that when brother make babi and leav a young mother to be a
pappi oh yeah yeah yeah and sinc we all came from a woman got our name from a woman an
d our game from a woman yeah yeah i wonder whi we take from our women whi we rape our
women do we hate our women whi whi whi whi whi whi i think it' time to kill for our wo
men time to heal our women be real to our women and if we don't we'll have a race

lda most similar
Tupac Shakur

        when i wa young me and my mama had beef seventeen year old kick out on the st
reet though back at the time i never thought i'd see her face aingt a woman aliv that
could take my mama' place suspend from school and scare to go home i wa a fool with th
e big boy break all the rule i shed tear with my babi sister over the year we wa poore
r than the other littl kid and even though we had differ daddi the same drama when thi
ng went wrong we'd blame mama i reminisc on the stress i caus it wa hell hug on my mam
```

Conclusion:

Picking a dataset that made sense for what we were trying to do was challenging at first. Initially, we had to decide how many artists to put in our datasets and what kind of parameters to use for our analysis. We think that the output we got from LDA was better than the TF-IDF, because with the TF-IDF there was only an 8% difference in the puritys for the 2 data sets. The LDA produced a 18% difference in purity on the same data sets which gives much more variation in the outputs. All in all, we learnt that sometimes data can be tricky.