Sabancı Üniversitesi

ENS492 Project Progress-II Report

Analysis of Billboard Top 100 List 1946-2014

14754 Sarp Terlemez

15563 Elif Meriç

18186 Hüseyin Güven

Supervised by İlker Birbil

11/2015

Sabanci University

Faculty of Engineering and Natural Sciences

1. Introduction

Our project's first step is to implement a code in Python successfully in order to collect popular song's lyrics from the Billboard Top 100 list between the years 1946-2014. After collecting lyrics and creating a database with these lyrics, we are going to create a word list about several concepts such as Sexism. In order to see the differences among these 68 years and distinguish these years, we decided to consider ruling presidents of USA and their influences to lyrics.

The challenges of this project is that to learn how to use the Python programming language since none of our group members was familiar to Python software and programming language so, in order to overcome these challenges, all of the group members should be able to work with this software and need to read literature reviews since we might have to come up with successful code and word cluster about the sexism concept.

2. Project Definition and Scope

2.1.Project Definition

Data mining, in general, is the computational process of analyzing large datasets involving methods at the intersection of statistics and database systems. Data mining, itself, is a subfield of computer sciences.

Our project is merely based on text mining among Billboard Magazine's Hot 100 songs in between the years 1946 and 2014. In order to collect all the lyrics of songs which are in the list of Billboard Magazine, we needed to develop a program which helps us to download the lyrics from internet and has written the codes using Python language. Furthermore, we classified the words that will be searched in the lyrics to get healthy statistical results.

2.2.Project Objective

Our main objective is collecting all song's lyrics from the Hot 100 list of Billboard's between the years 1946-2014 and analyse these lyrics in terms of sexism. In order to reach statistical and accurate results we needed to create word cluster which includes a lot of sexist words. After creating a sexist words cluster, our program is going to search all lyrics and get numerical information about specified words in our cluster. After searching and getting results from on hand lyrics dataset in terms of sexist words, we are going to analyze the results and interpret the results in terms of ruling presidents of USA for possible correlations.

2.3.Scope of the Project

Firstly, it is planned to describe the processes of programming and details about the codes. After giving brief information about the working principle of the program, selected words to search in the lyrics will be listed. Furthermore we will search the lyrics for every single word in our cluster about sexism. Numerical search results will be given according to years and ruling presidents of USA between 1946 and 2014. Numerical results will be analysed statistically and correlations between years and the number of sexist words used in lyrics are going to be analogised and results to be interpreted.

3. Project Design and Implementation

3.1 Project Design Data Collection

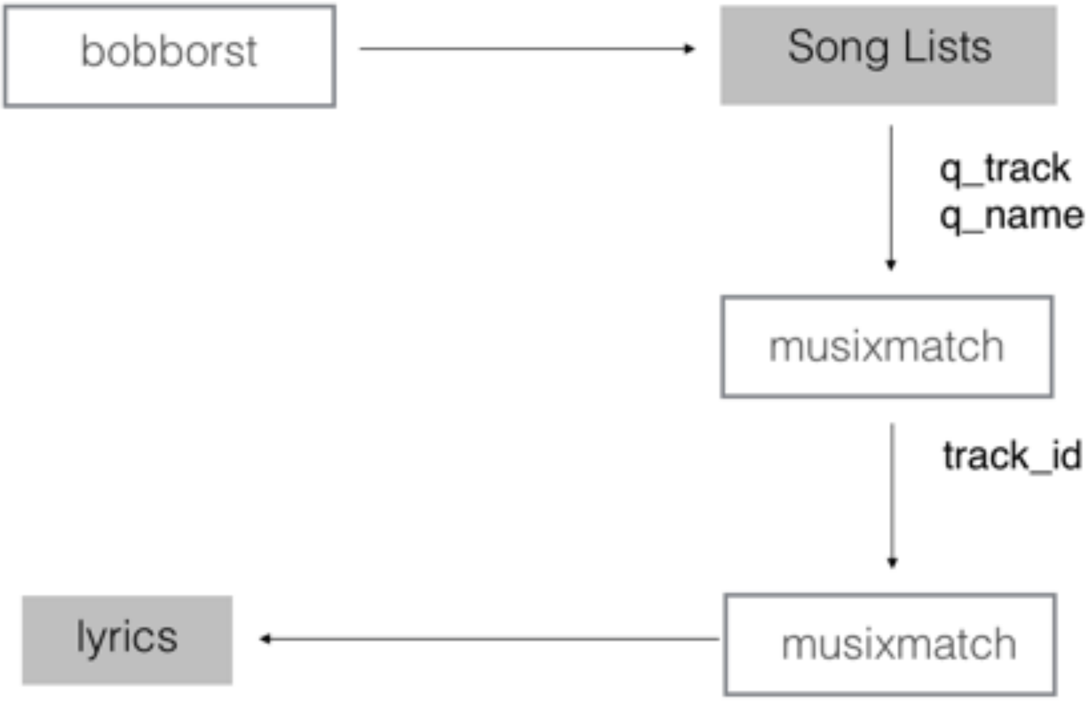
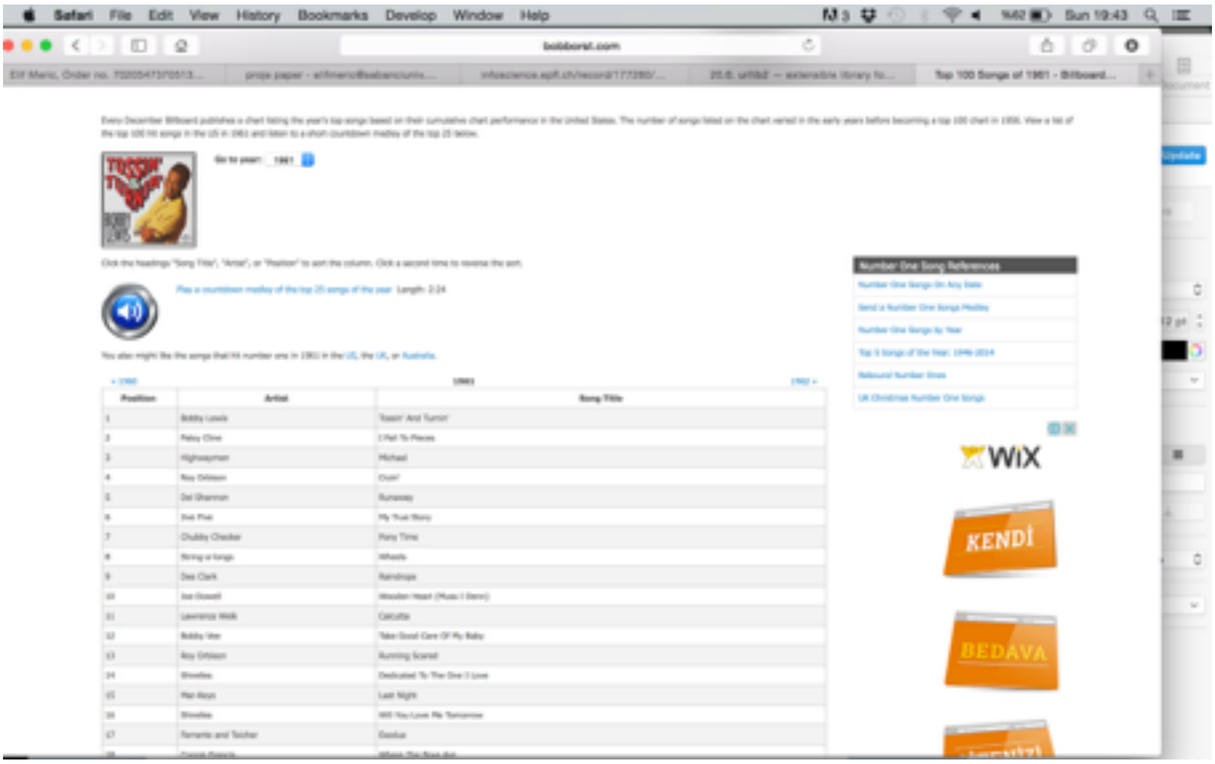


Figure 1 : Schematic of Data Collection Process

Since our objective to get lyrics between 1947-2014, first we found a website that contains a lists of top 100 : http://www.bobborst.com (Figure 2). This website provides us same top 100 list as official Billboard website but more flexible. This website is more easy to parse and get the datas.



Every December Billboard publishes a chart listing the year's top songs based on their cumulative chart performance in the United States. The number of songs listed on the chart varied in the early years before becoming a top 100 chart in 1956. View a list of the top 100 hit songs in the US in 1961 and listen to a short countdown medley of the top 25 below.

Go to year: 1961

Click the headings "Song Title", "Artist", or "Position" to sort the column. Click a second time to reverse the sort.

Play a countdown medley of the top 25 songs of the year Length: 2:24

You also might like the songs that hit number one in 1961 in the US, the UK, or Australia.

« 1960 1961 1962 »

Position	Artist	Song Title
1	Bobby Lewis	Tossin' And Turnin'
2	Patsy Cline	I Fall To Pieces
3	Highwaymen	Michael
4	Roy Orbison	Cryin'
5	Del Shannon	Runaway
6	Jive Five	My True Story
7	Chubby Checker	Pony Time
8	String-a-longs	Wheels
9	Dee Clark	Raindrops
10	Joe Dowell	Wooden Heart (Muss I Denn)
11	Lawrence Welk	Calcutta
12	Bobby Vee	Take Good Care Of My Baby
13	Roy Orbison	Running Scared
14	Shirelles	Dedicated To The One I Love
15	Mar-Keys	Last Night
16	Shirelles	Will You Love Me Tomorrow
17	Ferrante and Teicher	Exodus
18	Connie Francis	Where The Boys Are

Number One Song References

Number One Songs On Any Date

Send a Number One Songs Medley

Number One Songs by Year

Top 5 Songs of the Year: 1946-2014

Rebound Number Ones

UK Christmas Number One Songs

Safari File Edit View History Bookmarks Develop Window Help
Sun 19:43
bobborst.com
Top 100 Songs of 1961 - Billboard...
Update
WiX
KENDİ
BEDAVA

Figure 2 : Bobborst Top 100 / 1961

After list's collected, we need to find these songs' lyrics . Web services create APIs to external applications can collect them on websites. There are many useful websites for finding lyrics that provides API format like https://developer.musixmatch.com/. Musixmatch provides interface to get wanted data. These API methods provided by Musixmatch that we're going to use :

1. TRACK.SEARCH : we're going to use to search every song in our database.

Needed Parameters : *q_track* : track name, *q_artist* : artist name

track.search?q_track=*songname*&q_artist=*artistname*&f_has_lyrics=1

Since we've got list of song names and artist names from bobborst.com , we can search to find out if there is available lyrics. If response has available lyrics, we parse the json response to get track_id in order to get lyrics.

2. TRACK.LYRICS.GET: We're going to get the lyrics of the tracks on our lists.

Needed Parameters : *track_id*

track.lyrics.get?track_id=*trackid*

Since we've got list of track ids for each year's top 100 songs, we can finally get the lyrics from musixmatch.com . We stored lyrics year by year to search listed words.

3.2 Tools Used

Canopy : Enthought Canopy is Python development and analysis environment that provides easy scientific analysis and visualisation.

Beautiful Soup : Beautiful Soup is a Python library that designed for parsing HTMLs.

URLLIB2 : Library that defines functions and classes that make easier to open URLs.

GitHub : Website that where people build software and share. GitHub provides revision control system , that gives opportunity to control web-based projects development process and its history.

(ENS492 Project : **https://github.com/elifmeric/ENS492**)

4. Word Cluster and Concept Selection Process

Firstly, with considering the time interval we needed to define key concepts such as Sexism, Racism, Aggressiveness, Happiness, Consumerism, and Religiousness etc. These concepts are relatively easy to distinguish and determine in terms of words. For instance, “girl, bitch, slut, chick, hot, babe, fuck” these are the words in Sexism cluster. We picked these words because these are clearly sexist and make the songs sexist. We obtained necessary word information from several websites such as http://www.urbandictionary.com/define.php?term=sexism. After defining key concepts we are going to analyse our data in terms of ruling presidents of USA.

5. Further Progress

After implementation process, we created several graphs in order to distinguish frequencies and statistical results in terms of words and years. After having all lyrics of each year, we created a list which includes highlighted words and numerical results like words “girl, bitch, hot...etc”.

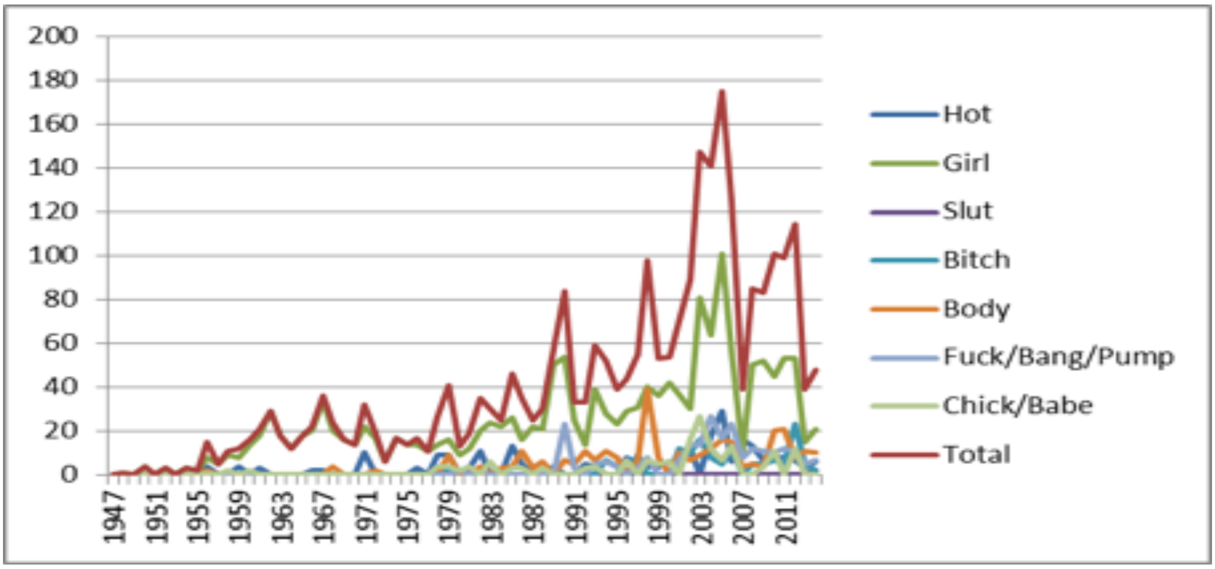


Figure 1

Figure 1 shows numerical usage data of each word in the lyrics between given years. This graph indicates that after 1955 sexist words began to take place in the lyrics and reached the pick point late 1990's and early 2000's. This is not a complete picture of all the words used in lyrics. Instead of that there are some selected words to be sure that there is a correlation between the years and number of sexist words used in music industry.

The other graph -which is shown below as Figure 2- is indicating the total number of sexist words take place in songs year by year. In the same way with first graph, this one also introduces the reality of increasing usage of sexist words especially after 1980's.

Bearing in mind that we do not have all the words to be searched in these graphs, there is still a big evidence of sexism among American song lyrics.

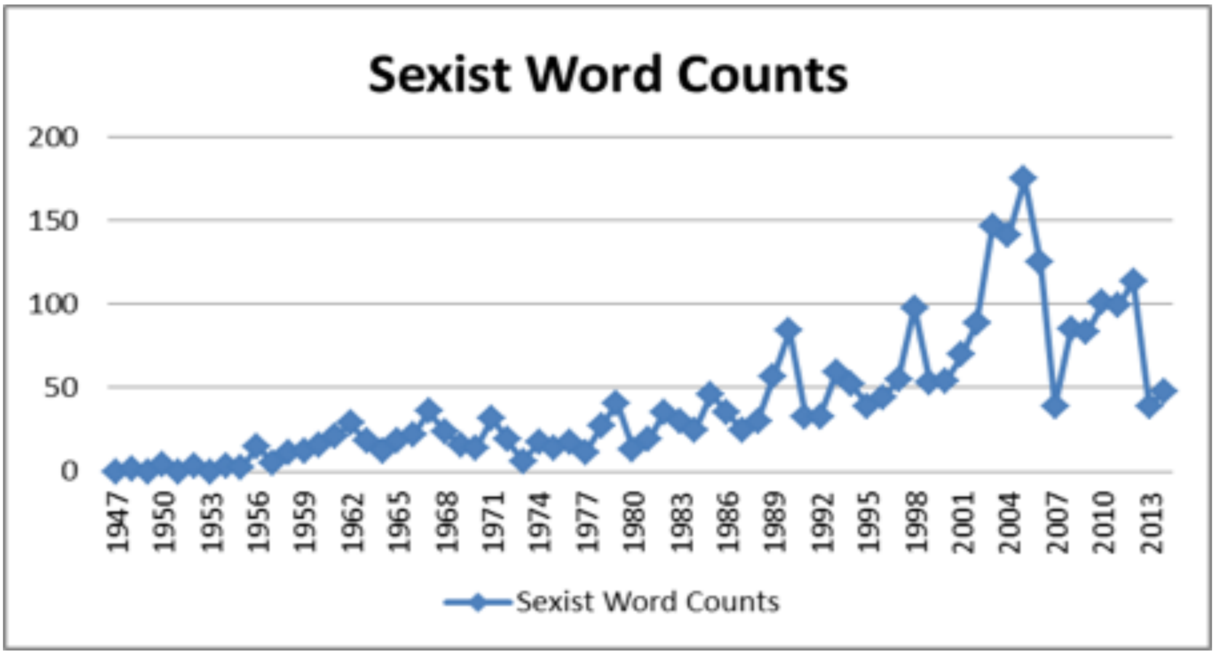


Figure 2

6. Conclusion

After getting all lyrics from the list, we created two graphs to clearly see the differences in frequencies between years. So we can clearly see that; after the beginning of 90's there is a significant increase in number of sexist words and there were almost zero sexist words until the beginning of 80's. After 80's there is increasing pattern until 2000's. After considering these significant changes in graphs we have decided to analyze these specified years. Between the years 2004-2007 the number of sexist words in songs reached its peak which is approximately 160 words in total.

In the future, we are going to add extra materials to existing word clusters about different concepts such as racism and we are going to analyze statistical results in terms of ruling presidents of USA. Main objective is finding correlations between dramatic changes and president's effects on popular song lyrics.

7. Time Plan

Essential Works For Project	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Getting General Information About Python	🟩	🟩	🟩					
Developing a Code For Lyrics				🟩	🟩	🟩	🟩	🟩
Creating a Words Cluster About Concepts				🟩	🟩	🟩		
Interpretation						🟩	🟩	🟩