SABANCI UNIVERSITY

ENS 491 Progress Report

Sarp TERLEMEZ / Elif MERİÇ / Hüseyin GÜVEN

İlker Bilbil

17.5.2015

Faculty of Engineering and Natural Science

ABSTRACT

In this project, we are going to analyze some popular music lyrics in terms of years. For this aim, we have decided to analyze Billboard magazine which has the music industry standard record chart in the United States for singles, published weekly by Billboard magazine. Chart rankings are based on radio play, streaming online and sales. We are going to analyze this Hot 100 chart and we are planning to reach these popular songs' lyrics. There are some key points for this process that we need to cover.

PROGRESS

For this project, we defined our procedure before the implementation part. Firstly, we need to get some knowledge about the MusiXmatch's API to reach the lyrics. API is a set of routines, protocols and tools for building software applications. An API indicates a software component in terms of inputs, outputs and underlying types which allows definitions and implementations to vary without compromising each other. Before collecting the data procedure, we need to understand API procedures, libraries and tools. After a research about the components like programmable web, a new programming language Python etc.

We decided to search for the correlation between lyrics and the concepts like consumerism, sexism and the religiousness in music industry in terms of USA presidents and their political influences.

Before the implementation part, we determined the procedures which are very essential for our sentiment analysis. Firstly, for data collection we need to access Billboard Hot 100 list among years so, first we need find a pattern between URLs of each year's documents

because our program should easily reach these websites and take the song names from official Billboard Hot 100 website https://www.billboard.com/biz/charts/year-end.

In addition to that, there are several websites which are very helpful for the access to most popular 40 songs between the years 1946-2014. We have decided to use this website not because it displays different Hot 40 information but because it's easier to parse the same information in more flexible manner. After finding a pattern between pages URLs, we need to store song and artist names in a data structure like array list to be able to query them in MusiXmatch website using the provided API. The pattern of the URLs is as follows.

http://www.bobborst.com/popculture/top-100-songs-of-the-year/?year=1946

http://www.bobborst.com/popculture/top-100-songs-of-the-year/?year=1947

***

http://www.bobborst.com/popculture/top-100-songs-of-the-year/?year=2014

After the creation of input collection, we need to find these songs' lyrics in API format. There are many useful websites for finding lyrics in API format like https://developer.musixmatch.com/. We are going to use this website to reach wanted songs' lyrics in terms of years. For this website, we are going to use this method to get wanted lyrics from the web. We need API key which will be provided from the previous website and we need track name which will be provided from the Bobborst.

http://api.musixmatch.com/ws/1.1/?apikey=APIKEY&?q_track=TRACKNAME&?f_has_lyrics=1

For this API provider website, we are allowed to reach only 2000 hits per day. However, we have 40 songs for each year, and we have 69 years, which gives us 2760 lyrics in each

execution of the program. Since the allowed number is less than this number, we might need to use an alternative method so we might need to learn the concept “web crawling”.

After the data analysis we are going to collect this data and create a statistical interference with using some visualization tools like histograms, bubble graphs etc. Before that we need to create a database of words which are related to the concepts for example;

- For sexism/racism concept: “chick, baby, Barbie, doll, etc.”

- For consumerism: “money, job, power etc.”

- For religiousness: “God, Jesus, bless, pray, heaven, hell etc.”

Finally we are planning to find some correlations between the lyrics and abovementioned concepts considering the presidential changes in USA. Finally; we have considered the previous researches and reached many visual results from the several websites for instance;

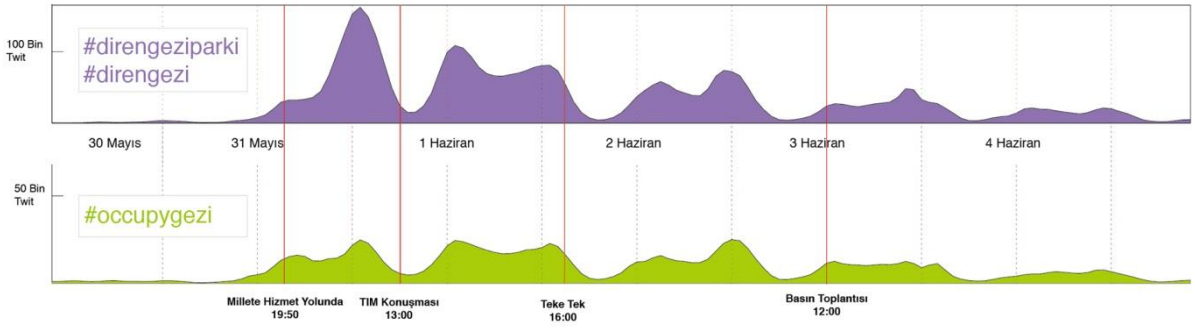

Figure 1

The figure above shows that the hashtag distribution of the tweets which are posted during the Gezi Park Protests against the government, chronologically.

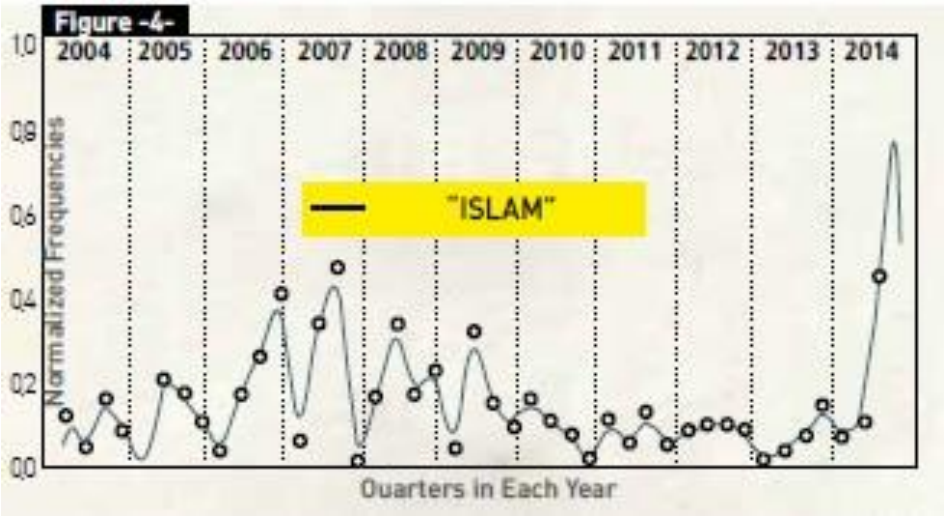
Figure -4-
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
1,0
0,8
0,6
0,4
0,2
0,0
Normalized Frequencies
"ISLAM"
Quarters in Each Year

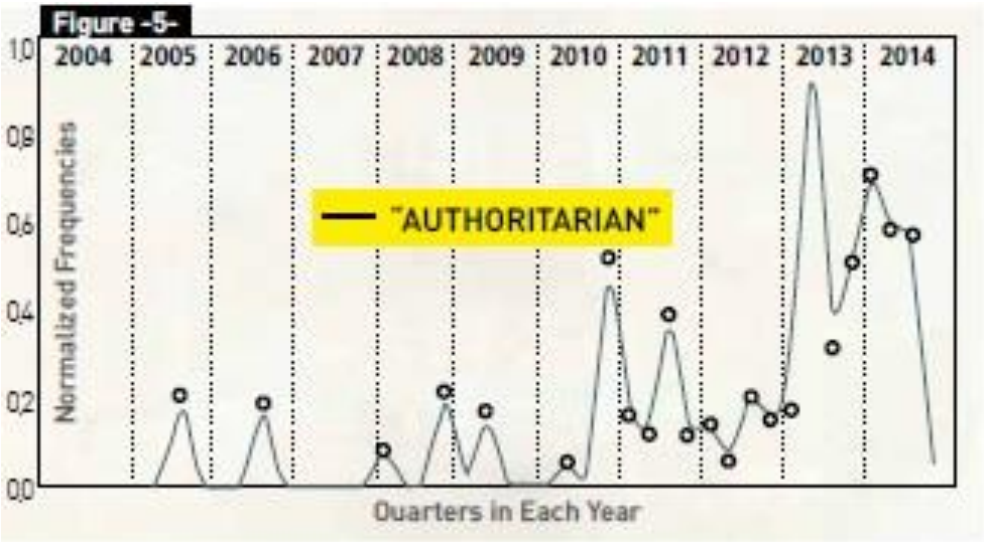
Figure -5-
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
1,0
0,8
0,6
0,4
0,2
0,0
Normalized Frequencies
"AUTHORITARIAN"
Quarters in Each Year

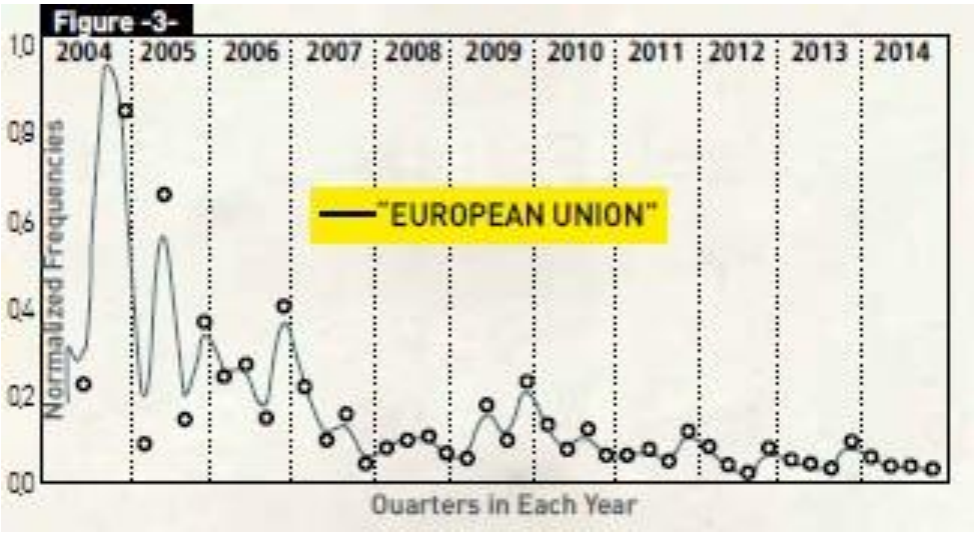
Figure -3-
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
1,0
0,8
0,6
0,4
0,2
0,0
Normalized Frequencies
"EUROPEAN UNION"
Quarters in Each Year

These graphs, which are taken from a previous work, indicate the sentiment analysis results from the articles about prime minister of Turkey, Erdogan, published in New York Times in terms of years. These results illustrate distributions of some specifically chosen words such as “authoritarian”, “Islam” and “European Union” chronologically.

We are planning to get these kinds of visual results, from our research with considering the presidents of USA.

CONCLUSION

With several techniques and tools that we are going to use in implementation part and as a group we are planning to reach accurate results about concepts like consumerism, sexism etc. and the sentiment analysis of Hot 100 song lyrics among years.

REFERENCES

1) http://www.radikal.com.tr/turkiye/twitter_millet_meclisinin_surpriz_siyaseti-1136382

2) http://www.hurriyetdailynews.com/erdogan-less-european-but-more-islamist.aspx?pageID=238&nID=73987&NewsCatID=396

3) http://dictionary.cambridge.org/grammar/british-grammar/sexist-language