Assignment 1: Analyzing and Tracking the Sentiment and Topics on Social Media

By

Mamun Mohammad (s3571301)

Table of Contents

Introduction:

Being an employee of Easy Analytics, I have assigned to analyze the sentiment and topics in social media of a reputed singer in the world known as “Katy Perry”. Katy Perry is an American singer, songwriter, and television judge. She uses social media to be connected with her fans worldwide. She is very interested to know what her followers are saying about her online what their feeling are towards image.

Since she is a celebrity there are always tweets about her and she has the most followers in the world followed by Justin Bieber and Barak Obama.

Lately, she has become one of the judges in American Idol and there are lot of good or bad comments about her in the twitter. So, my role will be analyzing sentiment and also have to do topic modelling collecting tweets from her account.

Data Collection

To collect the data, I have used Rest API for mining 10000 tweets from Katy Perry Twitter Account. I have used Rest API to get the data in a very short period, while the streaming API requires few days of live connection to retrieve the data

There are 2700 number of tweets have been collected from Katyperry’s twitter account.

The top 20 hashtags were collected using tweepy package in python and the hashtags are as below are as below:

Hashtags	occurences	Hashtags	occurences
#americanidol:	226	#rio2016:	12
#katyperry:	57	#witnesskp:	11
#photo:	43	#video:	8
#witnessthetour:	35	#katyperrybkk2018:	7
#win:	24	#spiritcooking:	7
#Follow:	23	#soitgoes:	5
#witness:	22	#mindmaze:	5
#update:	15	#witnessthetourmnl:	5
#rise:	15	#arianagrande:	4
#taylorswift:	12	#nickiminaj:	4

From this hashtag, table we can see that the most hashtags are energy, #americanidol:, #katyperry:, #photo:, #witnessthetour:.

The output for user mentions are as below:

Mentions	Occurences	Mentions	Occurences
katyperry:	2304	katydailybrasil:	123
musicnewsfact:	502	PopCrave:	122
ladygaga:	486	SenateMajLdr:	110
Zedd:	439	Adele:	89
yoderproduction:	164	kattyperryfam:	77
katyspics:	139	britneyspears:	69
BrunoMars:	134	LukeBryanOnline:	65
rihanna:	130	realDonaldTrump:	63
Drake:	126	BarackObama:	63
chartdata:	123	NancyPelosi:	61

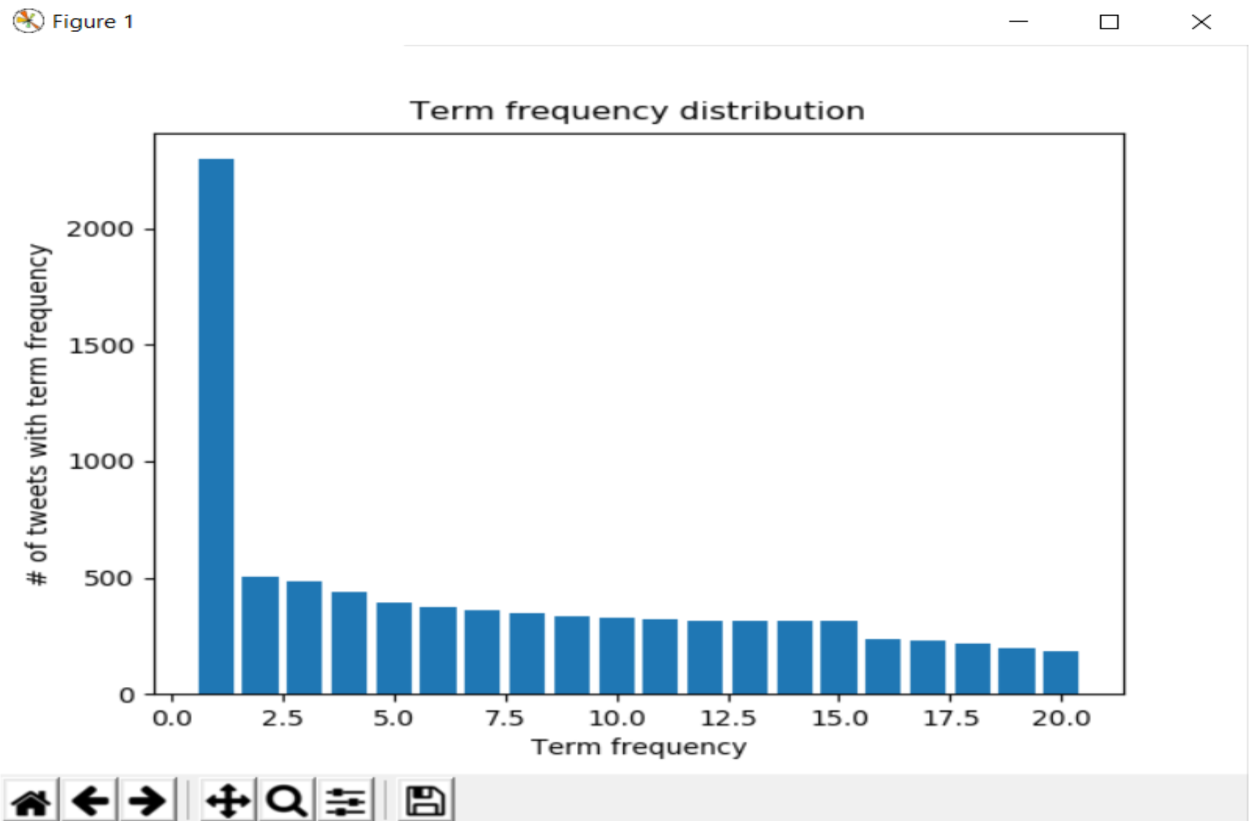
From the table we can see that how follwers are mentioning katyperry in her twitter timeline

The most common words from the collected data are as below:

Common words	occurences	Common words	occurences
@katyperri:	2296	way:	322
@musicnewsfact:	502	everybodi:	316
@ladygaga:	486	romanc:	315
@zedd:	438	calm:	313
love:	393	hot:	235
stan:	376	instagram:	232
gaga:	360	⍰:	215
song:	344	live:	196
thi:	332	kati:	185
bad:	325		

Here, from the output we can see @katyperri:, @musicnewsfact: @ladygaga: @zedd: these are the most common words Katyperry's followers used to retweet her.

The term frequency distribution is as below which shows how important these words are within the data collected

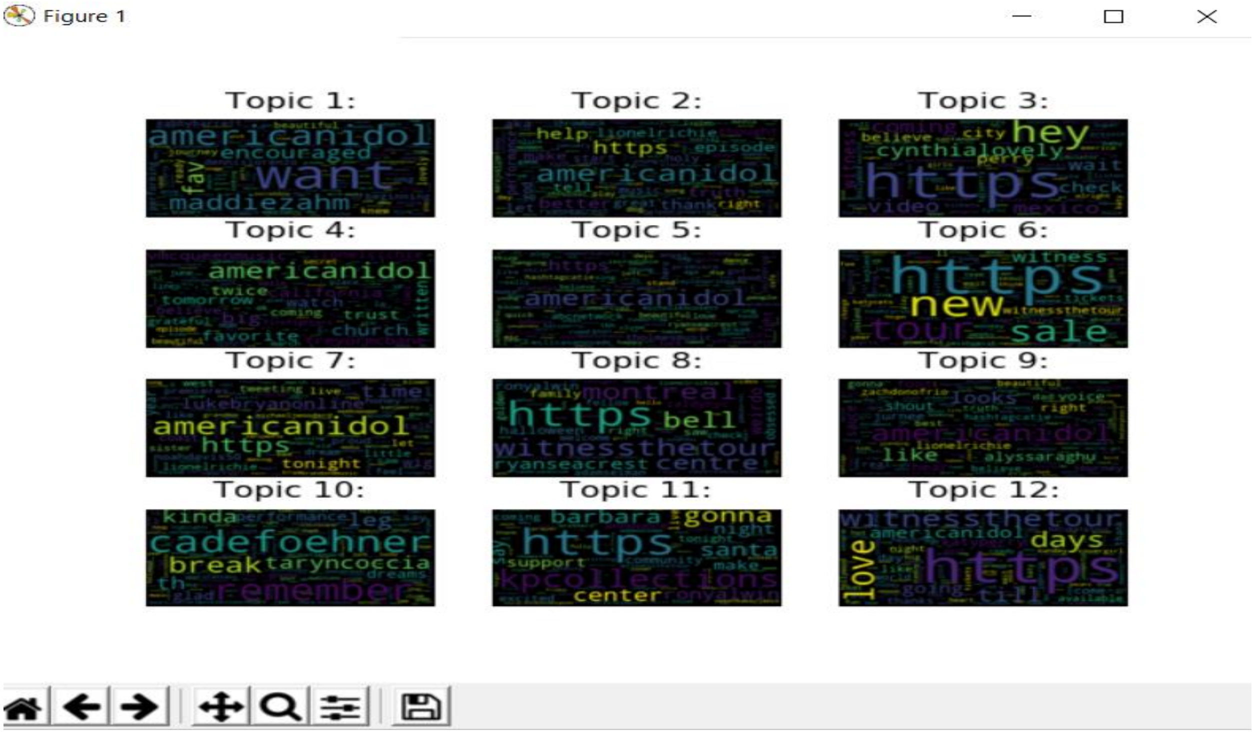


Pre-processing and Data Cleaning

After run through hashtags and using a package called word cloud, I have found out there are some foreign languages, texts and characters are exist in the data set which need to be cleaned to do further analysis.

Foreign Lanuages in twitter data

케이티페리:	3
ケイティ・ペリー:	3
케이티페리:	3



Type of preprocessing to perform:

I have used text preprocessing using nltk package, after I do preprocessing the most common words I have found here as below

Common words	Frequency
@katyperri:	2296
@musicnewsfact:	502
@ladygaga:	486
@zedd:	438
love:	393

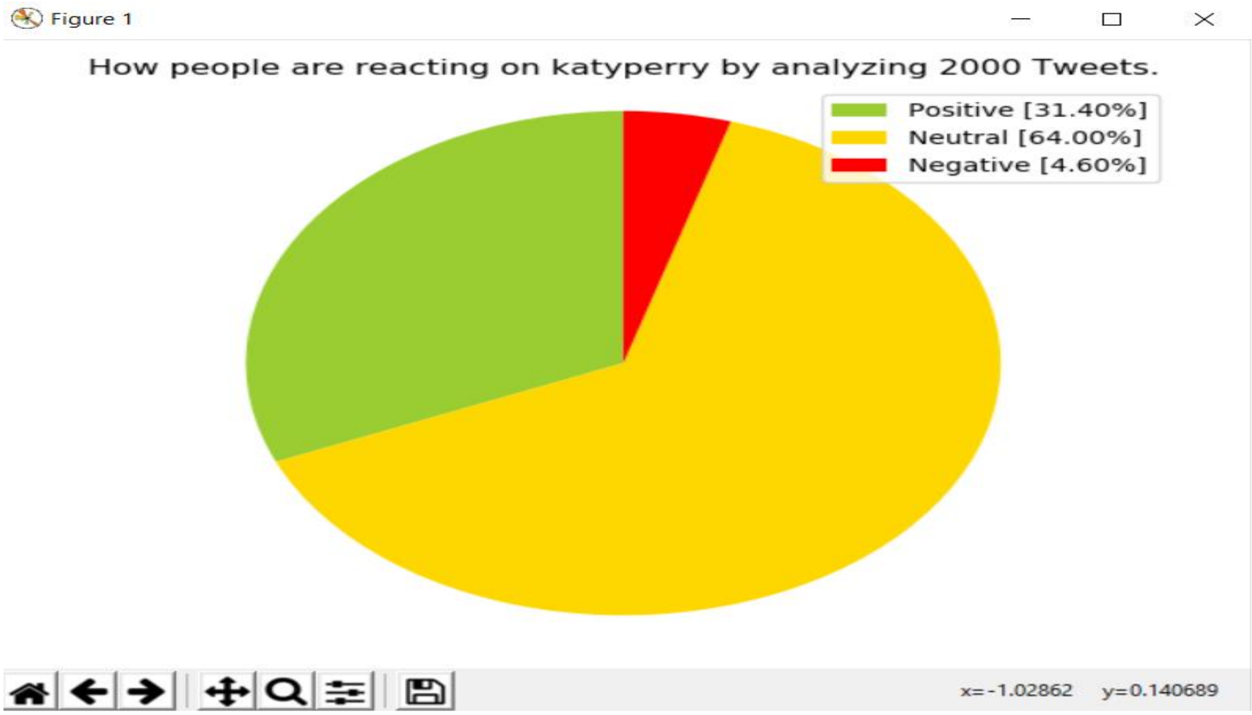
stan:	376
gaga:	360
song:	344
thi:	332
bad:	325
way:	322
n:	317
everybodi:	316
romanc:	315
calm:	313
hot:	235
instagram:	232
live:	196
kati:	185

Analysis Approach

Sentiment analysis—

I have used “textblob” package to do the sentiment analysis for 2000 tweets from Katy Perry’s twitter account. The strategy was to count every word and figure out whether that was negative, positive or neutral using the built in dictionary in “ texblob” package.

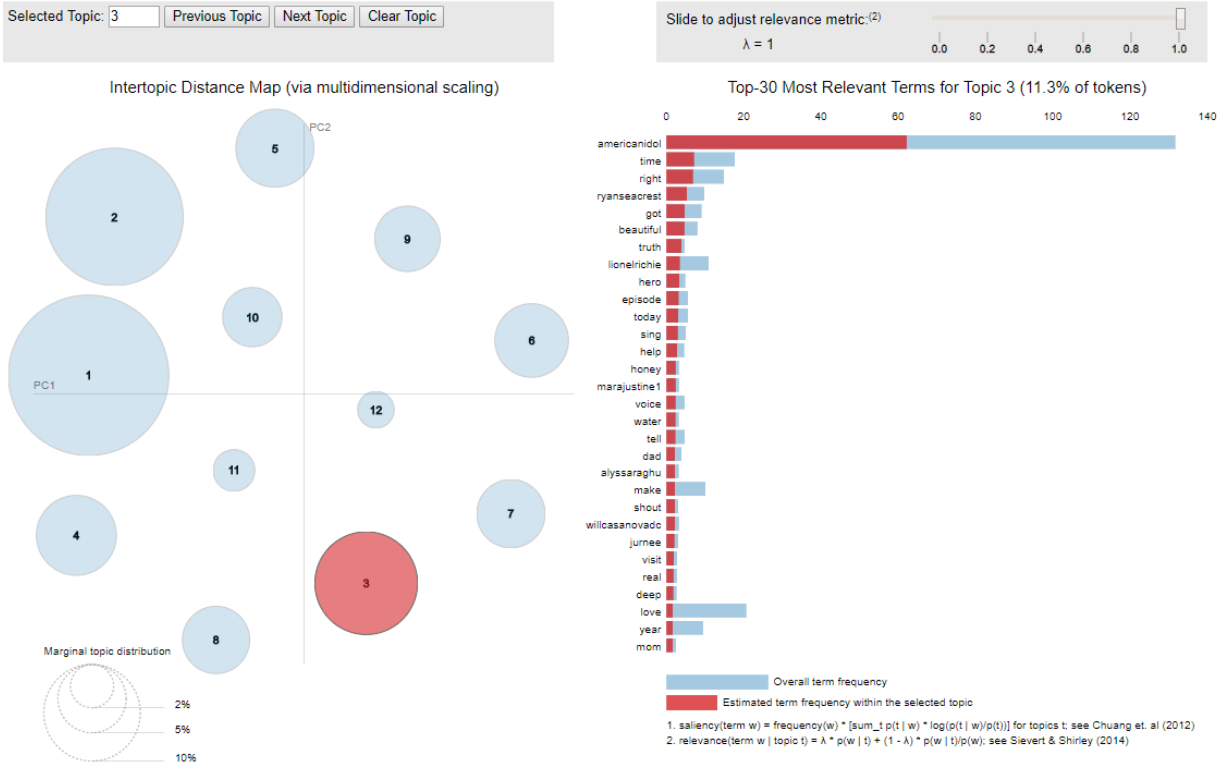
The output in pie chart is as below



Topic Modelling—

In my analysis I have used LDA model for topic modelling because LDA is a probabilistic model with interpretable topics. Since LDA gives categories for free.In this topic modelling I have taken 12 topics to analyze and getting the popular word for each topic.

The output from the topic modelling is as below



Selected Topic: 3 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 3 (11.3% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

From this topic modelling output we can see the most relavent terms

Analysis and insights

After analyze the tweets from Katy Parry's twitter account I have found there are less number of tweets considered as negative compared to positive tweets.Surprisingly.it can be seen that more than half of the words in twitter data was neutral.

Conclusion:

Katy parry is a renowned singer in America, but she has followers all over the world which makes her top famous personality in twitter till today. There are positive and negative tweets about her.

The analyses would be even better if the data can be more cleaned.