Data Analysis and Mining For Twitter Using R

Presented by:
Athanasiadou Maria

Under the guidance of
Dr. Koloniari Georgia

Department of Applied Informatics
University of Macedonia, Greece

June 25,2019

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

- Social structure of nodes-individuals or organizations

  that are linked to specific types of relationships:

  -friendship
  -affinity
  -knowledge
  -common interests

- A way of social networking is through social networking sites.
- Social networking sites are web services that allow people to create a public profile.
- Social data includes a variety of information and they constitute valuable source for analysis.

EXAMPLES OF SOCIAL NETWORKING SITES ARE:

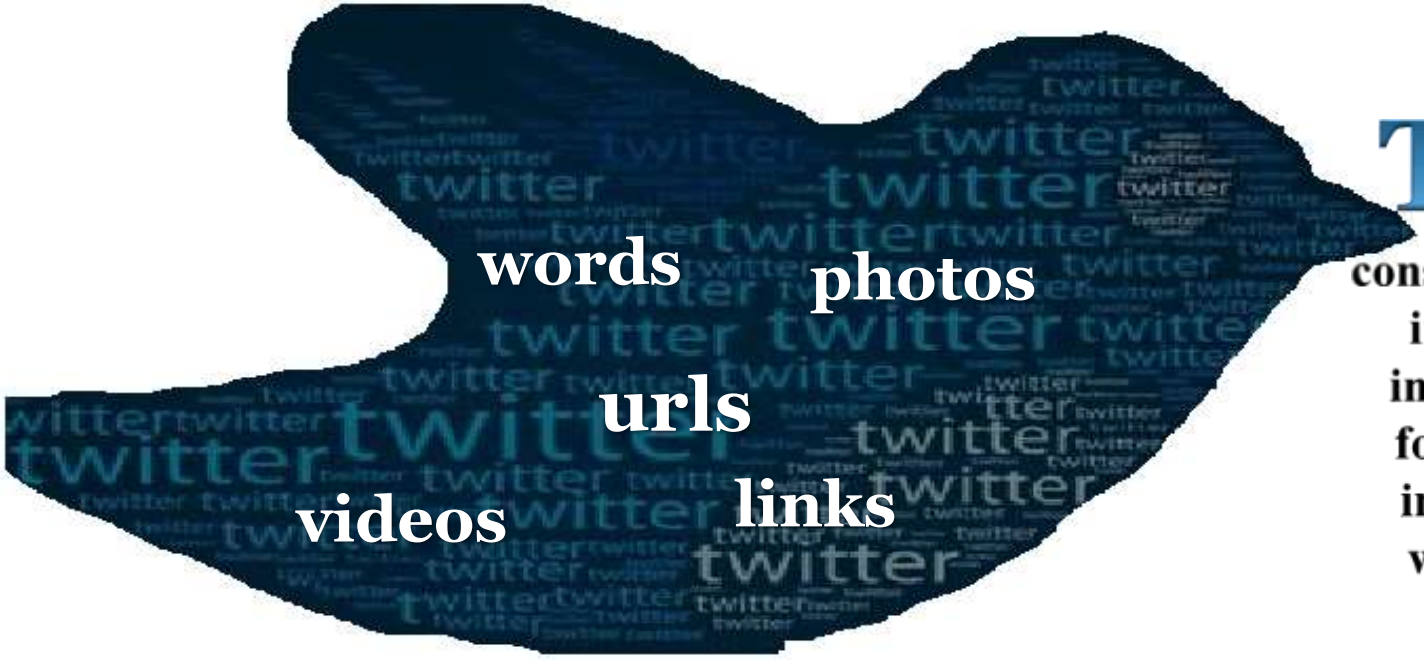
OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

Twitter

is one of the most popular online social networking and microblogging tools.

is described as “SMS of the Internet "



Tweets

consists of 140 characters in which users post instant messages, as a form of expression to interact with anyone within the network.

TWITTER TERMINOLOGY

#- Hashtag:

any word or phrase preceded by the # symbol, clicking it you can see tweets containing the same topic.

@ -Reply/Mention:

the @ symbol followed by a user in a tweet enables the direct delivery of the tweet to that user.

RT - Retweet:

the RT symbol in a tweet indicates that someone else's content is being re-posted.

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- **What is Twitter API?**
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

- Twitter allows you to interact with its data tweets using Twitter APIs.

- You can gather tweets with two possible methods:

  -Streaming API and

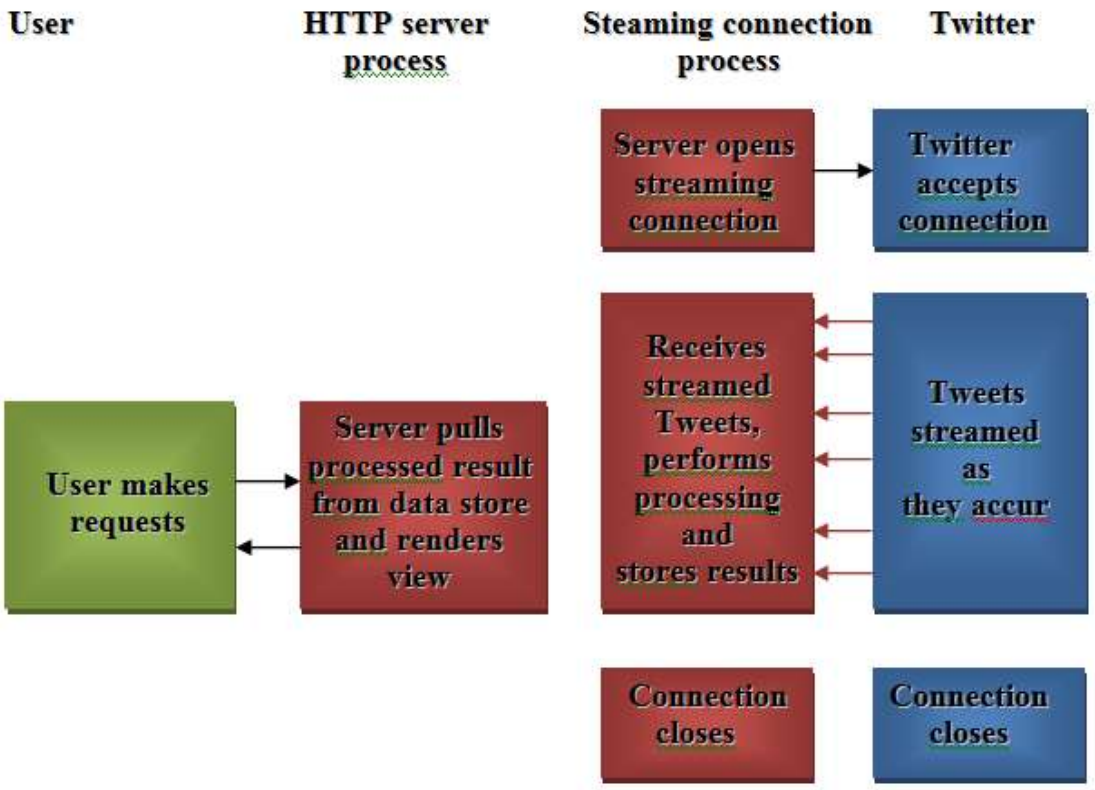
  -Search API

Their difference are in design and in the way they access Twitter data.

- Twitter APIs can be accessed only via authenticated requests using OAuth.

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- **Streaming API**
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work



- Streaming API →continuous real-time streams of tweets.
- Limitation: 1% sample of Twitter data(1 million tweets a day)

➢ Using the Streaming API method you can access data from an input query and fetch the search results as a stream. You can search for keywords, hashtags, userids.

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- **Goal of the thesis**
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

❖ To Export and exploit useful information from Twitter.

➢ Text mining techniques:

- text classification on Naive Bayes and Support Vector Machine
- text clustering on K-means, Latent Dirichlet Allocation

➢Exploratory data analysis

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- **Methodology**
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

Procedure for building a Twitter app:

1) Create a Twitter account at http://twitter.com
2) Create the Twitter app at http://apps.twitter.com

Python programming language:
- Collection, extraction and storing of the tweets
- Format of tweets: JavaScript Object Notation (JSON)

Atributes	Description
Created_at	The UTM time when this tweet was created
Text	The actual UTF-8 text of the status update
Screen_name	The screen name or alias that this user identifies themselves with
Location	The user-defined location for this account's profile
Followers_count	The number of users this account is following
Friends_count	The number of users this account is following, also known as their "followings"
Favourites_count	The number of Tweets this user has liked in the account's lifetime
Statuses_count	The number of Tweets (including retweets) issued by the user
Time_zone	A string describing the Time Zone this user declares
Lang	Indicates a BCP 47 language identifier corresponding to the machine-detected language of the Tweet text
Retweet_count	The number of times this tweet has been retweeted
Favorite_count	The number of times this tweet has been liked by Twitter user's

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- **Overview of R**
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

➢R is a system for statistical analyses and graphics

Basic features:

- Open source & free

- Interacts with other programming languages (Python, C/C++, Java)

- Runs on almost any standard computing platform and operating system (tablets, phone, game consoles)

- Includes over 15.000 libraries

- Is used among statisticians and data miners for developing statistical software and data analysis.

➢ Main packages: 1. tm→for text mining applications, 2.ggplot2→ for data visualization

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- **Dataset Description**
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

- Data collection: December 30, 2016 -Feb 04, 2017
- Subject of interest: Music genres
- Keywords: "classical music", "folk music", "pop music", "rap music", "rock music".

Dataset Statistics:

1º Dataset for Exploratory Data Analysis

2º Dataset for Text Mining

Music genres	#of Tweets	# of Tweets
Classical music	3415	2474
Folk music	2090	1500
Pop music	4243	2441
Rap music	3606	2319
Rock music	3446	1849
	16800	10583

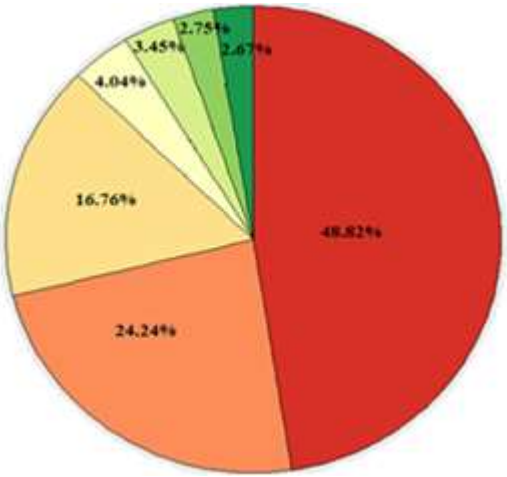
OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- **Statistics of Twitter Data**
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

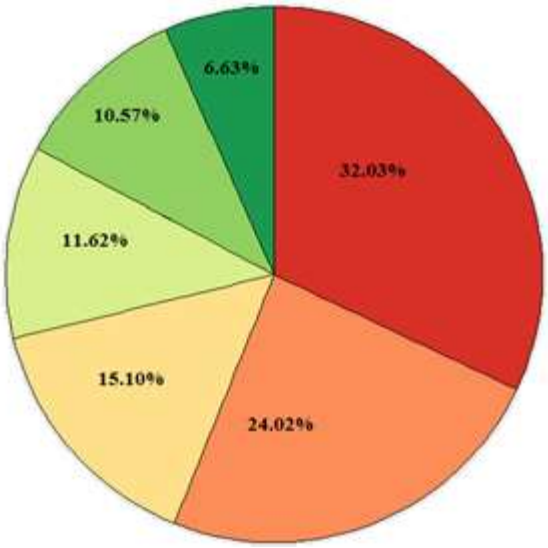
Measures of Location

	Min.	Median	Mean	Max.	Mode
Followers_count	0	531	11940.13	28715815	0
Friends_count	0	484	4090.18	5430083	0
Favourites_count	0	1445	10428.19	3555686	0
Statuses_count	1	6670	78040.9	164504281	4
Retweet_count	1	83	1951.56	76915	1
Favorite_count	0	90	2834.23	139460	0

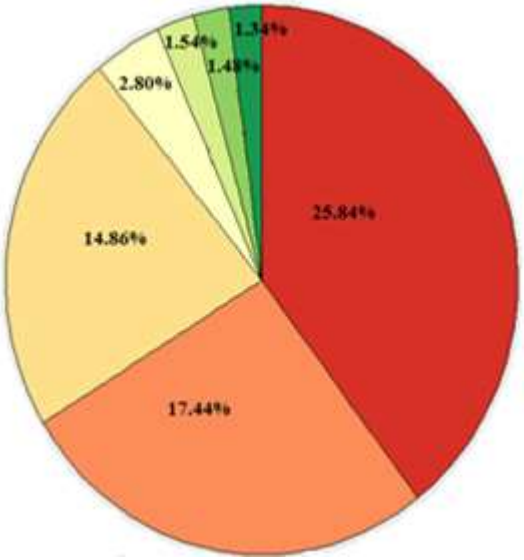
FOLLOWERS-FRIENDS-TWEETS OF USERS



- 48.82% users have less than 500 followers.
- 3.45% users have more than 20000 followers.
- Top user: 28715815 followers.



- All users have tweeted by a tweet.
- 6.63% of users have made less than 1000 tweets.
- 32.03% of users have made over than 20000 tweets.

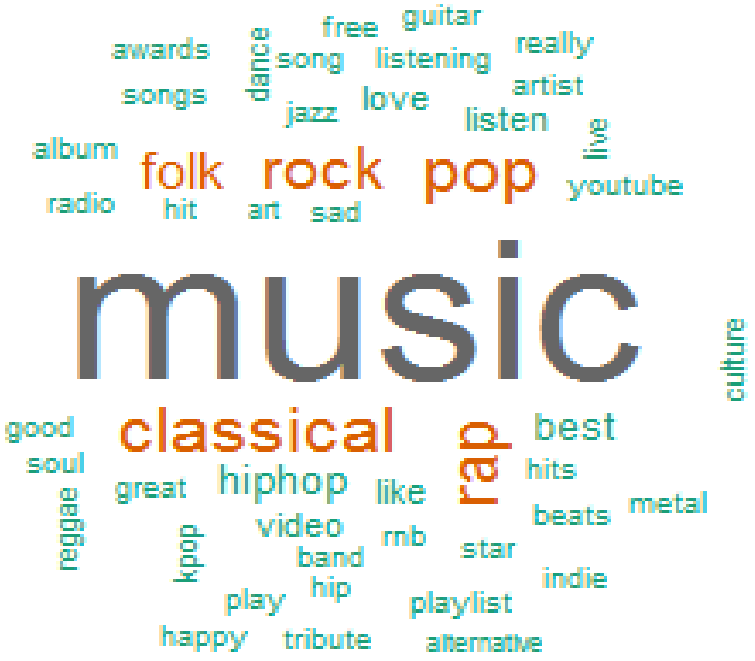


- 1.48% of users have no friends.
- 25.84% users have less than 5000 friends.
- 1.54% users have more than 20000 friends.
- Top user: 5430083 friends

• High no.of users have few followers & friends in contrast to the tweets they post.

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- **Visually overview of a text**
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work



Word	Frequency
music	9950
classical	2356
pop	2256
rap	2211
rock	2001
folk	1350

song													
songs	**rnb**	**love**	**pop**	**jazz**	**folk**	**hiphop**	**new**	**year**	**dance**	**good**	**listen**	**rap**	**this**
0.10	0.09	0.08	0.08	0.07	0.06	0.05	0.05	0.05	0.02	0.02	0.02	0.01	0.01

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- **Text preprocessing**
- Text classification
- Text clustering
- Conclusions- Future Work

```
xCorpus<-VCorpus(VectorSource(df1$text))
clean.corpus <- tm_map(xCorpus,content_transformer(tolower))
myStopList=read.table("C:/Users/pc/Desktop/stop.txt",header=FALSE,sep="\n",strip.whit
e=TRUE)
 clean.corpus<-tm_map(clean.corpus,removeWords,myStopList[,1])
 clean.corpus<-tm_map(clean.corpus,removeWords,stopwords("english"))
removeUsersNames<-(function(x)gsub("@\\w+", "", x))
clean.corpus<- tm_map(clean.corpus,content_transformer(removeUsersNames))
removeURL0<- (function(x) gsub("(f|ht)tp(s?)://\\S+", "", x, perl=T))
clean.corpus<-tm_map(clean.corpus,content_transformer(removeURL0))
clean.corpus <- tm_map(clean.corpus, removePunctuation)
```

Text before cleaning	Text after cleaning
I always enjoy discovering new music that ain't pop	always enjoy discovering new music aint pop
love this song https://t.co/yfSKnmUQjH	love song
@IAmLindsayJones the best classical music	best classical music

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- **Text classification**
- Text clustering
- Conclusions- Future Work

NAÏVE BAYES

- stemming
- Term frequency(tf)
- 10-cross validation
- 77 seconds

75%training dataset(7937 tweets)
25%test dataset (2646 tweets)

CA	Precision	Recall	F1	Kappa	Specificity	AUC
0.881	0.879	0.881	0.878	0.851	0.894	0.887

Naive

	Recall	Precision
classical_music	**0.930**	**0.914**
folk_music	0.901	0.873
pop_music	0.829	0.861
rap_music	0.903	0.906
rock_music	0.843	0.840

SVM

- kernel→radial
- stemming
- Term frequency-inverse document frequency(tf-idf)
- 10-cross validation
- 103 seconds

CA	Precision	Recall	F1	Kappa	Specificity	AUC
0.852	0.867	0.833	0.852	0.813	0.962	0.897

SVM

	Recall	Precision
classical_music	**0.917**	0.865
folk_music	0.80	**0.958**
pop_music	0.880	0.760
rap_music	0.863	0.894
rock_music	0.759	0.854

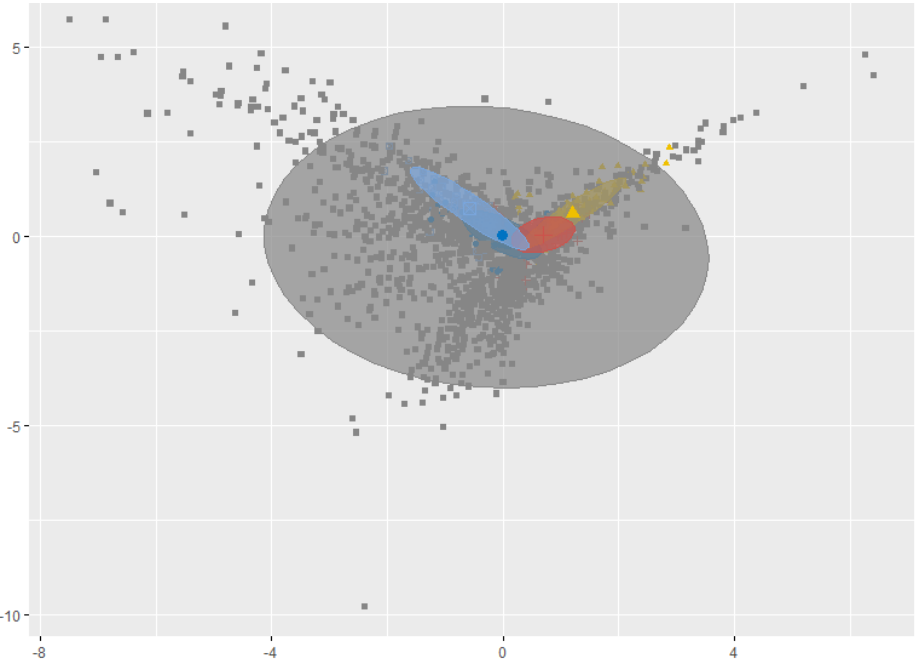
OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- **Text clustering**
- Conclusions- Future Work

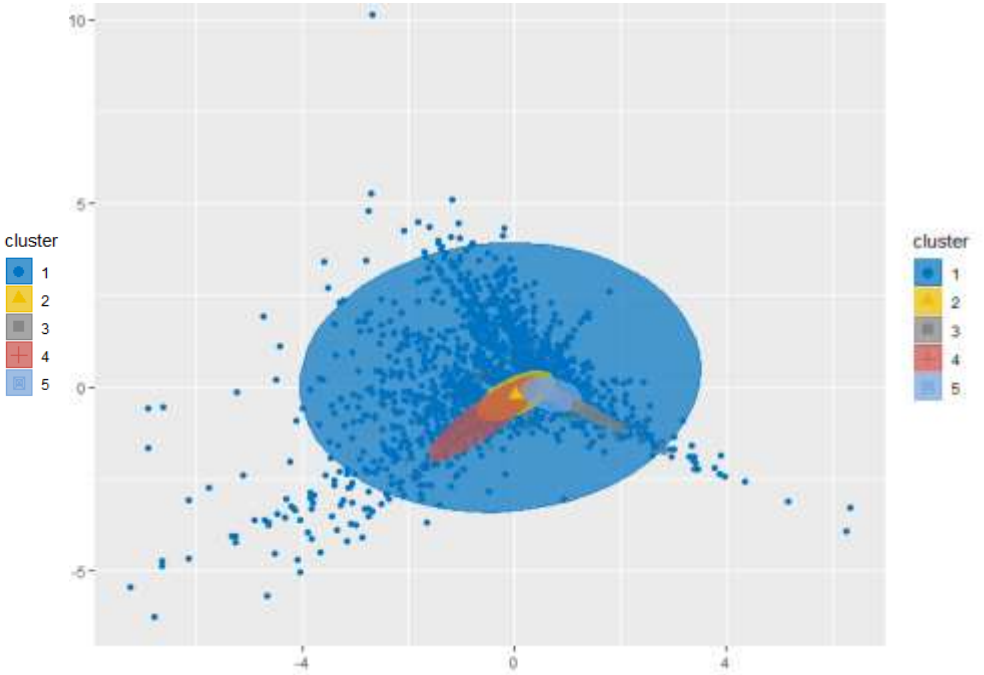
K-MEANS

- Jaccard Distance

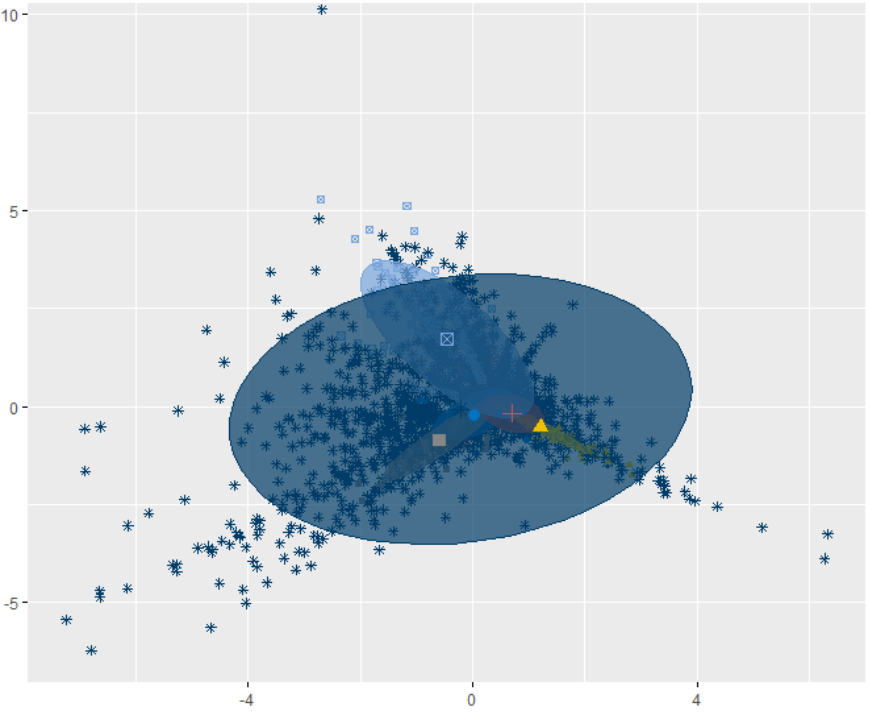
k=5,without stemming



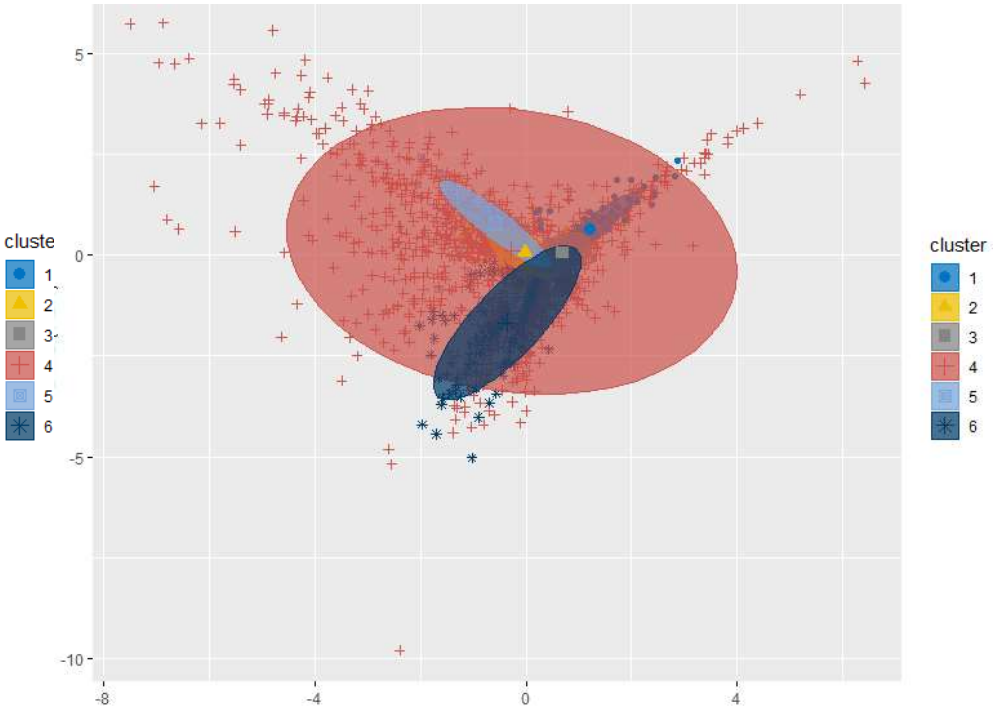
k=5,with stemming



k=6 with stemming



k=6 without stemming



LDA

Top 10 terms of each topic

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
rock	music	music	rap	pop
music	folk	classical	music	music
like	free	sad	hiphop	best
great	live	listening	video	love
radio	album	art	youtube	dance
good	awards	piano	star	rnb
band	guitar	track	playlist	jazz
tribute	acoustic	king	play	songs
indie	concert	amazing	beats	hits
metal	blues	friends	artist	soul

Topic probabilities by document

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
9437	**0.359**	0.156	0.156	0.156	0.171
7750	0.142	0.142	0.171	**0.2**	**0.3**
7077	0.190	**0.333**	0.158	0.158	0.158
10333	**0.241**	0.151	0.181	**0.242**	0.181
8380	0.161	0.161	**0.354**	0.161	0.161
10575	**0.272**	0.196	0.151	**0.212**	0.166
2932	0.159	0.159	0.159	**0.376**	0.144
7095	**0.242**	**0.227**	0.196	0.151	0.181
7067	0.158	0.174	0.158	0.158	**0.349**
2932	0.159	0.159	0.159	**0.376**	0.144

OUTLINE

- What are Social Networks and Social Data?
- What is Twitter
- What is Twitter API?
- Streaming API
- Goal of the thesis
- Methodology
- Overview of R
- Dataset Description
- Statistics of Twitter Data
- Visually overview of a text
- Text preprocessing
- Text classification
- Text clustering
- Conclusions- Future Work

- Best text classifier→ Naïve Bayes Algorithm

  Improvements in text classification algorithms:
  Naïve Bayes→ use of Laplace smoothing parameter
  SVM→ use of cost and gamma parameters

- Poor clustering → k-means & LDA

  It is observed that if words with similar content are used in data-selection filters, the algorithms cannot perform proper grouping of clusters.

- Built a web-based application appropriate for text classification or text clustering using some of the algorithms that we studied.

  - Shiny package

THANK YOU!