Twitter text mining with segmentation using R

Tomaž Kaštrun, Spar ICS GmbH, Spar Slovenija d.o.o.

Budapest BI Forum 2014

26.11.2014
INTRODUCTION

- BI Developer (MSSQL Server, C#, SAS, R, SAP, Py)
- 10+ ys experience databases
- 15+ ys experience data analysis and DM
- Working: Spar ICS Österreich, Spar Slovenija d.o.o.
- MCPT, MCT SQL Server
- tomaz.kastrun@gmail.com
- @tomaz_tsql
- http://tsqljokes.tumblr.com/
- Publishing articles, speaking at SQL events
- Coffee Lover, Fixie bikes junkie
AGENDA

- Creating credentials and twitter certificate
- Connecting to twitter account (@tomaz_tsql) using R
- Retrieving and storing twits
- Twitter text mining (with data visualization)
- Segmenting twits and segmenting followers (with data visualization)
- Presenting and visualizing data
AGENDA

- Creating credentials and twitter certificate
- Connecting to twitter account (@tomaz_tsql) using R
- Retrieving and storing twits
- Twitter text mining
- Segmenting twits and segmenting followers
Credentials & Certificate

- Install necessary R Libraries
- SSL CA Certificate
- Install Twitter Library
  - `install.packages("twitteR")`
  - `install.packages("plyr")`
  - `install.packages("stringr")`
  - `install.packages("ggplot2")`
  - `library(twitteR)`
  - `library(ROAuth)`
  - `library(plyr)`
  - `library(stringr)`
  - `library(ggplot2)`
AGENDA

- Creating credentials and twitter certificate
- Connecting to twitter account (@tomaz_tsql) using R
- Retrieving and storing twits
- Twitter text mining
- Segmenting twits and segmenting followers
Connecting to twitter account

- Establish secure connection

```r
# 2. Getting credentials and Twitter certificate

# Before
# Go to: https://apps.twitter.com/
# Login with your account
# Select Create New App
# Fill In: Application details
# OwnerID: 379167308
# Customer API Key: Tvbmz2zGFIuKKnv9iGHEQ1Ly0
# Customer API Secret: D32q3pdUdS5qFwu0zaq2Hjqyq4bd4AtGnD9sPMcmtJ3Kr77

requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL = "https://api.twitter.com/oauth/access_token"
authURL = "https://api.twitter.com/oauth/authorize"
consumerKey = "Tvbmz2zGFIuKKnv9iGHEQ1Ly0" #My Twitter account - generate id on apps.dev Twitter site
consumerSecret = "D32q3pdUdS5qFwu0zaq2Hjqyq4bd4AtGnD9sPMcmtJ3Kr77" #

twitCred <- OAuthFactory$new(consumerKey=consumerKey,
                             consumerSecret=consumerSecret,
                             requestURL=requestURL,
                             accessURL=accessURL,
                             authURL=authURL)
```
AGENDA

- Creating credentials and twitter certificate
- Connecting to twitter account (@tomaz_tsql) using R
- Retrieving and storing twits
- Twitter text mining
- Segmenting twits and segmenting followers
Retrieving and storing data

- Retrieving data (extracting tweets)
- Clean data (ahaaaaa!!!! DO IT!)
- Store data (CSV, Database, XML, binary file, ...)

Budapest BI Forum 2014

#budapestBI
AGENDA

- Creating credentials and twitter certificate
- Connecting to twitter account (@tomaz_tsql) using R
- Retrieving and storing twits
- Twitter text mining (with data visualization)
- Segmenting twits and segmenting followers (with data visualization)
Twitter text mining

- Unstructured text data
- Text categorization
- Text clustering
- Entity extraction
- Sentiment analysis
- Document summarization
- Behaviour prediction (based on text analysis)
- Text filtering (e.g.: spam)

Anybody noticed..? progress bar is not working? 😊
Word or two about Twitter...

- Twitter basic concepts
  - Following -> People on twitter I read their twitts (my „friends“)
  - Followers -> People on twitter that read my tweets
  - My twitts -> What I am writing
Case 1 – analysing text

- Microsoft PASS SUMMIT 2014
- Seattle, 4-7 November 2014
- MVP Event
- My statistics
  - 7 days of tracking twitts with hashtag #sqlpass or #summit2014
  - On general from 500-1500 tweets per 24/hours
  - Retrieving tweets, analysing tweets and doing visualization using R
  - Just analysis of text / text mining; text <- tweets per 24/hours
Case 1 – analysing text

2nd of November 2014
Huh.. Are we there yet?

Looking around the venue

Wii.. I’m registered!

Beer anyone?

registration behindthescenes

summit14 sqlpass
Case 1 – analysing text

4th of November 2014

Pre-cons!!!

Jeff, It’s been long!

Seattle, Seattle...

Beer tonight?

Budapest BI Forum 2014  #budapestBI
Case 1 – analysing text

4th of November 2014 – The night!

- Karaoke!!!
- Riding a bull!
- Party...
- Cowgirls
Case 1 – analysing text

4th of November 2014 – The night! And into the 5th....

Karaoke!!!
Riding a bull!
Party...
Cowgirls
Case 1 – analysing text

5th of November 2014

Keynote!!!

Sessions

sqlserver

query

azure

sqlpass summit14
Case 1 – analysing text

Snippet of R Code

```r
# SQL PASS WORD CLOUD from Twitter
sqlPass2014 <- searchTwitter(`"#summit14"`, n=800, caInfo="cacert.pem")
sqlpass2014_table <- do.call("rbind", lapply(sqlPass2014, as.data.frame))

sqlpass2014_table_last <- sqlpass2014_table[1:800,] # set number of rows to retrieve
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]
sqlpass2014_table_last = sqlpass2014_table_last [, -2]

r_sqlpass2014_text <- sapply(sqlpass2014_table_last, function(x) x$getText()) # r_sqlpass2014_text <- sqlpass2014_table_last
r_sqlpass2014_text_corpus <- Corpus(VectorSource(r_sqlpass2014_text))
r_sqlpass2014_text_corpus <- tm_map(r_sqlpass2014_text_corpus, tolower)
r_sqlpass2014_text_corpus <- tm_map(r_sqlpass2014_text_corpus, removePunctuation)
r_sqlpass2014_text_corpus <- tm_map(r_sqlpass2014_text_corpus, function(x) removeWords(x, stopwords()))
library(RColorBrewer)
pal2 <- brewer.pal(8, "Dark2")
wordcloud(r_sqlpass2014_text_corpus, min.freq=35, max.words=100, random.order=F, colors=pal2)
```

This is non-sense! ... But nevertheless.
Case 1 – Comparing days

Comparing 6th and 7th November

Most common frequent words

6th and 7th November
Case 2 – segmenting followers

- People I follow (cca 470) with my twitter account @tomaz_tsql
- Time-span: 5 days
- No hashtags #, no restrictions
- What I know: Only english speaking and mainly SQL gurus, data scientist, R developers. $\sum \sum = cca 300$ twitter accounts
- What I think I know: a lot of serious tweets on topics
- What I don’t know:
  - Are there any interesting stories?
  - Is it possible to determine who is reading books?
  - Can I find a group of SQL geeks that are also vivid runners?
  - Should I close my twitter account?
Case 2 – segmenting followers

„Segmenting“ my followers (based on ward/euclid/manhattan proximity) of statistics (number of friends and followers)

What does it say? Looks nice, but ...
Case 2 – segmenting followers

- Case of segmenting my followers (based on ward proximity of most frequent terms )

Frequent terms

Finding associations between the terms

Do some bootstrapping otherwise... or move to cloud
Case 2 – segmenting followers

Declaring sparsity of term-extraction matrix

95% Sparsity

98% Sparsity

Must tweak it!

Do we really want this word?
Case 2 – segmenting followers

... after some cleaning ... 4 clusters ...

... with cluster centers for most frequent words
Case 2 – segmenting followers

... playing with the 5 clusters
Case 2 – segmenting followers

Cluster visualization
Case 3 – text segmentation using sentimental analysis

- People I follow (cca 470) with my twitter account @tomaz_tsql
- Time-span: 1 days
- No hashtag, no restrictions, but tweet must have word sql
- Näive bayes classifier
Case 3 - sentimental analysis?

What is sentimental analysis of text

- A natural language processing of text
- Identifying and extracting subjective information
- Usually used as to classifying binary polarity; e.g.: happy – sad or 1 - 0
- No hashtag, no restrictions, but tweet must have word sql
- Näive bayes classifier
Case 3 - sentimental analysis?

6th and 7th November 2014

- Emotions:
  - Joy
  - Surprise
  - Sadness
  - Anger
  - Fear
Case 3 - sentimental analysis?

6th and 7th November 2014

- Polarity of Emotions:
  - Positive
  - Negative

![Bar chart showing emotional analysis of tweets with #summit2014 on Microsoft Summit 2014. The chart displays the number of tweets with different emotional polarities: negative, neutral, and positive. The positive sentiment is the most prevalent.]
Case 3 - sentimental analysis?

Code example:

```r
some_txt = gsub("RT\via\((?:\b\w*\@\w+)+\)\", "", some_txt)
some_txt = gsub("@\w+", "", some_txt)
some_txt = gsub("[[[:punct:]]]", "", some_txt)
some_txt = gsub("[[[:digit:]]]", "", some_txt)
some_txt = gsub("http\w+"", "", some_txt)
some_txt = gsub("\[\[\]\]"", "", some_txt)
some_txt = gsub("\b\s+\b\s+\b", "", some_txt)

# define "tolower error handling" function
try.error = function(x)
{
  # create missing value
  y = NA
  # tryCatch error
  try_error = tryCatch(tolower(x), error=function(e) e)
  # if not an error
  if (!inherits(try_error, "error"))
    y = tolower(x)
  # result
  return(y)
}

# lower case using try.error with sapply
some_txt = sapply(some_txt, try.error)

# remove NAS in some_txt
some_txt = some_txt[!is.na(some_txt)]
names(some_txt) = NULL

# classify emotion
class_emo = classify_emotion(some_txt, algorithm="bayes", prior=1.0)
# get emotion best fit
emotion = class_emo[,1]
# substitute NA's by "unknown"
emotion[is.na(emotion)] = "unknown"
```
Conclusion ...

False negative!

NO YOU CANNOT predict elections with Twitter

Limits of Electoral Predictions using Twitter
Questions?

While \{Questions=0\}

Feel free to drop me e-mail \(\text{tomaz.kastrun@gmail.com}\) if:

- You want to discuss text mining
- You want a copy of R code
- Want to drink \(x^n\) cups of coffee
- Want to transpose matrices or calculate bit parity
Thank you!