“Estonia” in Twitter and NY Times

Risto Hinno

Introduction and research question

Goal of this assignment is to understand what topics about Estonia are discussed in Twitter and NY Times. Research question is to find out if some of the topics discussed about Estonia in Twitter and NY Times are related to IT/technology/eGovernment. This one topic which estonians think they are famous for. Claim is subjective and only represents authors opinion.

Data gathering and analysis

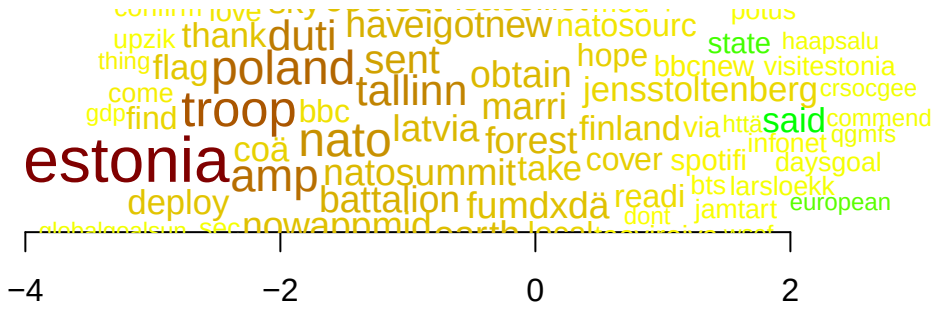
Data is gathered from NY Times and Twitter, searching for term “Estonia”. Code for gathering data (and further analysis) is here, because it is rather long for displaying.

In total 13 386 (from 2016.07.02 to 2016.07.12) tweets and meta data for 2002 NY Times articles (from 2006.01.01 to 2016.07.12) and 1096 articles body via web scraping (some links from metadata gave error 404) were gathered. It is important to keep in mind that timeframes for two datasets are different. For further analysis New York Times articles are compared with Twitter tweets. First both corporas are compared.

```
source("helpers.R");library(tm);library(corpustools)
EstoniantweetsDf=readRDS("EstoniaTweets.RDS");EstoniaNYT=readRDS("EstoniaNYT.RDS")
EstoniantweetCorpus <- VCorpus(VectorSource(cleanTweet(EstoniantweetsDf$text)))
dtmEstoniaTweet <- DocumentTermMatrix(EstoniantweetCorpus,
  control = list(stemming = TRUE, stopwords=T,removeNumbers = TRUE,
    removePunctuation = TRUE,  wordLengths = c(3, 140)))
NYTCorpus <- VCorpus(VectorSource(cleanTweet(EstoniaNYT$bodyTitle[!is.na(EstoniaNYT$titles)])))
dtmEstoniaNYT <- DocumentTermMatrix(NYTCorpus,
      control = list(stemming =TRUE,stopwords=T, removeNumbers = TRUE,
   removePunctuation = TRUE,wordLengths = c(3, 140)))
cmp = corpora.compare(dtmEstoniaNYT, dtmEstoniaTweet)
cmp = arrange(cmp, -chi)
```

From the plot it could be seen that NY Times is more (left side) related to security issues and twitter has bigger variety of topics.To find which topics are in texts, topic modelling is used. Number of topics were found by trying different topic numbers. If topics seemed to be too similar number of topics was reduced.

```
with(head(cmp, 100), plotWords(x=log(over), words = term, wordfreq = chi, random.y = T))
```




```
library(knitr) ;mNYT=lda.fit(dtmEstoniaNYT, K=5, alpha=.1);kable(terms(mNYT, 5))
```

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
art	one	russia	said	play
new	like	russian	countri	game
street	peopl	said	european	said
theater	time	nato	percent	team
center	said	state	year	first

```
mtweets=lda.fit(dtmEstoniaTweet, K=4, alpha=.1);kable(terms(mtweets, 5))
```

Topic 1	Topic 2	Topic 3	Topic 4
estonia	estonia	estonia	estonia
nato	tallinn	one	poland
russia	day	come	nato
troop	good	earth	troop
baltic	finland	forest	take

NYT topcis could be named following: art, people, Russia/NATO, government/economy, sports. Twitter data topics could be interpreted as security (Baltic cooperation), tourism, nature, security (cooperation with Poland).

Conclusions

As seen from the previous analysis twitter data did not involve topic regarding technology. Topic which NY Times and Twitter data shared was related to security/NATO. This evidence doesn't support claim that Estonia is famous in IT field. It must be bear in mind that data subset is from short period and conclusive conclusions could not be made based on this data. Also using differnet number of topics could give different results.