Lab: Text categorization: sport vs. politics (4h)

Build a binary classifier for tweets: sport vs. politics

1. decide input, output
2. decide solution assessment

→ ON PAPER !

3. decide (if any) how to obtain learning data
4. decide workflow and ML technique

R and Twitter

Package `twitteR`

- (Linux, possibly install `libcurl4-gnutls-dev` and `libssl-dev`)
- needs Twitter API registration: `https://apps.twitter.com/`
- `https://rayli.net/blog/data/newborn-app-using-twitter-and-r-data-analysis/`
- retrieve (`userTimeline("MaleLabTs", n = 20)`), convert in DF (`twListToDF(tweets)`)

Text mining in R

Package `tm`

- `http://www.rdatamining.com/docs/text-mining-with-r`
- load data from character vector

```
Corpus(VectorSource(tweets.df$text))
```

- to lowercase `tm_map(myCorpus, content_transformer(tolower))`
- remove punctuation

```
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*",
"", x)
tm_map(myCorpus, content_transformer(removeNumPunct))
```

- remove stop words

```
tm_map(myCorpus, removeWords, myStopwords)
```

- stemming `tm_map(myCorpus, stemDocument)` (requires package `snowballC`)
- getting as data

```
as.data.frame(t(as.matrix(TermDocumentMatrix(myCorpus))))
```