Text Mining in R

Christoph Hoffmann
Appstam Consulting GmbH
www.appstam.com

APPSTAM
Applied Statistics Machines

Copyrights and Trademarks

Agenda

Adapted Data Science Process

Case Study: Twitter Sentiment Analysis

Common Data Science Process

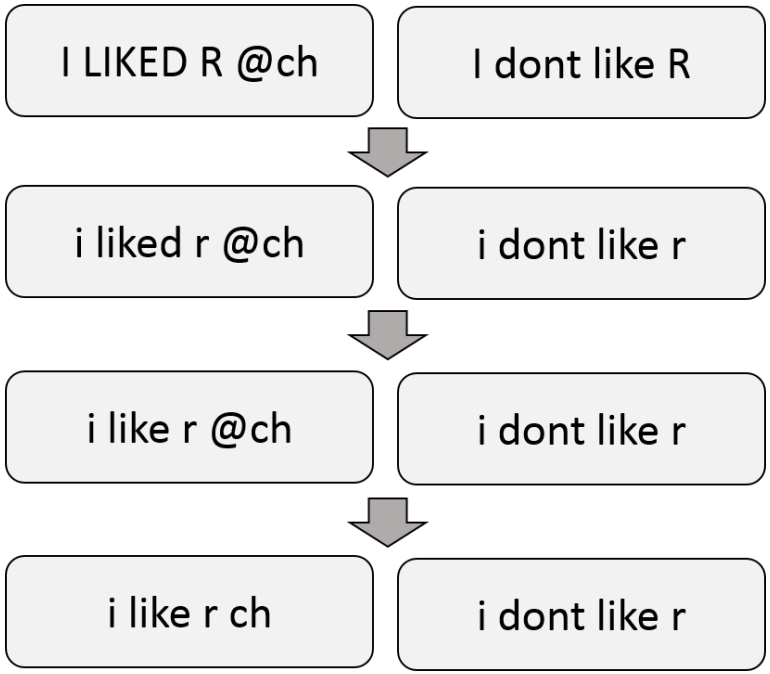
- Two way process
- Motivational questions
- Difference between Data Mining and Text Mining

Data Science Process: Acquisition

- Text data in machine readable format: books, mails, social networks, news
- Document standardization
- Raw text input - corpus - collection of text documents

```
library("tm")
corpus <- VCorpus(DirSource("C:/Users/E2/Desktop/eRum/text"))
lapply(corpus,as.character)

$document1.txt
"I LIKED R @ch"

$document2.txt
"I dont like R"
```

Data Science Process: Preparation



- Lower cases
- Stemming ("SnowballC")
- Remove stopwords

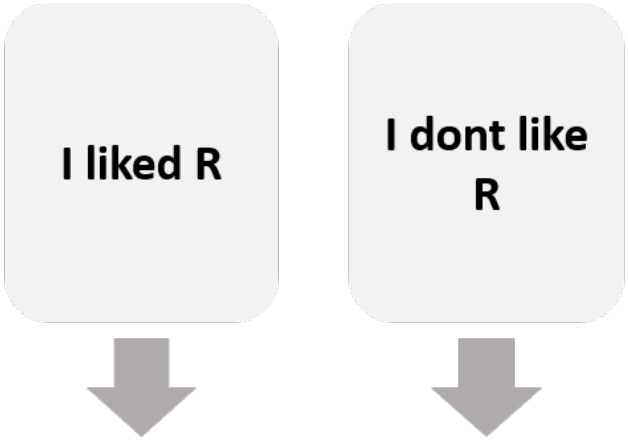
```
tm_map(corpus, content_transformer(tolower))

tm_map(corpus, stemDocument)
```

Data Science Process: Preparation

Name	Income	Gender	District	Credit
Christoph Hoffmann	800 €	M	10553	0
Annelise Schmitt	2600 €	F	10551	1
...	.. .	...	...	...
Max Mustermann	1500 €	NA	10538	1

- Highly structured data
- Starting point for modelling
- Create similar structure for Text Mining

Data Science Process: Preparation



	I	liked	R	dont	like
Doc 1	1	1	1	0	0
Doc 2	1	0	1	1	1

- Create Document-Term-Matrix
- Basis for summary statistics ("wordcloud")

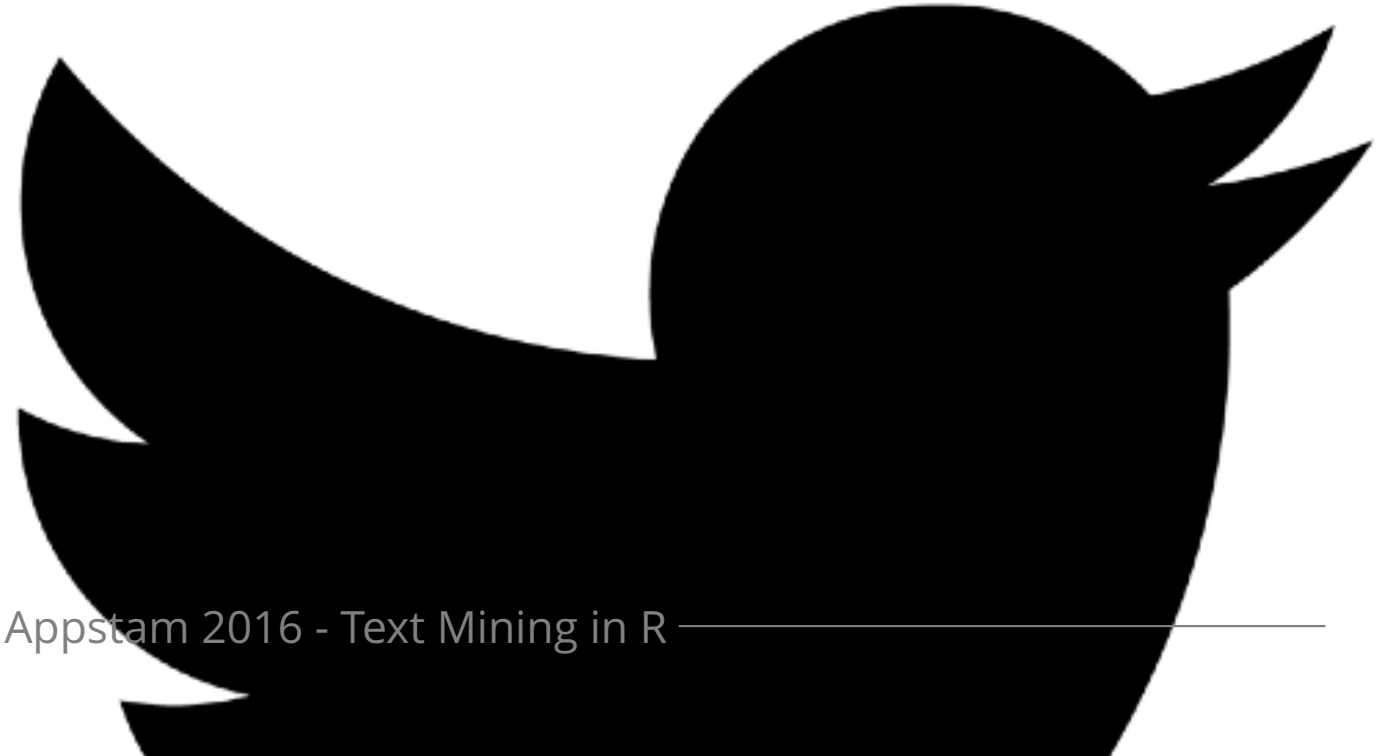
```
dtm <- DocumentTermMatrix(corpus)

findFreqTerms(dtm, 100)
```

Data Science Process: Modelling

Applications	
Classification	Map documents to different classes
Clustering	Organize documents into similar groups
Information Retrieval	Similar and most relevant documents
Information Extraction	Extract data from unstructured format

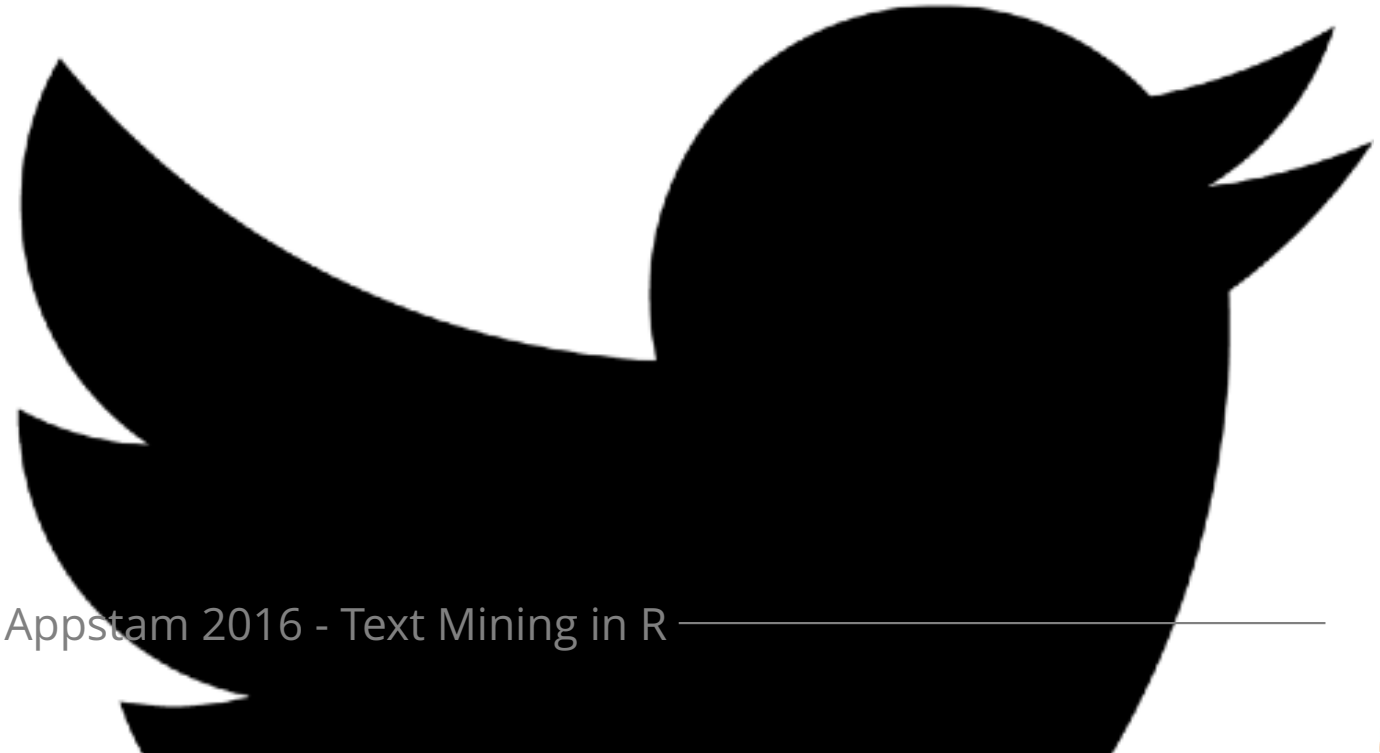
Acquisition: Twitter API

Create application:	https://apps.twitter.com/
Twitter's Enterprise API:	GNIP
Sample Data:	www.crowdflower.com/data-for-everyone/



Acquisition: Twitter API

```
library(twitteR)
setup_twitter_oauth("4 identification keys")
tweetsDonald <- userTimeline("realDonaldTrump", n=3200)
```



Modelling: Naive Bayes Classification

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{1}$$

$$P(d|c) = P(w_1, \ldots, w_n|c) = P(w_1|c) \times P(w_2|c) \times \cdots \times P(w_n|c) \tag{2}$$

$$\hat{P}(c) = \frac{N_c}{N_d} \tag{3}$$

$$\hat{P}(w_j|c) = \frac{|w_j \text{ in } c|}{|w \text{ in } c|} \tag{4}$$