1 Understanding the problem

Tasks to accomplish

Obtaining the data - Can you download the data and load/manipulate it in R?

Familiarizing yourself with NLP and text mining - Learn about the basics of natural language processing and how it relates to the data science process you have learned in the Data Science Specialization.

1.1 Obtaining the data

Download the data from Amazon S3.

Reference to the website which is the original source of the corpus, maintained by Hans Christensen

The dataset contains news, blogs and tweets in four different languages, English, German, Russian and Finish.

References: Corpus website Dataset description. Lines, sizes, etc. tm and tau libraries Corpus statistics

Questions to consider

What do the data look like?

Summary statistics about the data sets.

basic summaries of the three files? Word counts, line counts and basic data tables

The english Corpus has three datasets, with the following statistics:

Twitter: Small sentence(s), maximum number of characters observed is 213. There are 167 million characters, in 30 million words, in 2.3 million tweets. Blogs: Paragraphs. Multiple sentences per blog. Largest sentence has 40835 characters. In total, this dataset has 37 million words in less than a million lines. News: Paragraphs. Multiple sentences. Largest sentence has 11384 characters, total words are 30 million in 1 million lines.

```
> docs<-Corpus(DirSource(file.path(".", "dataset", "en_US")))
```

Where do the data come from?

Twitter Blogs News

Can you think of any other data sources that might help you in this project?

Mailboxes. Facebook/googleplus/linkedin. Twitter stream.

1.2 NLP and text mining

Familiarizing yourself with NLP and text mining - Learn about the basics of natural language processing and how it relates to the data science process you have learned in the Data Science Specialization.

What are the common steps in natural language processing?

The NLP pipeline involves the following steps:

- EOS detection. Are the 3 datasets are already categorized like this?

- Tokenization. In the four languages, tokens are words, splitted by space. In the languages that use pictograms, there is no space to seperate the tokens in sentences.
- Profanity filtering.
- Part-of-speech tagging. Tag tokens by nouns, verbs, etc.
- Chunking. Grammar based analysis of the tagged tokens, not statistical analysis.
- Extraction

What are some common issues in the analysis of text data?

What is the relationship between NLP and the concepts you have learned in the Specialization?

2 Data acquisition and cleaning

Tokenization - identifying appropriate tokens such as words, punctuation, and numbers. Writing a function that takes a file as input and returns a tokenized version of it.

Profanity filtering - removing profanity and other words you do not want to predict.

Loading the data

The load of the data in R has been done in the Corpus data structure, provided by the text mining framework library, tm. That loads the corpus in to the memory.

Data frame is not a good data type to load the text, because it is prone to dimentionality problems. Corpus is using lists.

2.1 Tokenization

2.2 Profanity filtering

3 Exploratory analysis