Information Retrieval Project 1, Group 11

Michael Merki, Julien Romero, Markus Greiner

November 2016

# Installation

See README.md how to install

# Preprocessing / Tokenizing

We started with the provided tokenizer that is part of ReutersRCVParser. For the 50’000 training documents, this resulted in a vocabulary size of 9.2 Mio tokens for the given 150 K words.

The performance of the classifiers during runtime is strongly influenced by the vocabulary size. So the goal was to reduce vocabulary size without negative influence on precision.

In a second step, the following tokens were matched using regular expressions and replaced numbers, dates, ordinals, punctuation and underlines with standard tokens:

// regular expressions defined statically  
**val** *rDate* = ("^\\d+[/-]\\d+[/-]\\d+$".r, "<DATE>")  
**val** *rUSPhone* = ("^\\d{3}\\W\\d+{3}\\W\\d{4}$".r -> "<USPHONE>")  
**val** *rNumber* = ("^[-]?\\d+([.,]\\d+)\*$".r -> "<NUMBER>")  
**val** *rTwoNum* = ("^\\d+[-/=]\\d+$".r -> "<TUMBER>")  
**val** *rOrdinal* = ("^\\d+(th|1st|2nd|3rd)$".r -> "<ORDINAL>")  
**val** *rPunct* = ("[,;.:]$".r -> " <PUNCT>") // if it is like "end.", should return "end <PUNCT>"  
**val** *rLine* = ("--+".r -> "") // underlines like -----------

In addition, the list of english stop words of the nltk[[1]](#footnote-1) was taken and all of these words replaced by a single token.

These two measures reduced the vocabulary by 135 words.

In a further step, the provided Porter stemmer was applied, reducing the vocabulary by over 25’000 words .

During a final step, the terms in the training set with frequency > 30’000 were manually inspected, and these words were added to the stop-word list:

**val** *HighFreqWords* = *List*(  
 //"share", // -> 32117,  
 //"bank", // -> 30300,  
 //"million", // -> 45624,  
 "would", // -> 35039,  
 //"percent", // -> 53790,  
 "year", // -> 40026,  
 //"market", // -> 35931,  
 "said", // -> 155872,  
 "new" // -> 33655  
)

All these stop-words were converted to static regular expressions to speed up matching.

Overview vocabulary size

50’000 documents, total 9’208’748 tokens.

|  |  |
| --- | --- |
| Tokenizing | Size vocabulary |
| Simple white-space tokenizer | 150’125 |
| Replace number-pattern matching  Replace stop-words | 149’990 |
| Porter Stemmer  Replace most frequent words | 124’695 |

1. <http://www.nltk.org/book/ch02.html> [↑](#footnote-ref-1)