Exercise: Text Analytics, LDA (Latent Dirichlet Allocation)

Reuters21578 is loaded, documents are pre-processed:

- Transformed to lower case
- Problematic symbols are removed
- Punctuation is removed
- Strip digits
- Stop words are removed, also words that appear many times, with no sense
- Stemming

21.000 documents are loaded, we will only pick a few to work with them faster, only documents from some topics are selected, after observing the existing topics some of the most popular are selected:

```
topicsInterest<-c("crude", "grain", "trade", "wheat", "ship", "corn", "oilseed","coffee","gold","money-supply",
"sugar","jobs","livestock")
```

Initially:

```
> Reuters21578
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 21578
```

Then few topics are selected:

```
> Reuters21578
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 2831
```

A bunch of words are removed, to find more interesting conclusions:

```
# remove unwanted resulting words

> myStopWords<-c("mths", "qtli", "nine", "avg", "billion", "shrs", "div", "cts", "shr", "net", "loss", "rev", "share"
, "mln","pct","dlrs","week","year","month","januari","februari","march","april","may","juny","july","ago","decemb")
> doc <- tm_map(stem.doc, removeWords, myStopWords)
```

Pre-processing also includes conversion to lower case, removing punctuation, stopwords and special characters as well as numbers.

The DTM (Document Term Matrix) is created, from which we can already extract some information:

```
# top words
> freq <- colSums(as.matrix(dtm))
> ord <- order(freq, decreasing = 1)
> freq[head(ord)]

     oil   export    trade    wheat    price    sugar   import     last     rose     gold   barrel     with     bank     were      nil
43.66663 40.61775 38.69861 38.37866 34.26687 29.22267 28.10607 27.12391 26.97342 26.92745 26.49221 26.39832 26.38077 26.37577 25.71869

   grain    total   depart  product    coffe
25.64301 24.97490 24.79160 24.74104 24.02716
```

```
# tail words
> freq[tail(ord)]
     cant  momentum      gear      hint   nervous    barney
0.3319242 0.3300684 0.3162873 0.3092594 0.3038089 0.2820412

# Words frequency
```

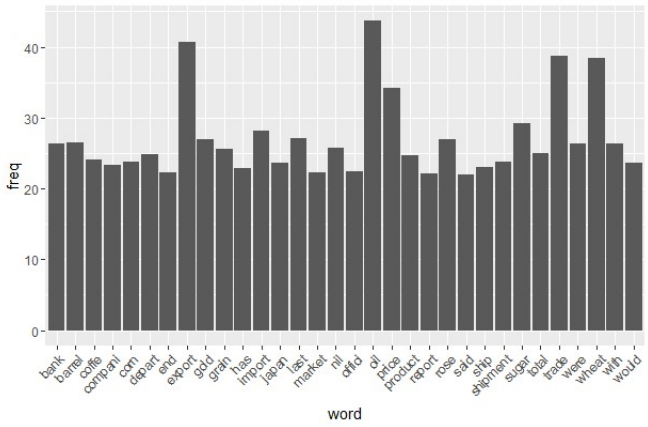

Figure 1 – Words frequency

We can as well conform a word cloud:

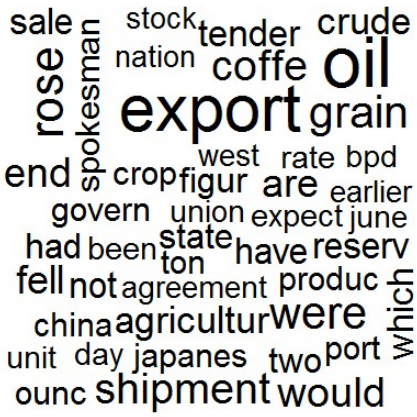

Figure 2 – Word Cloud

After obtaining the DTM (Document Term Matrix), and removing empty words rows to avoid error when running *lda,* a LDA model can be built:

```
> ldaOut <-LDA(new_dtm, 10, control=list(alpha=0.1))
```

Running LDA provides us an unsupervised model with the following data:

```
# Which documents got assigned to which topic
> as.matrix(topics(ldaOut))
…
20682   10
20699    8
20709    6
20719    7
20721    6
20723    2
20738    2
20756    7
20757   10
20763    9
20774    7
20778    7
20787    1
…
```

```
# Which words are most representative for each topic
```

```
> as.matrix(terms(ldaOut,10))
```

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
[1,]	"the"	"the"	"the"	"the"	"the"	"the"	"the"	"the"	"the"	"the"
[2,]	"and"	"tonn"	"said"	"and"	"said"	"oil"	"said"	"and"	"and"	"billion"
[3,]	"said"	"for"	"and"	"said"	"and"	"said"	"and"	"said"	"said"	"and"
[4,]	"trade"	"and"	"for"	"crop"	"coffe"	"price"	"ship"	"compani"	"trade"	"from"
[5,]	"japan"	"said"	"that"	"was"	"export"	"and"	"gulf"	"gold"	"that"	"said"
[6,]	"that"	"export"	"price"	"area"	"for"	"barrel"	"was"	"will"	"would"	"bank"
[7,]	"japanes"	"wheat"	"market"	"for"	"quota"	"for"	"that"	"oil"	"for"	"februari"
[8,]	"for"	"reuter"	"will"	"tonn"	"produc"	"crude"	"reuter"	"for"	"not"	"januari"
[9,]	"with"	"from"	"this"	"last"	"sugar"	"opec"	"for"	"reuter"	"import"	"was"
[10,]	"offici"	"nil"	"are"	"reuter"	"price"	"bpd"	"had"	"from"	"propos"	"rose"

Even if the R script to build this model should be refined (as unfortunately we do not get to remove words like *the*, *and*, *for* or *that*), and then some topics do not look clear enough, we could define topics as:

Topic 1	*japan trade*
Topic 2	*wheat export*
Topic 3	*market price*
Topic 4	-
Topic 5	*coffee/sugar export/production*
Topic 6	*oil/crude barrels*
Topic 7	*oil shipping from gulf*
Topic 8	*gold and oil companies*
Topic 9	*trade/import*
Topic 10	*bank*