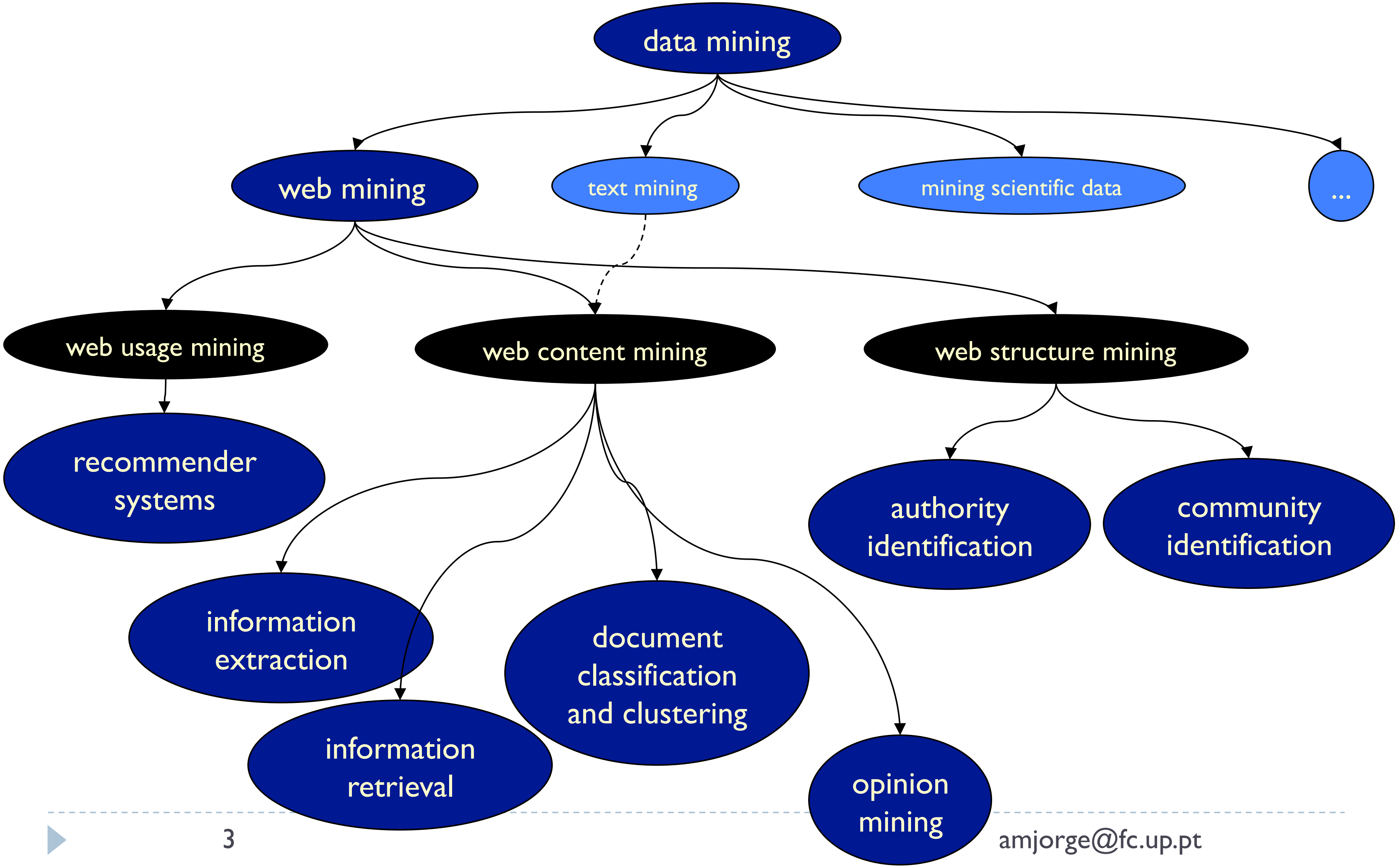
Text Mining

Alípio Jorge, DCC-FC, Universidade do Porto
amjorge@fc.up.pt

Overview

- Text Mining introduction
  - concepts
- Example task of clustering documents
  - stopword removal

Knowledge (sort of) tree



What is Text Mining

- Text mining (TM)
  - extracting useful information from a collection of documents
- wrt data mining:
  - data sources are unstructured or semi-structured documents.
- TM involves:
  - Basic pre-processing / TM operations, such as
    - identification / extraction of representative features
  - identification of complex patterns
    - e.g. relationships between previously identified concepts
- TM exploits techniques / methodologies from
  - data mining, machine learning, information retrieval,
  - corpus-based computational linguistics

Concepts

- Corpus
  - collection of documents
- Static / Dynamic

- Text documents can be :
  - unstructured
    - i.e. free-style text
      (but from a linguistic perspective they are really structured objects)
  - weakly structured
    - adhering to some pre-specified format,
      - □ scientific papers, business reports, legal memoranda, news stories etc.
  - semistructured
    - exploiting heavy document templating or style sheets.
      - □ html, xml, latex

Document representation

- Feature based representation
  - each document is transformed into a set of features
  - vector model

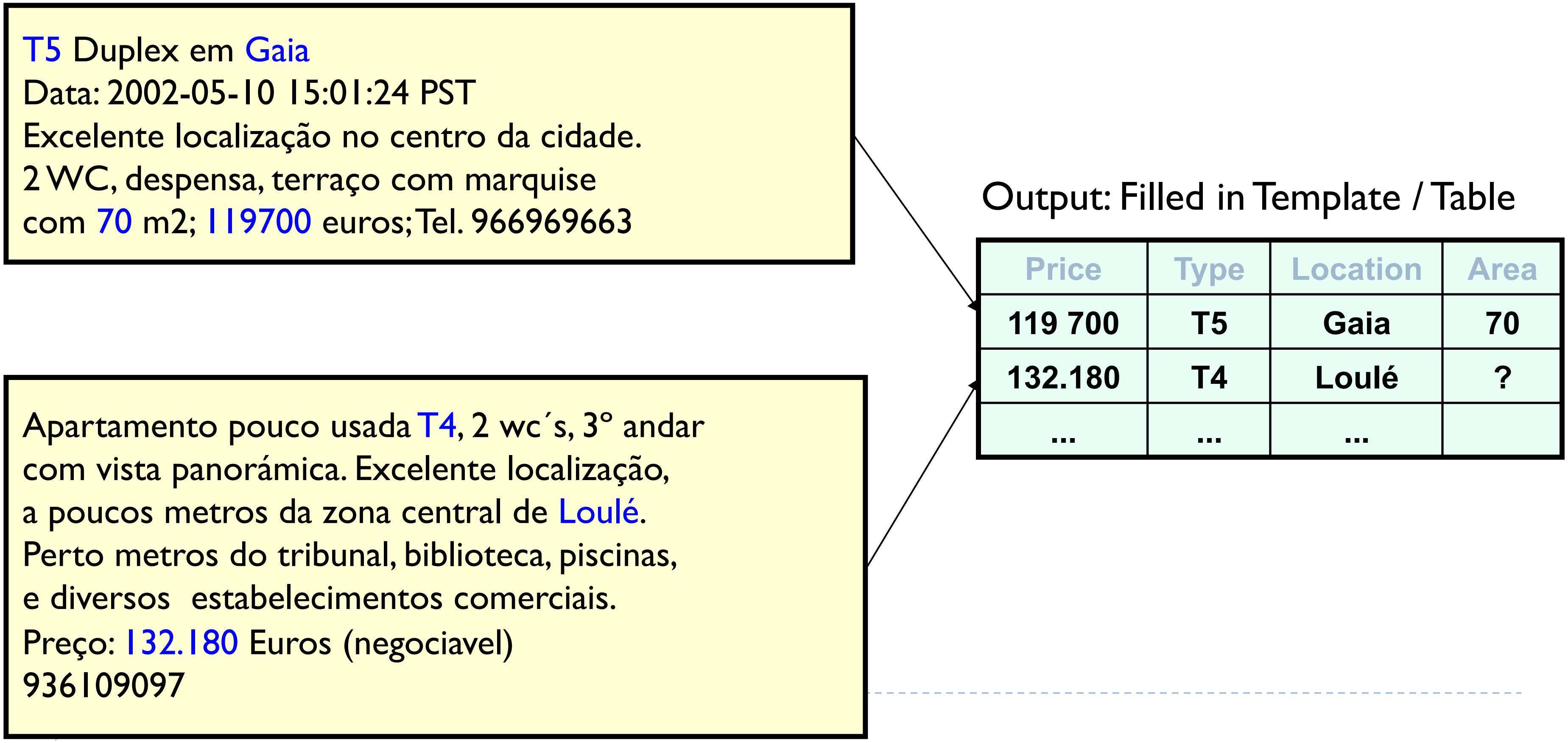
- Features
  - Words
    - bag-of-words representation
  - Terms
    - including multi-words
      - □ "white house"
  - Concepts
    - concept "car" can be represented by different terms
      - □ car, automobile, vehicle, sports car
      - □ synonimy, polysemy

Common Text Mining Tasks

- Information Retrieval
- Clustering / organization of documents
- Document classification (categorization)
- Information extraction

Information Extraction

- IE involves identification of certain entities in the text, their extraction and representation in a pre-specified format (e.g. a table).



T5 Duplex em Gaia
Data: 2002-05-10 15:01:24 PST
Excelente localização no centro da cidade.
2 WC, despensa, terraço com marquise
com 70 m2; 119700 euros; Tel. 966969663

Apartamento pouco usada T4, 2 wc´s, 3° andar
com vista panorámica. Excelente localização,
a poucos metros da zona central de Loulé.
Perto metros do tribunal, biblioteca, piscinas,
e diversos estabelecimentos comerciais.
Preço: 132.180 Euros (negociavel)
936109097

Output: Filled in Template / Table

Price	Type	Location	Area
119 700	T5	Gaia	70
132.180	T4	Loulé	?
...	...	...	

(some) Advanced Text Mining Tasks

- Concept co-occurrence
  - Quantification of co-occurrence
  - e.g. Association mining with terms or concepts in texts
- Summarization
  - summarize one text
  - summarize a document collection
- Keyword extraction
  - characteristic keywords
- Sentiment Analysis / Opinion Mining
  - written film reviews
  - discussions in forums about a product or idea

Clustering: one example task

- We have a collection of documents and we want to automatically organize it by dividing it into homogeneous groups or a hierarchy that can be more easily browsed by a user.

- Our collection:
  - 50 news articles from the reuters news agency. These articles belong to the same topic "acquisitions".
- Approach:
  - vectorize
  - cluster

Clustering

- Loading the data
  - this data set comes with package tm
  - it is already packaged as a Corpus
    - some previous steps will be needed for other sources
  - transform docs into a document x term matrix (TF)

```
> data(acq)
> inspect(acq)
> dtm <- DocumentTermMatrix(acq)
> dtm
<<DocumentTermMatrix (documents: 50, terms: 2103)>>
Non-/sparse entries: 4135/101015
Sparsity           : 96%
Maximal term length: 21
Weighting          : term frequency (tf)
```

Clustering

- inspecting

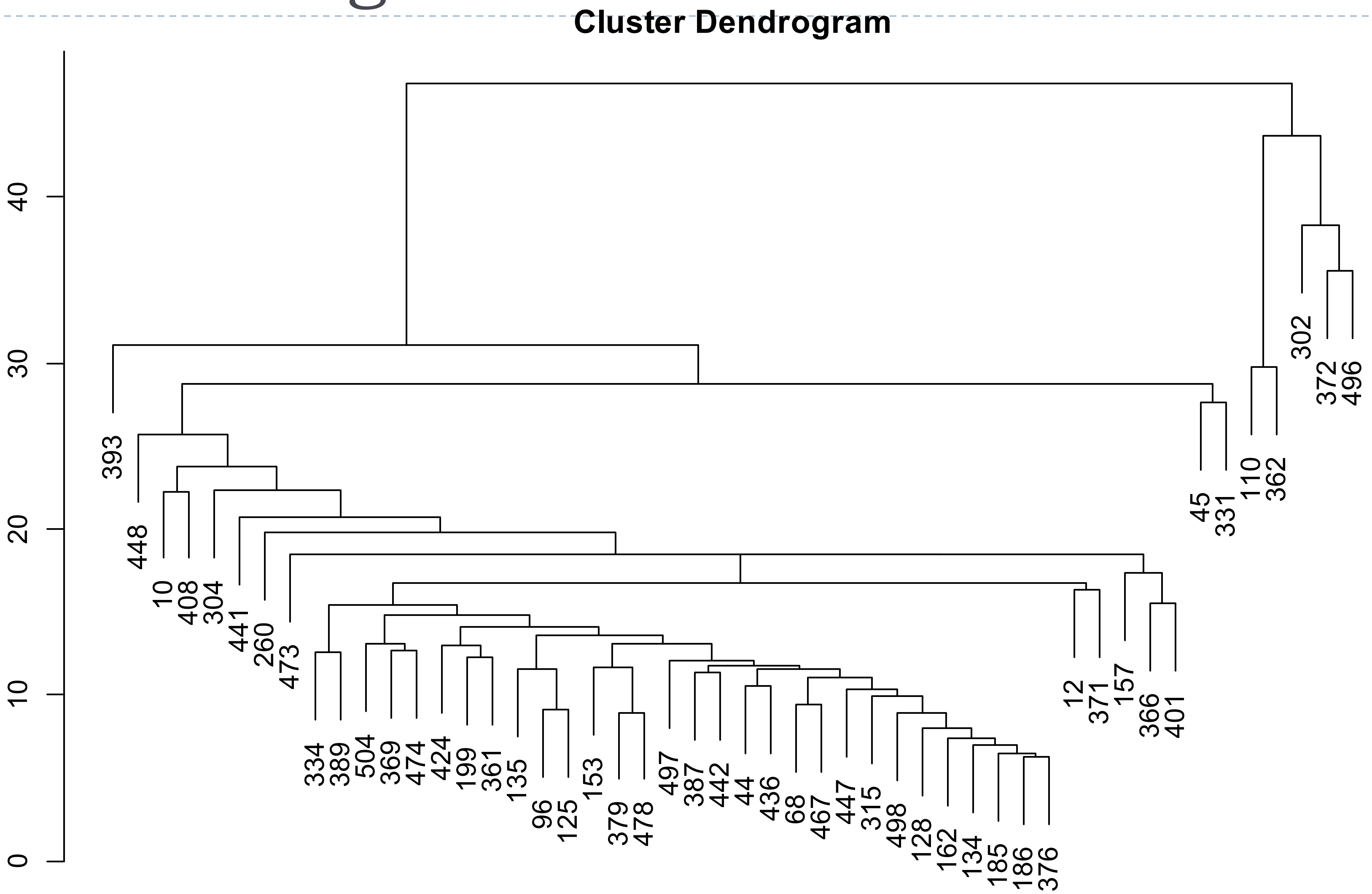
```
> inspect(acq)
A corpus with 50 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID

$`reut-00001.xml`
Computer Terminal Systems Inc said
it has completed the sale of 200,000 shares of its common
stock, and warrants to acquire an additional one mln
shares, to
<Sedio N.V.> of Lugano, Switzerland for 50,000 dlrs.
    The company said the warrants are exercisable for five
years at a purchase price of .125 dlrs per share.
```

Clustering

- applying R's hierarchical clustering
  - build a distance matrix
    - what is the distance measure?
  - call hclust
  - then we can plot the results
    - dist, hclust and plot are from R's base set of functions

```
> DistM <- dist(dtm)
> Tree <- hclust(DistM)
> plot(Tree)
```

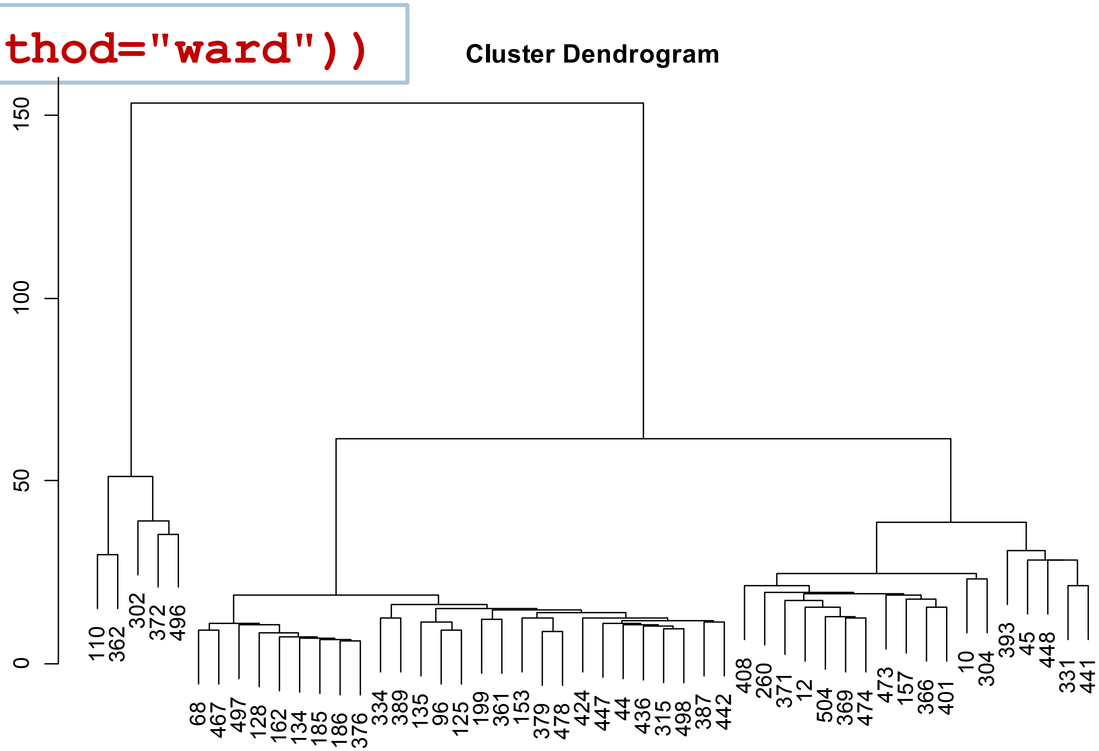
Clustering


Cluster Dendrogram
0
10
20
30
40
393
448
10
408
304
441
260
473
334
389
504
369
474
424
199
361
135
96
125
153
379
478
497
387
442
44
436
68
467
447
315
498
128
162
134
185
186
376
12
371
157
366
401
45
331
110
362
302
372
496

Clustering

- improving cluster balance
  - change the method in hclust
    - average, single, ward, ...

```
> plot(hclust(DistM,method="ward"))
```



Clustering

- getting 3 clusters from clustering tree

```
> ClustKey <- cutree(hclust(DistM,method="ward.D"),3)
> ClustKey
 10  12  44  45  68  96 110 125 128 134 135 153 157 162 
  1   1   2   1   2   2   3   2   2   2   2   2   1   2 
185 186 199 260 302 304 315 331 334 361 362 366 369 371 
  2   2   2   1   3   1   2   1   2   2   3   1   1   1 
372 376 379 387 389 393 401 408 424 436 441 442 447 448 
  3   2   2   2   2   1   1   1   2   2   1   2   2   1 
467 473 474 478 496 497 498 504 
  2   1   1   2   3   2   2   1 
```

Clustering

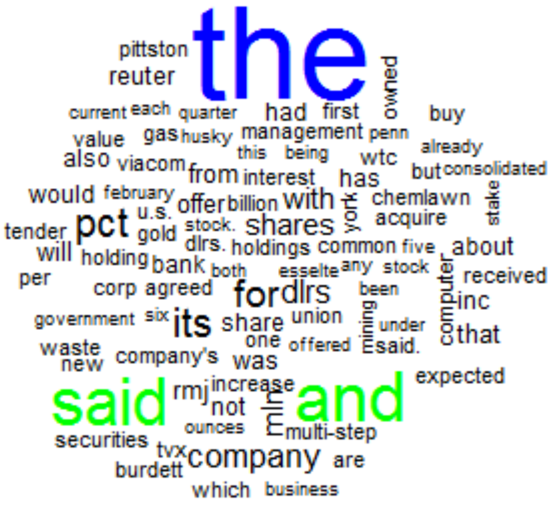
- Characterizing the clusters
  - top tfwords per cluster

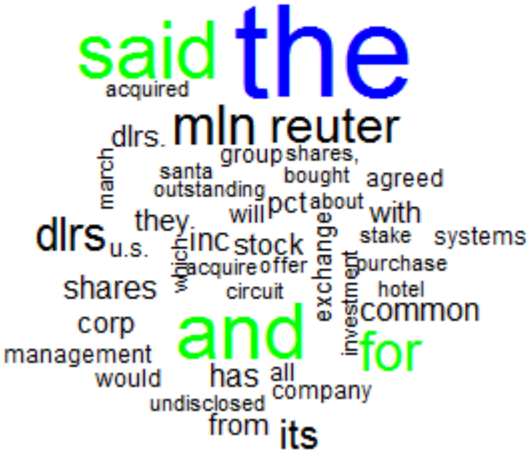
```
> c1 <- dtm[ClustKey==1,]
> sumtf1 <- apply(c1,2,sum)
> sumtf1[order(sumtf1,decreasing=T)[1:30]]
     the     said      and      for      its      mln 
     186       98       88       50       49       40 
  reuter     dlrs      pct   shares      has  company 
      39       35       28       27       25       24 
    with   common      inc     from     will    stock 
      23       22       19       16       16       15 
   would     corp    dlrs.    offer     they    about 
      15       14       14       13       13       11 
  agreed exchange     that      buy    owned    said. 
      11       11       11        9        9        9 
```

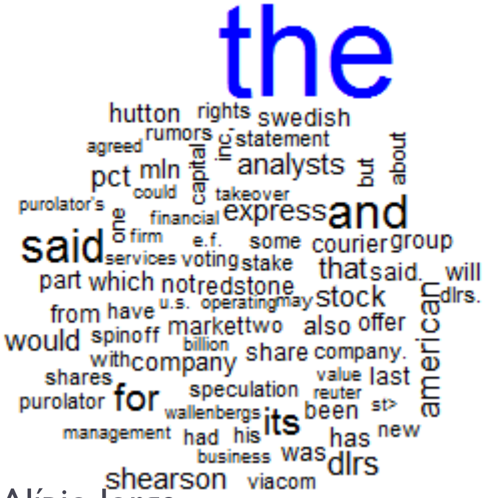
Clustering

- Characterizing the clusters
  - view wordclouds

```
> library(wordcloud)
> wordcloud(names(sumtf1), sumtf1,
+           col=c('black','green','blue'),min.freq=5)
> wordcloud(names(sumtf2), sumtf2,
+           col=c('black','green','blue'),min.freq=5)
```

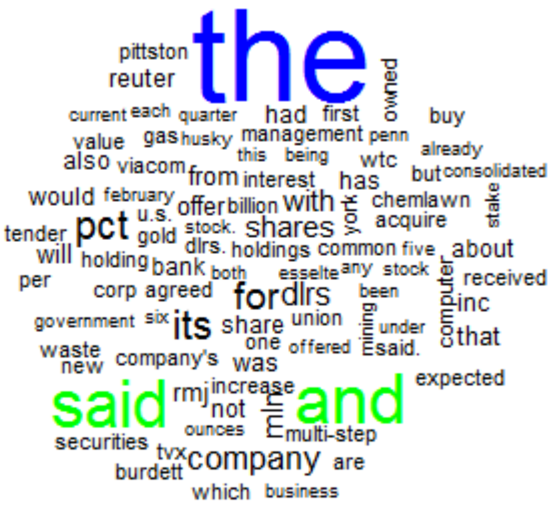


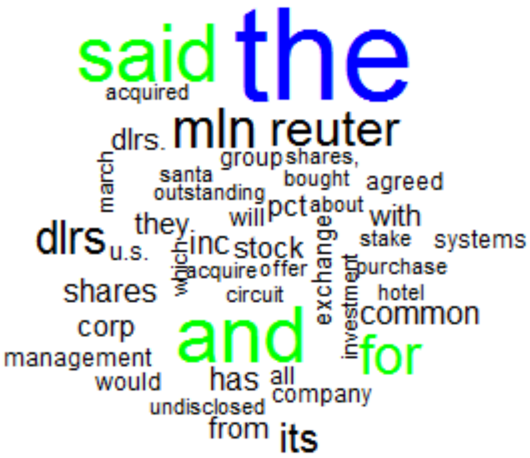


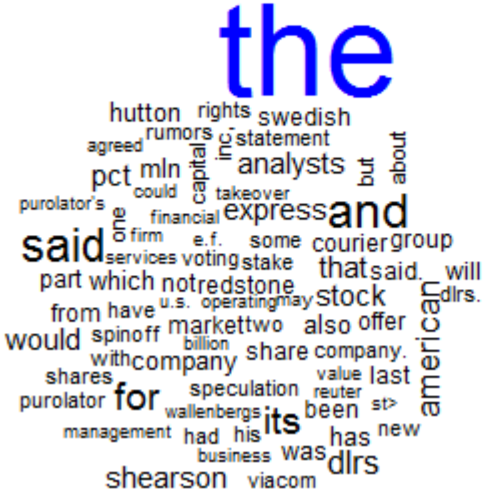


Clustering

- Characterizing the clusters
  - we can see that most frequent words are the same and not very specific
  - these are typically "stopwords"







Stopword removal

- Stopwords
  - "frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents."
  - articles, prepositions and conjunctions are natural candidates.

```
> stopwords("en")
  [1] "a"           "about"       "above"       "across"
  [5] "after"       "again"       "against"     "all"
  [9] "almost"      "alone"       "along"       "already"
 [13] "also"        "although"    "always"      "am"
 [17] "among"       "an"          "and"         "another"
 [21] "any"         "anybody"     "anyone"      "anything"
 [25] "anywhere"    "are"         "area"        "areas"
 [29] "aren't"      "around"      "as"          "ask"
...
```

Stopword removal

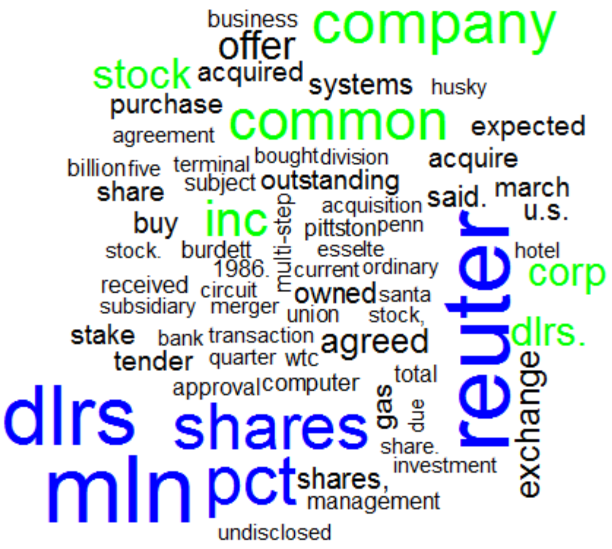
- Stopwords
  - remove columns from dtm that correspond to stopwords

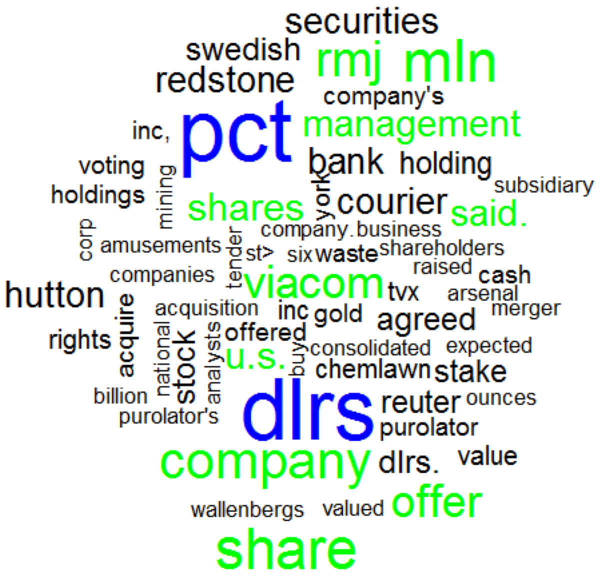
```
> dtms <- dtm[,setdiff(colnames(dtm),stopwords("en"))]
> ncol(dtm)
[1] 2007
> ncol(dtms)
[1] 1843
```

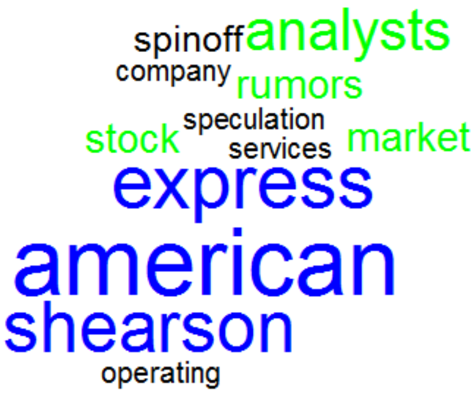
Stopword removal

- Stopwords

```
> Tree<-hclust(dist(dtms),method="ward")
> plot(Tree) # check if 3 clusters is still a good idea
> k<-cutree(Tree,3)
> words1<-apply(dtm[k==1,],2,sum)
```







Stopword removal

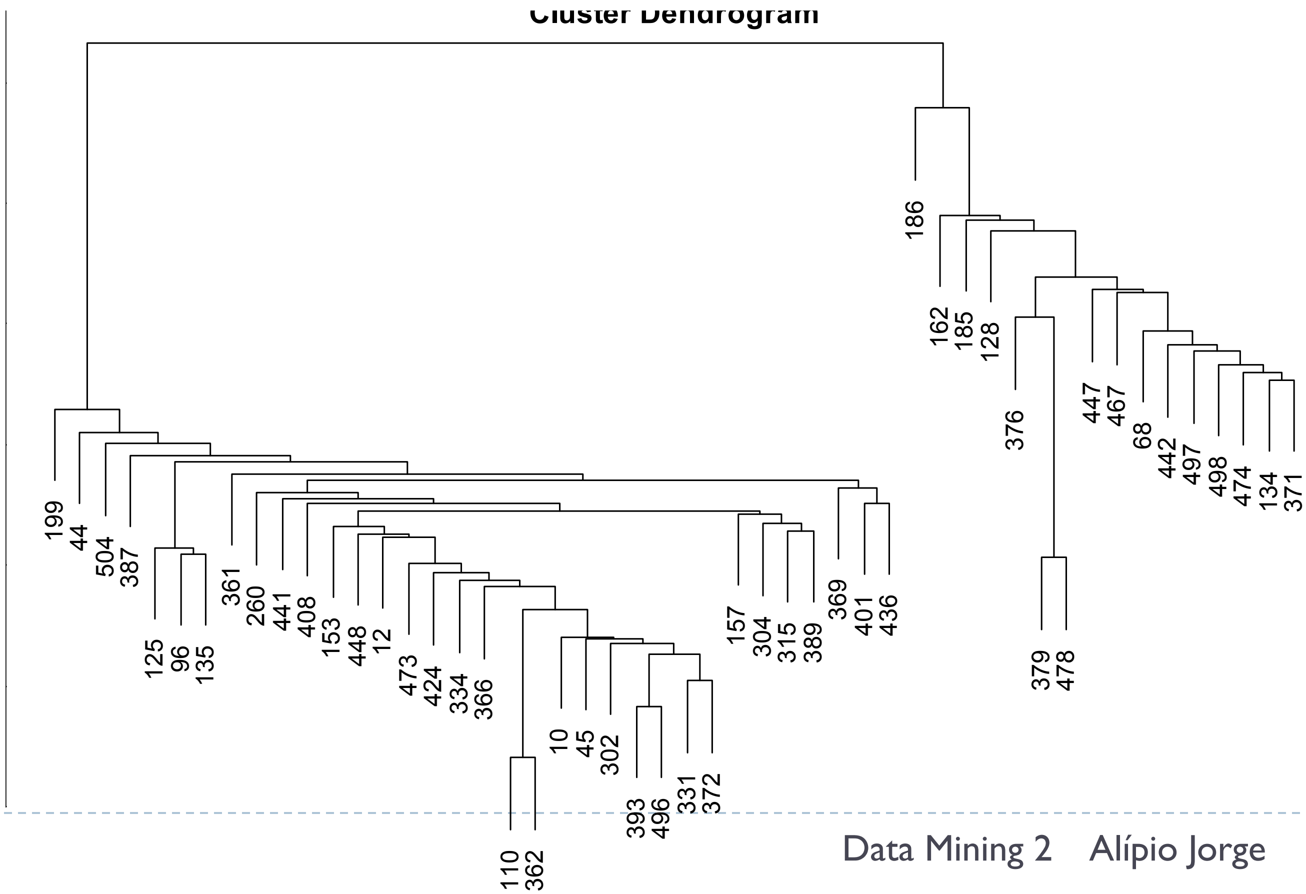
- stopwords were clearly in the way
- but we could try TF-IDF
  - it is supposed to be able to penalize words that are common to many docs
- and we have been using euclidean distance – not cosine
  - euclidean is default method of the function dist

- Let's try these paths then
  - without stopword removal first

TF-IDF (from the beginning)

- have a look at the tree

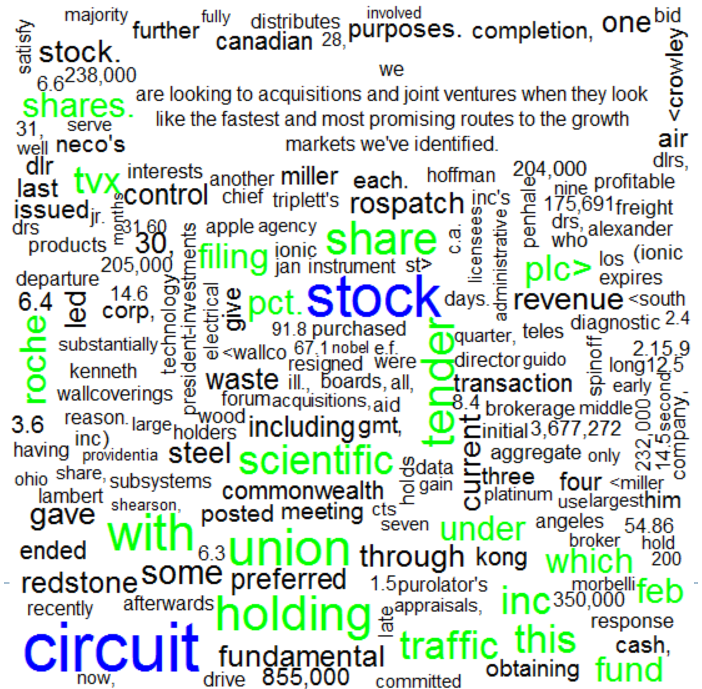
```
> dtm<-weightTfIdf(DocumentTermMatrix(acq))
> plot(hclust(dist(dtm),method="ward"))
```

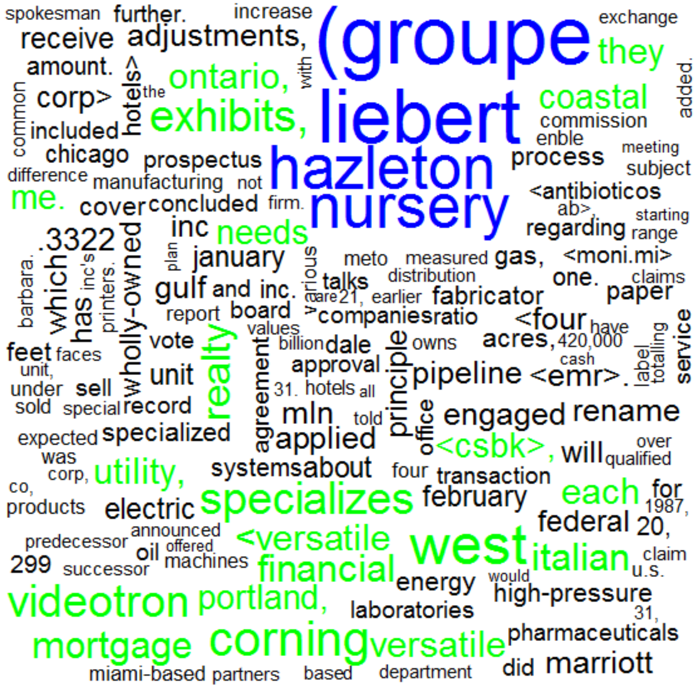


TF-IDF (from the beginning)

- build the wordclouds: no stopwords

```
> Tree<-hclust(dist(dtm),method="ward")
> k<-cutree(Tree,2)
> words1<-apply(dtm[k==1,],2,sum)
> words2<-apply(dtm[k==2,],2,sum)
> wordcloud(names(words1),words1*100,
+       col=c('black','green','blue'),min.freq=5)
```

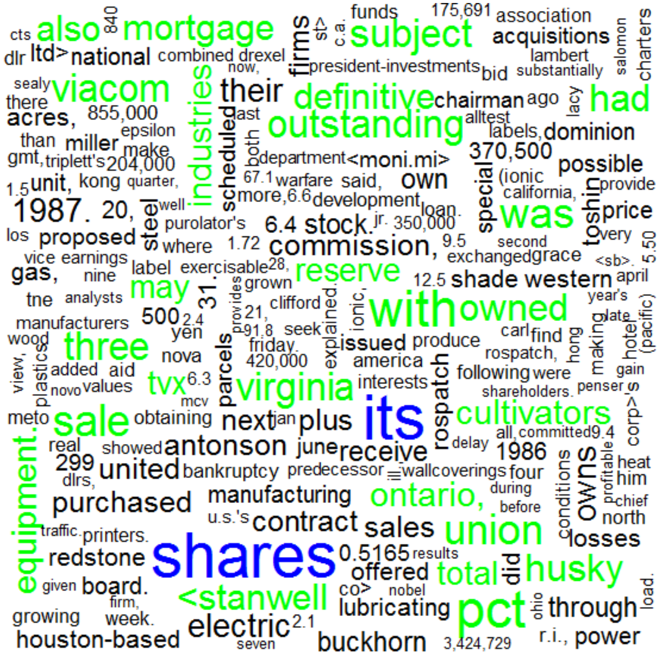


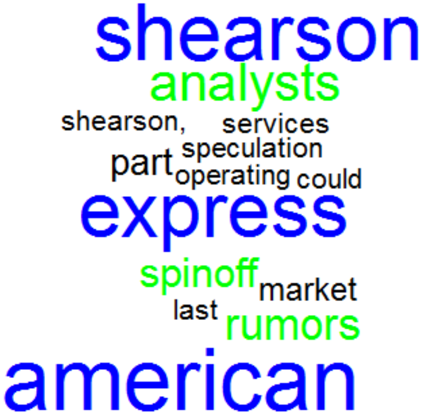


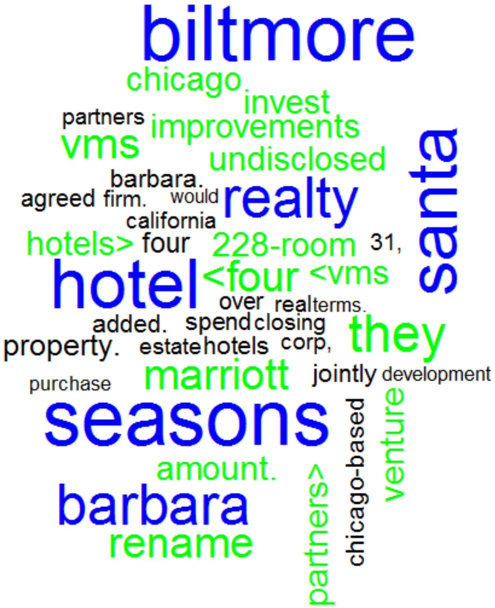
TF-IDF with cosine distance

- use function ”mycosdist" previously defined

```
> Tree<-
hclust(proxy::dist(as.matrix(dtm),method=mycosdist),method="ward.D")
> plot(Tree)
> k<-cutree(Tree,3)
> words1<-apply(dtm[k==1,],2,sum)
> wordcloud(names(words1),words1*100,
+           col=c('black','green','blue'),min.freq=5)
```







Clustering example: summary

- Applying hierarchical clustering to text
- Using TF and TF-IDF schemes
- Removing stopwords
- Using euclidean and cosine distance
- Using wordclouds

Exercises

- Combine corpora “acq” and “crude” from tm.
  - Apply clustering and see if there are two natural clusters and if the wordclouds characterizing the clusters are indicative of the content.
  - Variants
    - Use TF, no stop words, euclidean distance
    - Use TF, with stop words, euclidean distance
    - Use TF-IDF, with and without stop words, euclidean distance
    - Use TF-IDF, with and without stop words, cosine distance
  - Produce an Rmd report with your commands and results.

```
> data(crude)
# notice that c is specially defined in tm as tm_combine
> docs <- c(acq,crude)
```

Resources

- Books
  - Web Data Mining, Bing Liu, Springer, 2007
  - Mining the World Wide Web, Chang, G., Healey, M., McHugh, J., Wang, J., Kluwer Academic Press, 2001.
  - Modern Information Retrieval, Ricardo Baeza-Yates and Berthier Ribeiro-Neto
- Slides
  - Pavel Brazdil's on text mining