Szövegbányászat

Áttekintés

A dokumentumban csupán a szövegbányászat azon területét vizsgáljuk meg, amely a nagy méretű szöveges dokumentumok áttekintését és értelmezését, a dokumentumok metrika mentén történő összehasonlítását teszi lehetővé. E célt elsősorban a dokumentumok ún. Bag-of-Words modell szerinti ábrázolásával érhetjük el.

Tekintsük az alábbi két dokumentumot, amelyek mind egy-egy mondatot tartalmaznak:

```
text<-c("This is one hell of a sentence.", "This is another sentence, not the previous sentence.")
```

Az R nyelvben a szövegbányászatot a `tm` csomag segítségével végezhetjük el. Ehhez egy ún. *corpus*-t kell létrehoznunk:

```
corpus<-Corpus(VectorSource(text))
print(corpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
```

Láthatjuk, hogy a corpus az egyes dokumentumok kezelését végzi, jelen esetben két dokumentumot tartalmaz. A corpus-t részletesebben is megvizsgálhatjuk:

```
inspect(corpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1] This is one hell of a sentence.
## [2] This is another sentence, not the previous sentence.
```

```
inspect(corpus[[1]])
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 31
##
## This is one hell of a sentence.
```

A corpus-on különböző szűréseket végezhetünk:

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

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
# kisbetűvé alakítjuk
corpus<-tm_map(corpus,removePunctuation)        # eltávolítjuk az írásjeleket
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
corpus<-tm_map(corpus,stripWhitespace) # eltüntetjük a többszörös szóközöket
```

```
## Warning in tm_map.SimpleCorpus(corpus, stripWhitespace): transformation drops
## documents
```

```
inspect(corpus[[2]])
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 50
##
## this is another sentence not the previous sentence
```

Ezek alapján elkészíthetjük a bag-of-words térbeli leírását a dokumentumoknak, az ún. `document term matrix`-ot (amely alapesetben a *term frequency* súlyozást alkalmazza, azaz az egyes szavak előfordulási gyakoriságát jegyzi):

```
dtm<-DocumentTermMatrix(corpus)
inspect(dtm)
```

```
## <<DocumentTermMatrix (documents: 2, terms: 8)>>
## Non-/sparse entries: 10/6
## Sparsity           : 38%
## Maximal term length: 8
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs another hell not one previous sentence the this
##    1       0    1   0   1        0        1   0    1
##    2       1    0   1   0        1        2   1    1
```

Láthatjuk, hogy itt a szavak gyakoriságát (egyfajta hisztogramként) számoljuk össze. A dokumentumban található szavakat gyakoriság szerint lekérdezhetjük:

```
findFreqTerms(dtm,2)
```

```
## [1] "sentence" "this"
```

Vagy éppen ábrázolhatjuk őket gyakoriság szerint:

```
plot.wordcloud<-function(dtm) {
  m<-as.matrix(dtm)
  v<-sort(colSums(m),decreasing=T)

  suppressWarnings(wordcloud(words = names(v), freq = v, min.freq = 1,
            max.words=100, random.order=FALSE, rot.per=0.35,
            colors=brewer.pal(8, "Dark2")))
}
plot.wordcloud(dtm)
```

Gyakori feladat, hogy az egyes szavak súlyozását meg kell változtatni, például TF-IDF súlyozásra. A TF-IDF súlyozás képlete:

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log \frac{N}{|\{d \in D : t \in d\}|}$$

```r
dtm_tfidf<-weightTfIdf(dtm)
inspect(dtm_tfidf)
```

```
## <<DocumentTermMatrix (documents: 2, terms: 8)>>
## Non-/sparse entries: 6/10
## Sparsity           : 62%
## Maximal term length: 8
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
## Sample             :
##     Terms
## Docs   another hell       not  one  previous sentence       the this
##    1 0.0000000 0.25 0.0000000 0.25 0.0000000        0 0.0000000    0
##    2 0.1428571 0.00 0.1428571 0.00 0.1428571        0 0.1428571    0
```

Tekintsünk egy összetettebb adathalmazt, amely a Reuters hírügynökség 20 cikkét tartalmazza a kőolajról:

```r
data("crude")
print(crude)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 20
```

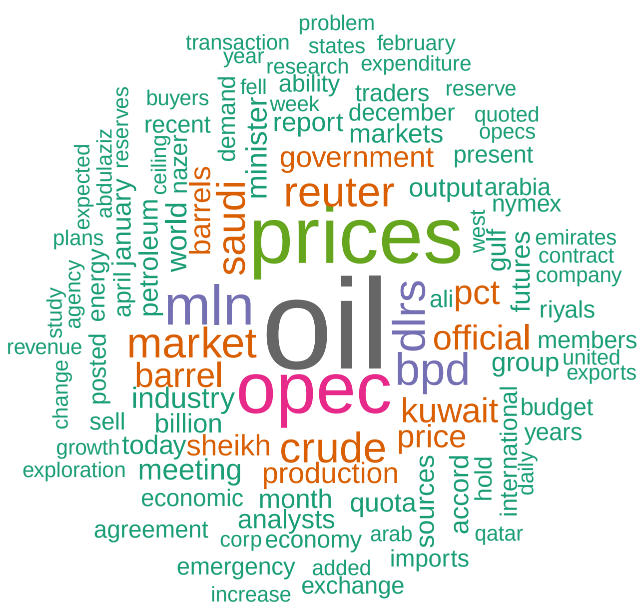
```r
corpus<-crude

clean_corpus<-function(corpus)
{
  corpus<-tm_map(corpus,content_transformer(tolower))
  corpus<-tm_map(corpus,removeNumbers)
  corpus<-tm_map(corpus,removePunctuation)
  corpus<-tm_map(corpus,removeWords,stopwords('SMART'))
  corpus<-tm_map(corpus,stripWhitespace)
  corpus
}
corpus<-clean_corpus(corpus)
```

```r
dtm<-DocumentTermMatrix(corpus)
inspect(dtm)
```

```
## <<DocumentTermMatrix (documents: 20, terms: 862)>>
## Non-/sparse entries: 1461/15779
## Sparsity           : 92%
## Maximal term length: 16
## Weighting          : term frequency (tf)
## Sample             :
##      Terms
## Docs  bpd crude dlrs market mln oil opec prices reuter saudi
##   144   4     0    0      3   4  12   13      5      1     0
##   236   7     2    2      0   4   7    6      5      1     0
##   237   0     0    1      0   1   3    1      1      1     0
##   242   0     0    0      2   0   3    2      2      1     1
##   246   0     0    0      0   0   5    1      1      1     0
##   248   2     0    4      8   3   9    6      9      1     5
##   273   8     5    2      1   9   5    5      5      1     7
##   489   0     0    1      0   3   4    0      2      1     0
##   502   0     0    1      0   3   5    0      2      1     0
##   704   0     0    0      2   0   3    0      3      1     0
```

```r
plot.wordcloud(dtm)
```




```r
findFreqTerms(dtm,10)
```

```
##  [1] "barrel"     "barrels"    "bpd"        "crude"      "dlrs"
##  [6] "government" "industry"   "kuwait"     "market"     "meeting"
## [11] "minister"   "mln"        "official"   "oil"        "opec"
## [16] "pct"        "price"      "prices"     "production" "reuter"
## [21] "saudi"      "sheikh"     "world"
```

```r
findAssocs(dtm,"oil",0.7)
```

```
## $oil
##      opec     named      late    prices    winter   markets  analysts agreement
```

```
##      0.87      0.81      0.79      0.79      0.79      0.78      0.77      0.76
## emergency    buyers     fixed
##      0.74      0.71      0.71
```

```r
findAssocs(dtm,"winter",0.7)
```

```
## $winter
##      late agreement    market     fixed     named       oil      opec  analysts
##      1.00      0.88      0.81      0.79      0.79      0.79      0.77      0.76
##   markets emergency
##      0.73      0.72
```

Példa (USA elnöki beszédek)

Első lépésként, tekintsük át az adatbázist, ismerjük meg az adatokat.

```r
conn<-dbConnect(RSQLite::SQLite(),"/opt/datasets/sotu.db")
dbListTables(conn)
```

```
## [1] "speach"
```

```r
dbListFields(conn,"speach")
```

```
## [1] "date"       "president"  "title"      "url"        "transcript"
```

```r
dbGetQuery(conn,"SELECT date,president,title FROM speach LIMIT 1")
```

```
##         date       president
## 1 2018-01-30 Donald J. Trump
##                                                                      title
## 1 Address Before a Joint Session of the Congress on the State of the Union
```

```r
dbGetQuery(conn,"SELECT president,count(*) FROM speach GROUP BY president")
```

```
##                president count(*)
## 1        Abraham Lincoln        4
## 2         Andrew Jackson        8
## 3         Andrew Johnson        4
## 4           Barack Obama        8
## 5      Benjamin Harrison        4
## 6        Calvin Coolidge        6
## 7      Chester A. Arthur        4
## 8        Donald J. Trump        2
## 9   Dwight D. Eisenhower       10
## 10 Franklin D. Roosevelt       13
## 11       Franklin Pierce        4
## 12           George Bush        4
## 13        George W. Bush        8
## 14     George Washington        8
## 15        Gerald R. Ford        3
## 16      Grover Cleveland        8
## 17       Harry S. Truman        8
## 18        Herbert Hoover        4
## 19        James Buchanan        4
## 20         James K. Polk        4
## 21        James Madison        8
## 22          James Monroe        8
```

```
## 23          Jimmy Carter        7
## 24            John Adams        4
## 25       John F. Kennedy        3
## 26     John Quincy Adams        4
## 27            John Tyler        4
## 28     Lyndon B. Johnson        6
## 29      Martin van Buren        4
## 30      Millard Fillmore        3
## 31         Richard Nixon       12
## 32         Ronald Reagan        8
## 33   Rutherford B. Hayes        4
## 34    Theodore Roosevelt        8
## 35      Thomas Jefferson        8
## 36      Ulysses S. Grant        8
## 37     Warren G. Harding        2
## 38   William Howard Taft        4
## 39    William J. Clinton        8
## 40      William McKinley        4
## 41        Woodrow Wilson        8
## 42        Zachary Taylor        1
```

```
dbDisconnect(conn)
```

Hasonlítsuk össze Barack Obama és George W. Bush beszédeit, emeljük ki a legfontosabb különbségeket!

```
conn<-dbConnect(RSQLite::SQLite(),"/opt/datasets/sotu.db")
results_obama<-dbGetQuery(conn,paste0("SELECT * from speach where president = 'Barack Obama'"))
results_trump<-dbGetQuery(conn,paste0("SELECT * from speach where president = 'George W. Bush'"))
dbDisconnect(conn)

results<-rbind(results_obama, results_trump)
results$president <- as.factor(results$president)

content<-gsub("[^a-zA-Z0-9 .]","",results$transcript)
corpus<-Corpus(VectorSource(content))
corpus<-clean_corpus(corpus)
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents

## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents

## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("SMART")):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, stripWhitespace): transformation drops
## documents
```

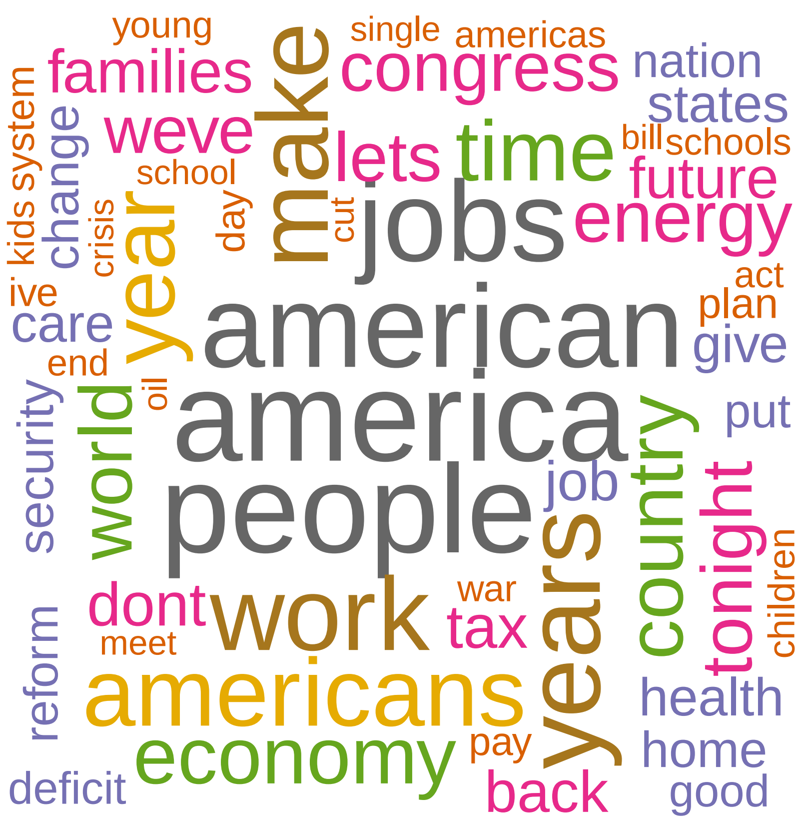
```
dtm<-DocumentTermMatrix(corpus)
print(dtm)
```

```
## <<DocumentTermMatrix (documents: 16, terms: 6996)>>
## Non-/sparse entries: 21130/90806
## Sparsity           : 81%
```

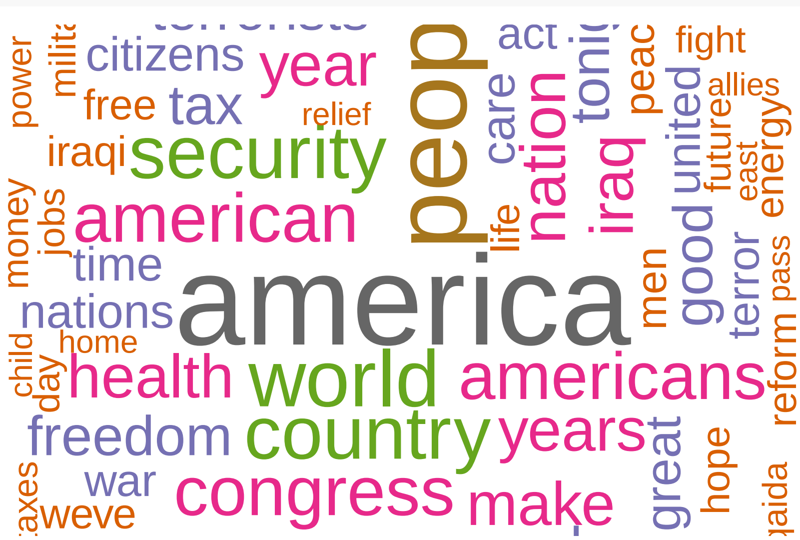
```
## Maximal term length: 31
## Weighting          : term frequency (tf)
```

```r
dtm_list<-split(dtm,results$president)

plot.wordcloud(dtm_list$`Barack Obama`)
```




```r
plot.wordcloud(dtm_list$`George W. Bush`)
```



Láthatjuk, hogy a beszédek kulcsszavai nagyon hasonlóak. Alkalmazzuk a TF-IDF súlyozást, hogy erőteljesebben tudjunk fókuszálni a különbségekre!

```r
dtm<-DocumentTermMatrix(corpus, control=list(weighting=weightTfIdf))
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom functions are
## ignored
```

```r
print(dtm)
```

```
## <<DocumentTermMatrix (documents: 16, terms: 6996)>>
## Non-/sparse entries: 19898/92038
## Sparsity           : 82%
## Maximal term length: 31
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

```r
dtm_list<-split(dtm,results$president)

plot.wordcloud(dtm_list$`Barack Obama`)
```


```r
plot.wordcloud(dtm_list$`George W. Bush`)
```

Láthatjuk, hogy az egyes beszédek jelentősen eltérnek, és jellemzően az adott érára jellemző szavakat tartalmazzák. Tekintsünk meg pár tipikus asszociációt a beszédekből:

```
findAssocs(dtm_list$`Barack Obama`,"minimum",0.9)
```

```
## $minimum
##       prices     americaa       amount capabilities    daughters       engage
##         0.95         0.93         0.93         0.93         0.93         0.93
##         feet     fulltime     hospital        hours        sandy       syrian
##         0.93         0.93         0.93         0.93         0.93         0.93
##         wage     shoulder          gun     stronger
##         0.92         0.91         0.90         0.90
```

```
findAssocs(dtm_list$`Barack Obama`,"banks",0.95)
```

```
## $banks
##    largest     ignore    healthy   affected   captured   extended    failure
##       0.97       0.97       0.96       0.95       0.95       0.95       0.95
##    familys      gains identified     leaves       lend      madam   overcome
##       0.95       0.95       0.95       0.95       0.95       0.95       0.95
##   preserve preventive    recover  restoring  shortterm    surplus      watch
##       0.95       0.95       0.95       0.95       0.95       0.95       0.95
##  yesterday
##       0.95
```

```
findAssocs(dtm_list$`George W. Bush`,'drug',0.92)
```

```
## $drug
##    reduced   marriage    propose    patriot    meeting abstinence  placement
##       0.94       0.94       0.94       0.94       0.93       0.93       0.93
```

```
findAssocs(dtm_list$`George W. Bush`,'saddam',0.964)
```

```
## $saddam
##           hussein            threat         improving         treatment
##              0.99              0.97              0.97              0.97
##           arsenal              bold              date         discovery
##              0.96              0.96              0.96              0.96
##             ebola          enormous         existence       frustration
##              0.96              0.96              0.96              0.96
```

```
##         goodness            grasp             line           mobile
##             0.96             0.96             0.96             0.96
##      participate          profits         profound          putting
##             0.96             0.96             0.96             0.96
##         refugees      regulations           season           simple
##             0.96             0.96             0.96             0.96
##            stake           strike            turns         villages
##             0.96             0.96             0.96             0.96
##            yemen       aggression        alliances         companys
##             0.96             0.96             0.96             0.96
##         constant         decision          denying       friendship
##             0.96             0.96             0.96             0.96
##         hardship            hasnt             hunt          imagine
##             0.96             0.96             0.96             0.96
##         material            moved            paris       population
##             0.96             0.96             0.96             0.96
##            ports       prohibited            risen           slowly
##             0.96             0.96             0.96             0.96
##             sums        threatens             tide           untold
##             0.96             0.96             0.96             0.96
##           waited        announced         battered             bear
##             0.96             0.96             0.96             0.96
##            boost           church       efficiency         exercise
##             0.96             0.96             0.96             0.96
##              lay       litigation       negotiated       physicians
##             0.96             0.96             0.96             0.96
##    possibilities            rural           soviet          strikes
##             0.96             0.96             0.96             0.96
##         strongly             ties         visiting          western
##             0.96             0.96             0.96             0.96
##         arsenals             burn           deeply        describes
##             0.96             0.96             0.96             0.96
##        devastate          educate          gravest            moves
##             0.96             0.96             0.96             0.96
##          ordered        proposing         reckless        targeting
##             0.96             0.96             0.96             0.96
##            acres           agency         ambition              die
##             0.96             0.96             0.96             0.96
##        enlisting             fate             gate     intimidation
##             0.96             0.96             0.96             0.96
##             labs            model           notice           relied
##             0.96             0.96             0.96             0.96
##          warfare         airports          burning          crucial
##             0.96             0.96             0.96             0.96
##         dictator        enriching        excessive       laboratory
##             0.96             0.96             0.96             0.96
##            merge      surrounding          familys      immediately
##             0.96             0.96             0.96             0.96
##        louisiana            seeks           builds        documents
##             0.96             0.96             0.96             0.96
##          possess        reckoning       tremendous         abortion
##             0.96             0.96             0.96             0.96
##        accompany        accounted             acid        afflicted
##             0.96             0.96             0.96             0.96
```

```
##             agent             alarm          aluminum        americayou
##              0.96              0.96              0.96              0.96
##           analyze          antidrug    antiretroviral          appeared
##              0.96              0.96              0.96              0.96
##        assembling            atomic         attempted         awareness
##              0.96              0.96              0.96              0.96
##            banned           banning             baton         benchmark
##              0.96              0.96              0.96              0.96
##           binding         bioshield            bitter       blackmailed
##              0.96              0.96              0.96              0.96
##          blocking          bombings         botulinum           buffalo
##              0.96              0.96              0.96              0.96
##       bureaucrats          canister         caribbean           carries
##              0.96              0.96              0.96              0.96
##          casualty         cataloged           clarity           coached
##              0.96              0.96              0.96              0.96
##              cole commandandcontrol     companionship       concessions
##              0.96              0.96              0.96              0.96
##         concluded       confessions          confound          conquest
##              0.96              0.96              0.96              0.96
##        considered      conspiracies           consult         contained
##              0.96              0.96              0.96              0.96
##          contempt           contest          contrary             crate
##              0.96              0.96              0.96              0.96
##          credibly           cruelty            cruise         deceiving
##              0.96              0.96              0.96              0.96
##       declaration          declines         defectors          defended
##              0.96              0.96              0.96              0.96
##           designs         diagnosis          dictates      difficulties
##              0.96              0.96              0.96              0.96
##          director         disarming         disclosed        disfigured
##              0.96              0.96              0.96              0.96
##          dividend        domination             doses             dread
##              0.96              0.96              0.96              0.96
##            drills          dripping          duration            easily
##              0.96              0.96              0.96              0.96
##         elaborate         embassies            employ             entry
##              0.96              0.96              0.96              0.96
##     environmental          estimate             evade           exhaust
##              0.96              0.96              0.96              0.96
##        experiment         explained       explanation           fairest
##              0.96              0.96              0.96              0.96
##        fatherless            fights      fingerprints             fires
##              0.96              0.96              0.96              0.96
##           flights           forests             fumes           gaining
##              0.96              0.96              0.96              0.96
##              germ         gibraltar          guidance           hamburg
##              0.96              0.96              0.96              0.96
##            healed        hesitation            hiding      highstrength
##              0.96              0.96              0.96              0.96
##         hitlerism              hmos      hopelessness            hormuz
##              0.96              0.96              0.96              0.96
##            horror   hydrogenpowered          imminent            incite
##              0.96              0.96              0.96              0.96
```

```
##         incometax           infants         infection       inoculating
##              0.96              0.96              0.96              0.96
##        inspection          instance       instructing       integration
##              0.96              0.96              0.96              0.96
##       intensified         interview          investor   invulnerability
##              0.96              0.96              0.96              0.96
##          iranians             irons            junior            knowwe
##              0.96              0.96              0.96              0.96
##           lawsuit        legaliraqs           lengths            lethal
##              0.96              0.96              0.96              0.96
##     lifeextending             links            liters          location
##              0.96              0.96              0.96              0.96
##         logistics            loving          mandates           manmade
##              0.96              0.96              0.96              0.96
##             mercy             milan        militarism          miracles
##              0.96              0.96              0.96              0.96
##        miraculous  misunderstanding        mobilizing        monitoring
##              0.96              0.96              0.96              0.96
##          mounting         munitions           mustard        mutilation
##              0.96              0.96              0.96              0.96
##      nationalized         ninetytwo            object         obstacles
##              0.96              0.96              0.96              0.96
##          obtained          onethird         operative           opinion
##              0.96              0.96              0.96              0.96
##          orphaned          overlook            oxygen      partialbirth
##              0.96              0.96              0.96              0.96
##         permitted      perseverance           placing            plague
##              0.96              0.96              0.96              0.96
##           planned          politely            posing            powell
##              0.96              0.96              0.96              0.96
##           powered             prize          prospect           pursued
##              0.96              0.96              0.96              0.96
##        quantities              rape           rations        recovering
##              0.96              0.96              0.96              0.96
##    recriminations       reluctantly       reorganized         represses
##              0.96              0.96              0.96              0.96
##         requested          resolute       respiratory        restrained
##              0.96              0.96              0.96              0.96
##         restraint            resume            reveal           revival
##              0.96              0.96              0.96              0.96
##          richness           risking             rouge             ruins
##              0.96              0.96              0.96              0.96
##            ruling        sanitizing            sanity             sarin
##              0.96              0.96              0.96              0.96
##         scattered         scavenger         screeners            search
##              0.96              0.96              0.96              0.96
##          secretly            seldom           sensors          sentence
##              0.96              0.96              0.96              0.96
##           shadowy       shareholder             ships          showroom
##              0.96              0.96              0.96              0.96
##     significantly         singapore             sites          smallpox
##              0.96              0.96              0.96              0.96
##              soul           sparing            staffs        stagnation
##              0.96              0.96              0.96              0.96
```

```
##        statements          straits        strangers         suddenly
##              0.96             0.96             0.96             0.96
##              sued       sufficient         suitable   systematically
##              0.96             0.96             0.96             0.96
##          taxation          tongues             tons        torturing
##              0.96             0.96             0.96             0.96
##            towers            toxin     transforming         treasured
##              0.96             0.96             0.96             0.96
##          treaties          triumph            tubes           twelve
##              0.96             0.96             0.96             0.96
##           upwards              uss            utter             vial
##              0.96             0.96             0.96             0.96
##       victimsonly            virus          viruses          weakest
##              0.96             0.96             0.96             0.96
##         whirlwind         wildlife        witnesses    wonderworking
##              0.96             0.96             0.96             0.96
```

Próbáljuk meg kategorizálni és összehasonlítani az összes beszédet az 1990 évek eleje óta!

```
conn<-dbConnect(RSQLite::SQLite(),"/opt/datasets/sotu.db")
results<-dbGetQuery(conn,"SELECT * from speach where date>'1900.01.01'")
dbDisconnect(conn)

content<-gsub("[^a-zA-Z0-9 .]","",results$transcript)
corpus<-Corpus(VectorSource(content))
corpus<-clean_corpus(corpus)
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents

## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents

## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("SMART")):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, stripWhitespace): transformation drops
## documents
```

```
dtm<-DocumentTermMatrix(corpus)
print(dtm)
```

```
## <<DocumentTermMatrix (documents: 132, terms: 26581)>>
## Non-/sparse entries: 191568/3317124
## Sparsity           : 95%
## Maximal term length: 32
## Weighting          : term frequency (tf)
```

```
dtm<-removeSparseTerms(dtm,0.95)
print(dtm)
```

```
## <<DocumentTermMatrix (documents: 132, terms: 5919)>>
## Non-/sparse entries: 154221/627087
## Sparsity           : 80%
## Maximal term length: 17
## Weighting          : term frequency (tf)
```

```
plot.wordcloud(dtm)
```


```
dtm<-DocumentTermMatrix(corpus, control=list(weighting = weightTfIdf))
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom functions are
## ignored
```

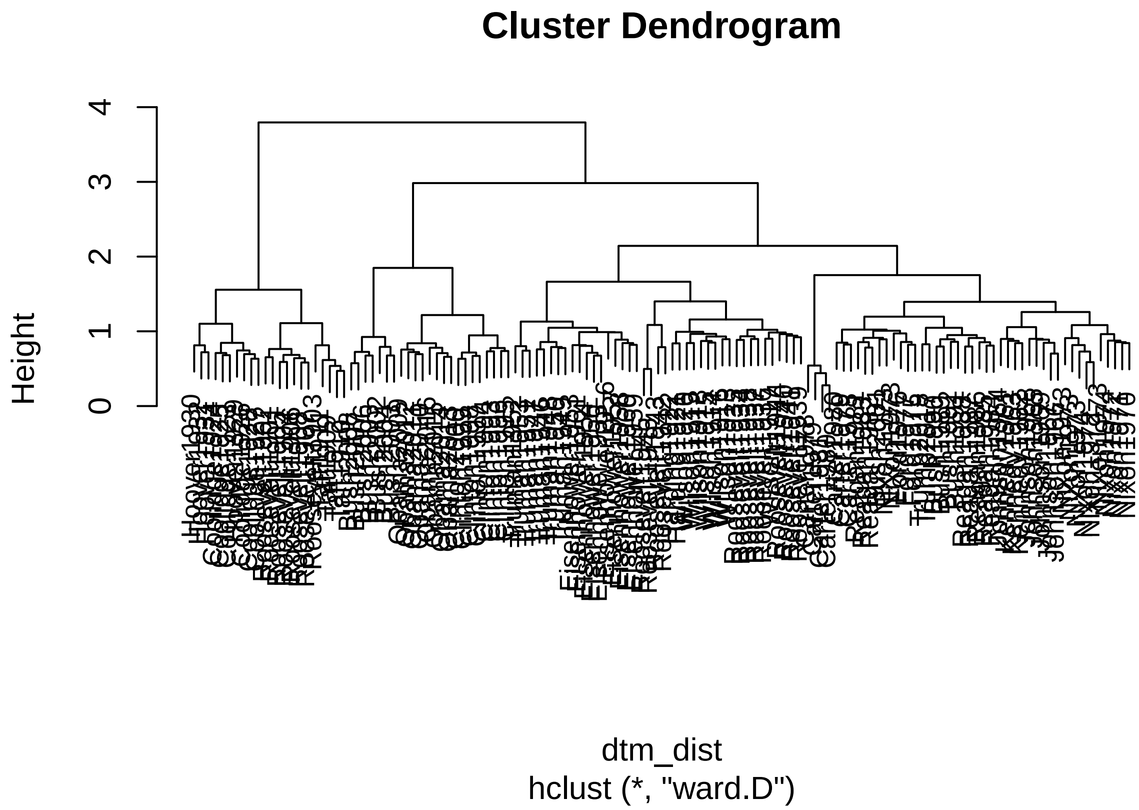
```
lastnames<-sapply(strsplit(results$president, ' '), function(x) x[length(x)])
years<-format(as.Date(results$date),'%Y')

dtm$dimnames$Docs <- paste0(lastnames,years)
```

Készítsük el a távolságmátrixot, és aztán alkalmazzunk hierarchikus klaszterezést az egyes beszédekre.

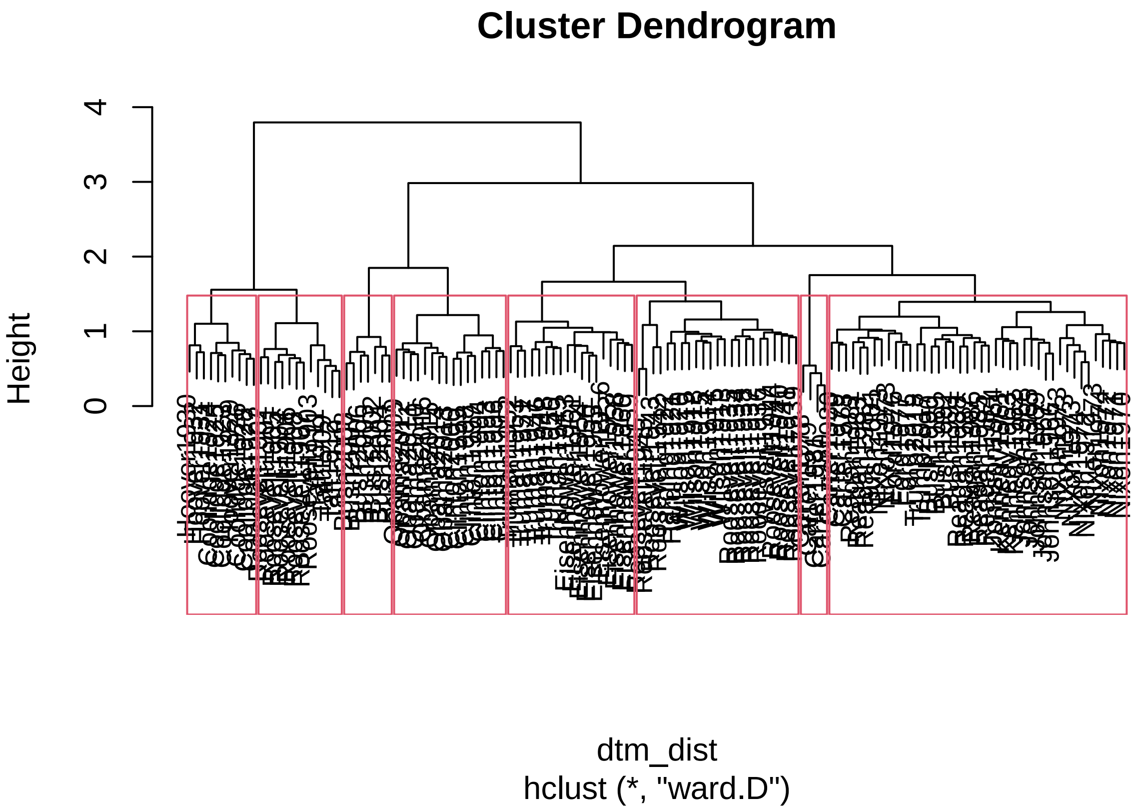
```
dtm_mat<-as.matrix(dtm)
cosine_distance<-function(mat)
{
  dist<-mat/sqrt(rowSums(mat*mat))
  dist<-dist %*% t(dist)
  as.dist(1-dist)
}
dtm_dist<-cosine_distance(dtm_mat)
```

```
h<-hclust(dtm_dist,method="ward.D")
plot(h,cex=0.8)
```



Ábrázoltuk a dendogrammot, célszerű kiválasztani egy vágatot, tekintsünk 8 klasztert:

```
plot(h,cex=0.8)
cls<-rect.hclust(h,8)
```



Ha ezeket a klasztereket megvizsgáljuk, és ábrázoljuk a hozzá tartozó tipikus szófelhőket, akkor azt látjuk, hogy minden érának jól megkülönböztethető szókészlete és kulcsszavai voltak:

```
sapply(cls, function(l) {
  plot.wordcloud(dtm[l,])
  names(l)
  })
```

```
## [[1]]
##  [1] "Hoover1932"   "Hoover1931"   "Hoover1930"   "Hoover1929"   "Coolidge1928"
##  [6] "Coolidge1927" "Coolidge1926" "Coolidge1925" "Coolidge1924" "Coolidge1923"
##
## [[2]]
##  [1] "Taft1912"      "Taft1911"      "Taft1910"      "Taft1909"
##  [5] "Roosevelt1908" "Roosevelt1907" "Roosevelt1906" "Roosevelt1905"
##  [9] "Roosevelt1904" "Roosevelt1903" "Roosevelt1902" "Roosevelt1901"
##
## [[3]]
## [1] "Bush2008" "Bush2007" "Bush2006" "Bush2005" "Bush2004" "Bush2003" "Bush2002"
##
## [[4]]
##  [1] "Obama2016"   "Obama2015"   "Obama2014"   "Obama2013"   "Obama2012"
##  [6] "Obama2011"   "Obama2010"   "Obama2009"   "Clinton2000" "Clinton1999"
## [11] "Clinton1998" "Clinton1997" "Clinton1996" "Clinton1995" "Clinton1994"
## [16] "Clinton1993"
##
## [[5]]
##  [1] "Eisenhower1961" "Eisenhower1960" "Eisenhower1959" "Eisenhower1958"
##  [5] "Eisenhower1957" "Eisenhower1956" "Eisenhower1956" "Eisenhower1955"
```

```
##  [9] "Eisenhower1954" "Eisenhower1953" "Truman1953"     "Truman1952"
## [13] "Truman1951"     "Truman1950"     "Truman1949"     "Truman1948"
## [17] "Truman1947"     "Truman1946"
##
## [[6]]
##  [1] "Roosevelt1945" "Roosevelt1945" "Roosevelt1944" "Roosevelt1943"
##  [5] "Roosevelt1942" "Roosevelt1941" "Roosevelt1940" "Roosevelt1939"
##  [9] "Roosevelt1938" "Roosevelt1937" "Roosevelt1936" "Roosevelt1935"
## [13] "Roosevelt1934" "Harding1922"   "Harding1921"   "Wilson1920"
## [17] "Wilson1919"    "Wilson1918"    "Wilson1917"    "Wilson1916"
## [21] "Wilson1915"    "Wilson1914"    "Wilson1913"
##
## [[7]]
## [1] "Carter1981" "Carter1980" "Carter1979" "Carter1978"
##
## [[8]]
##  [1] "Trump2018"   "Trump2017"   "Bush2001"    "Bush1992"    "Bush1991"
##  [6] "Bush1990"    "Bush1989"    "Reagan1988"  "Reagan1987"  "Reagan1986"
## [11] "Reagan1985"  "Reagan1984"  "Reagan1983"  "Reagan1982"  "Reagan1981"
## [16] "Carter1980"  "Carter1979"  "Carter1978"  "Ford1977"    "Ford1976"
## [21] "Ford1975"    "Nixon1974"   "Nixon1974"   "Nixon1973"   "Nixon1973"
## [26] "Nixon1973"   "Nixon1973"   "Nixon1973"   "Nixon1973"   "Nixon1972"
## [31] "Nixon1972"   "Nixon1971"   "Nixon1970"   "Johnson1969" "Johnson1968"
## [36] "Johnson1967" "Johnson1966" "Johnson1965" "Johnson1964" "Kennedy1963"
## [41] "Kennedy1962" "Kennedy1961"
```