<u>Assignment 5</u>

Part 1

Case: The CEO has charged your analytics groups with a task: "What are the characteristics of an organization that adapts well to data analytics?" The CEO's intention is to restructure the company to foster adoption and advancement of data analytics capabilities across units.

Steps: Find a list of 30 successful data-driven companies; Create a corpus of their mission statement; Create a corpus of their core values; Analyze the corpus and provide insight on how to structure a firm for data-analysis readiness; Are there any other data-driven approaches you would recommend the CEO to implement?

LOADING REQUIRED PACKAGES

```
Console ~/Desktop/
> library("quanteda", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
quanteda version 0.9.8.5


Attaching package: 'quanteda'

The following object is masked from 'package:base':

    sample

> library("slam", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
> library("stm", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
stm v1.1.3 (2016-01-14) successfully loaded. See ?stm for help.
> library("tm", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
Loading required package: NLP

Attaching package: 'NLP'

The following object is masked from 'package:quanteda':

    ngrams


Attaching package: 'tm'

The following objects are masked from 'package:quanteda':

    as.DocumentTermMatrix, stopwords
```

EXPLORING THE LIST OF 30 DATA-DRIVEN COMPANIES

```
> library("openNLP", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
> #load .csv file with news articles
> url<-
+   orgcorpus<- read.csv("/Users/anushiarora/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/Organization Information - Sheet1.csv",
+                        header=TRUE, stringsAsFactors=FALSE)
> dim(orgcorpus)
[1] 30  3
> names(orgcorpus)
[1] "Organization.Name" "Mission.Statement" "Core.Values"
```

```
> head(orgcorpus)
  Organization.Name
1    InsightSquared
2          Trifacta
3          Cloudera
4             Birst
5        Gainsight
6            Google
                                                                                                                         Mission.Statement
1                                                     We want to solve big business problems, with the best minds, in an insanely collaborative and supportive environment.
2                                                                                  Enable you to quickly transform Big Data from a burden into a strategic asset.
3                                      Bringing the power of Hadoop, MapReduce, and distributed storage to companies of all sizes in the enterprise, Internet and government sectors.
4 At Birst we\x89.re giving business meaning to data \x89\xdb\xd2 by enabling business users to analyze all of their data, from all types of sources, to solve real problems. Fast.
5                                                                                                Prevent churn, Increase Upsell and Create happy customers for life.
6                                                                                         To organize the world\x89.s information and make it universally accessible and useful.

Core.Values
1
Problem Solvers; Taking Ownership; Team Driven; Positive; Nimble, Work Hard
2
Passionate; Commited; Honest; Innovators; Providing Best Solutions
3
Open to new ideas; Believe in transperency with customer;Creating growth opportunities for employees; optimizing globally; Innovator; Courageous
4
Unique perspective; Innovative; Efficient
5
The Golden Rule; Success of all Stakeholders; Child-like Joy
6 Focus on the user and all else will follow; It is best to do one thing really, really well; Fast is better than slow; Democracy on the web works; You don't need to be at your desk to need an an
swer; You can make money without doing evil; There's always more information out there; The need for information crosses all border; You can be serious without a suit; Great just isn't good enoug
h
> str(orgcorpus)
'data.frame':   30 obs. of  3 variables:
 $ Organization.Name: chr  "InsightSquared" "Trifacta" "Cloudera" "Birst" ...
 $ Mission.Statement: chr  "We want to solve big business problems, with the best minds, in an insanely collaborative and supportive environment." "Enable you to quickly transform Big Data from a
burden into a strategic asset." "Bringing the power of Hadoop, MapReduce, and distributed storage to companies of all sizes in the enterprise, Internet and gove"| __truncated__ "At Birst we\x89.r
e giving business meaning to data \x89\xdb\xd2 by enabling business users to analyze all of their data, from al"| __truncated__ ...
 $ Core.Values      : chr  "Problem Solvers; Taking Ownership; Team Driven; Positive; Nimble, Work Hard" "Passionate; Commited; Honest; Innovators; Providing Best Solutions" "Open to new ideas; B
elieve in transperency with customer;Creating growth opportunities for employees; optimizing globally; Inno"| __truncated__ "Unique perspective; Innovative; Efficient" ...
```

CORPUS OF MISSION STATEMENT

```
Console ~/Desktop/
> # create a corpus for Mission Statements
> require(quantedа)
> missioncorpus<- corpus(orgcorpus$Mission.Statement,
+                        docnames=orgcorpus$Organization.Name)
> names(missioncorpus)   #to explore the output of the corpus function: "documents" "metadata"  "settings"  "tokens"
[1] "documents" "metadata"  "settings"  "tokens"
> summary(missioncorpus)  #summary of corpus
Corpus consisting of 30 documents.

             Text Types Tokens Sentences
   InsightSquared    20     21         1
         Trifacta    14     15         1
         Cloudera    20     25         1
            Birst    27     37         2
        Gainsight    12     12         1
           Google    14     15         1
           Visier    20     21         1
            Domo     23     26         1
 Tableau Software    24     26         3
        MarkLogic    32     44         2
      Informatica    10     11         1
          Platfora    13     13         1
              SAP     9      9         1
    SAS Institute    11     11         1
        Microsoft    13     14         1
              EMC    11     11         1
 Alpine Data Labs    25     25         1
          GoodData    16     19         1
              Qlik     8      8         1
        Salesforce    21     23         1
   Neo Technology    16     16         1
          Teradata    16     17         1
            Oracle    30     34         1
              Dell    18     21         1
  Hewlett-Packard    23     25         1
   TIBCO Software    13     14         1
            Splunk    13     13         1
          1010data    28     33         1
         SnapLogic    39     45         1
               IBM     4      4         1
```

```
Source:  /Users/anushiarora/Desktop/* on x86_64 by anushiarora
Created: Wed Nov 16 17:50:50 2016
Notes:

> head(missioncorpus)
                                                
```

```
Console ~/Desktop/
> #to create a custom dictionary list of stop words
>
> swlist = c("s", "make", "use", "see", "enabl", "big", "best", "way", "can", "everi", "re", "leverag", "want", "mind", "insan",
+            "bring", "type", "mean", "ve", "everyth", "even", "ipaa", "line")
>
> dfm.stem<- dfm(cleanmissioncorpus, toLower = TRUE,
+                ignoredFeatures = c(swlist, stopwords("english")),
+                verbose=TRUE,
+                stem=TRUE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 30 documents
   ... indexing features: 249 feature types
   ... removed 50 features, from 197 supplied (glob) feature types
   ... stemming features (English), trimmed 18 feature variants
   ... created a 30 x 181 sparse dfm
   ... complete.
Elapsed time: 0.048 seconds.
```

TOP 15 FREQUENT TERMS IN DFM OF MISSION STATEMENT

```
> topfeatures.stem<-topfeatures(dfm.stem, n=50)
> topfeatures.stem
       data        busi      custom       world        help       organ       peopl       enabl     compani
         12          10           9           6           6           5           5           4           4
      drive        valu   transform       power    enterpris       creat       empow      better       innov
          4           4           3           3           3           3           3           3           3
      deliv   technolog        solv     problem     environ        give  understand       manag    intellig
          3           3           2           2           2           2           2           2           2
      never      import        idea       solut      improv       everi     industri relationship      leverag
          2           2           2           2           2           2           2           2           2
    connect      servic    platform      applic        mind      insan     collabor     support       quick
          2           2           2           2           1           1           1           1           1
     burden     strateg       asset       bring      hadoop
          1           1           1           1           1
```

GENERATING WORD CLOUD FOR MISSION STATEMENT

```
> library(wordcloud)
> set.seed(142)
> dark2 <- brewer.pal(8, "Set1")
> freq<-topfeatures(dfm.stem, n=500)
>
>
> wordcloud(names(freq),
+           freq, max.words=200,
+           scale=c(3, .1),
+           colors=brewer.pal(8, "Set1"))
```

GENERATING BIGRAMS

```
Console ~/Desktop/
> #dfm with bigrams
>
> cleanmissioncorpusbigram <- tokenize(missioncorpus,
+                         removeNumbers=TRUE,
+                         removePunct = TRUE,
+                         removeSeparators=TRUE,
+                         removeTwitter=FALSE,
+                         ngrams=2, verbose=TRUE)
Starting tokenization...
  ...preserving Twitter characters (#, @)...total elapsed: 0 seconds.
  ...tokenizing texts...total elapsed:  0.00100000000020373 seconds.
  ...replacing Twitter characters (#, @)...total elapsed: 0 seconds.
  ...creating ngrams...total elapsed: 0.00499999999919964 seconds.
  ...replacing names...total elapsed:  0.00100000000020373 seconds.
Finished tokenizing and cleaning 30 texts.
>
> dfm.bigram<- dfm(cleanmissioncorpusbigram, toLower = TRUE,
+                  ignoredFeatures = c(swlist, stopwords("english")),
+                  verbose=TRUE,
+                  stem=FALSE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 30 documents
   ... indexing features: 457 feature types
   ... removed 355 features, from 197 supplied (glob) feature types
   ... created a 30 x 102 sparse dfm
   ... complete.
Elapsed time: 0.028 seconds.
```

```
> topfeatures.bigram<-topfeatures(dfm.bigram, n=50)
> topfeatures.bigram
    business_problems insanely_collaborative supportive_environment      quickly_transform       strategic_asset
                    1                      1                       1                      1                     1
     hadoop_mapreduce     distributed_storage     enterprise_internet     government_sectors      giving_business
                    1                      1                       1                      1                     1
     business_meaning       enabling_business        business_users           solve_real        real_problems
                    1                      1                       1                      1                     1
       problems_fast         prevent_churn        churn_increase       increase_upsell        create_happy
                    1                      1                       1                      1                     1
    happy_customers universally_accessible       empower_leaders      executives_manage        drives_value
                    1                      1                       1                      1                     1
traditional_business  business_intelligence  intelligence_systems          help_people          data_eight
                    1                      1                       1                      1                     1
         eight_words       drive_everything         team_inspired         people_search       search_create
                    1                      1                       1                      1                     1
communicate_research       help_businesses       businesses_grow          better_place       leading_data
                    1                      1                       1                      1                     1
     data_management     management_company     empower_customers fact-based_enterprises        sas_delivers
                    1                      1                       1                      1                     1
     delivers_proven       proven_solutions       drive_innovation    improve_performance       empower_every
                    1                      1                       1                      1                     1
```

CORRELATION BETWEEN THE TOPS 3 FREQUENCY TERMS IN MISSION STATEMENTS

```
> #specifying a correlation limit of 0.5
> dfm.tm<-convert(dfm.stem, to="tm")
> findAssocs(dfm.tm,
+            c("data", "busi", "help"),
+            corlimit=0.5)
$data
enabl
 0.55

$busi
   solv problem   birst    mean    user  analyz    type   sourc    real    fast  execut     ten billion  dollar   spent
   0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52    0.52
 tradit  system
   0.52    0.52

$help
    idea     team   inspir    chang   search communic research     grow    safer    place    peopl
    0.73     0.70     0.70     0.70     0.70     0.70     0.70     0.70     0.70     0.70     0.56
```

PROCESS THE DATA OF MISSION STATEMENTS

```
Console ~/Desktop/
> #Process the data of Mission Statementfor analysis.
> temp<-textProcessor(documents=orgcorpus$Mission.Statement, metadata = orgcorpus)
Building corpus...
Converting to Lower Case...
Removing stopwords...
Removing numbers...
Removing punctuation...
Stemming...
Creating Output...
> names(temp)  # produces:  "documents", "vocab", "meta", "docs.removed"
[1] "documents"    "vocab"        "meta"         "docs.removed"
> meta<-temp$meta
> vocab<-temp$vocab
> docs<-temp$documents
> out <- prepDocuments(docs, vocab, meta)
Removing 147 of 192 terms (147 of 290 tokens) due to frequency
Removing 1 Documents with No Words
Your corpus now has 29 documents, 45 terms and 143 tokens.
> docs<-out$documents
> vocab<-out$vocab
> meta <-out$meta
```

EXPLORING TOP 15 TOPICS OF MISSION STATEMENTS

```
Console ~/Desktop/
> #running stm for top 15 topics
> top15 <-stm(docs , vocab ,
+             K=15,
+             verbose=TRUE,
+             data=meta,
+             max.em.its=25)
Beginning Initialization.
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 1 (approx. per word bound = -4.067)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 2 (approx. per word bound = -3.982, relative change = 4.068e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 3 (approx. per word bound = -3.806, relative change = 2.458e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 4 (approx. per word bound = -3.744, relative change = 1.624e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 5 (approx. per word bound = -3.700, relative change = 1.182e-02)
```

```
Console ~/Desktop/
Topic 1: help, empow, innov, use, understand
 Topic 2: best, intellig, solv, problem, big
 Topic 3: custom, big, platform, improv, solut
 Topic 4: make, deliv, better, use, solut
 Topic 5: transform, enterpris, problem, use, understand
 Topic 6: data, enabl, problem, solv, give
 Topic 7: creat, servic, connect, data, platform
 Topic 8: world, leverag, solv, improv, help
 Topic 9: busi, way, manag, solv, problem
 Topic 10: compani, can, manag, solut, peopl
 Topic 11: organ, technolog, see, understand, improv
 Topic 12: drive, enabl, give, solv, problem
 Topic 13: valu, relationship, industry_, solut, connect
 Topic 14: peopl, never, import, idea, understand
 Topic 15: power, innov, environ, solut, improv
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 6 (approx. per word bound = -3.665, relative change = 9.296e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 7 (approx. per word bound = -3.638, relative change = 7.519e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 8 (approx. per word bound = -3.614, relative change = 6.545e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 9 (approx. per word bound = -3.591, relative change = 6.460e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 10 (approx. per word bound = -3.568, relative change = 6.382e-03)
```

```
Console ~/Desktop/
Topic 1: help, empow, innov, creat, idea
 Topic 2: best, intellig, problem, solv, environ
 Topic 3: custom, big, platform, deliv, solut
 Topic 4: make, deliv, better, use, innov
 Topic 5: transform, enterpris, big, problem, busi
 Topic 6: data, enabl, big, solv, platform
 Topic 7: servic, connect, creat, industry_, platform
 Topic 8: world, improv, leverag, technolog, idea
 Topic 9: busi, way, manag, solv, intellig
 Topic 10: compani, solut, manag, deliv, world
 Topic 11: organ, technolog, industry_, innov, see
 Topic 12: drive, give, enabl, platform, solv
 Topic 13: valu, relationship, industry_, connect, solut
 Topic 14: peopl, help, never, import, can
 Topic 15: power, innov, platform, environ, leverag
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 21 (approx. per word bound = -3.420, relative change = 2.757e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 22 (approx. per word bound = -3.409, relative change = 3.010e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 23 (approx. per word bound = -3.399, relative change = 3.104e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 24 (approx. per word bound = -3.389, relative change = 3.038e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Model Terminated Before Convergence Reached
```

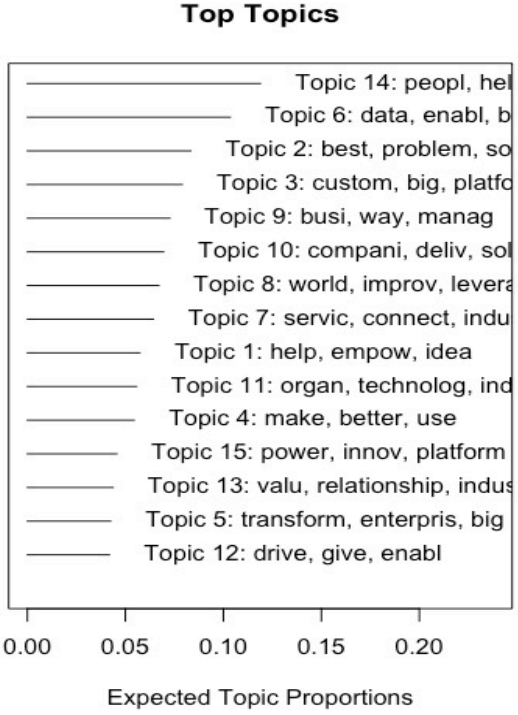
TOP 15 TOPICS OF MISSION STATEMENT

```
Console ~/Desktop/
> topics <-labelTopics(top15 , topics=c(1:15))
> topics
Topic 1 Top Words:
        Highest Prob: help, empow, idea, creat, innov, way, use
        FREX: empow, help, creat, idea, innov, way, better
        Lift: empow, idea, help, way, creat, innov, use
        Score: empow, help, idea, way, creat, innov, use
Topic 2 Top Words:
        Highest Prob: best, problem, solv, environ, intellig, big, busi
        FREX: best, problem, solv, environ, intellig, big, manag
        Lift: best, problem, solv, environ, intellig, big, busi
        Score: best, intellig, problem, environ, solv, big, busi
Topic 3 Top Words:
        Highest Prob: custom, big, platform, creat, deliv, innov, solut
        FREX: custom, big, platform, creat, innov, enterpris, deliv
        Lift: custom, big, platform, creat, deliv, innov, solut
        Score: custom, big, platform, creat, deliv, innov, solut
Topic 4 Top Words:
        Highest Prob: make, better, use, innov, deliv, creat, understand
        FREX: make, better, use, way, innov, creat, idea
        Lift: make, use, better, innov, deliv, creat, understand
        Score: make, use, better, deliv, innov, creat, understand
Topic 5 Top Words:
        Highest Prob: transform, enterpris, big, busi, empow, problem, manag
        FREX: transform, enterpris, big, manag, way, creat, drive
        Lift: transform, enterpris, big, empow, busi, problem, manag
        Score: transform, enterpris, big, busi, empow, problem, manag
Topic 6 Top Words:
        Highest Prob: data, enabl, big, give, solv, platform, problem
        FREX: data, enabl, big, platform, enterpris, better, creat
        Lift: enabl, data, big, give, solv, platform, problem
        Score: data, enabl, big, give, solv, platform, problem
Topic 7 Top Words:
        Highest Prob: servic, connect, industry_, creat, big, platform, better
        FREX: servic, connect, industry_, creat, better, big, enterpris
        Lift: servic, connect, industry_, creat, platform, big, better
        Score: connect, servic, industry_, big, creat, platform, better
```

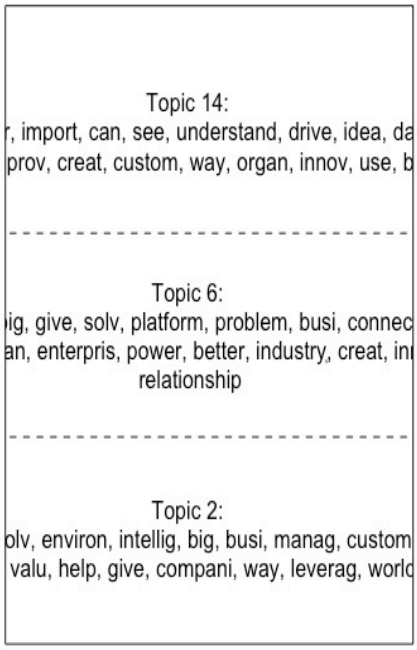
```
Console ~/Desktop/
Topic 8 Top Words:
        Highest Prob: world, improv, leverag, technolog, idea, help, deliv
        FREX: improv, world, leverag, technolog, idea, help, innov
        Lift: improv, leverag, world, technolog, idea, help, deliv
        Score: world, leverag, improv, technolog, idea, help, peopl
Topic 9 Top Words:
        Highest Prob: busi, way, manag, intellig, valu, transform, solv
        FREX: busi, way, manag, intellig, big, creat, drive
        Lift: busi, way, manag, intellig, solv, valu, transform
        Score: busi, way, manag, intellig, valu, solv, transform
Topic 10 Top Words:
        Highest Prob: compani, deliv, solut, manag, world, relationship, improv
        FREX: compani, deliv, solut, manag, innov, enterpris, drive
        Lift: compani, solut, deliv, manag, world, relationship, improv
        Score: compani, deliv, solut, manag, world, relationship, improv
Topic 11 Top Words:
        Highest Prob: organ, technolog, industry_, innov, creat, data, see
        FREX: organ, technolog, creat, innov, industry_, better, help
        Lift: organ, technolog, industry_, innov, creat, see, data
        Score: organ, technolog, industry_, innov, creat, see, can
Topic 12 Top Words:
        Highest Prob: drive, give, enabl, platform, solv, data, problem
        FREX: drive, give, platform, manag, big, innov, enabl
        Lift: give, drive, enabl, platform, solv, problem, manag
        Score: give, drive, enabl, platform, solv, problem, data
Topic 13 Top Words:
        Highest Prob: valu, relationship, industry_, connect, solut, servic, organ
        FREX: valu, relationship, industry_, creat, innov, connect, technolog
        Lift: relationship, valu, industry_, connect, solut, servic, technolog
        Score: valu, relationship, industry_, connect, servic, solut, organ
Topic 14 Top Words:
        Highest Prob: peopl, help, never, import, can, see, understand
        FREX: never, peopl, import, can, help, see, understand
        Lift: never, import, can, peopl, see, understand, help
        Score: peopl, import, never, help, can, see, understand
Topic 15 Top Words:
        Highest Prob: power, innov, platform, environ, give, deliv, leverag
        FREX: power, innov, platform, environ, big, enterpris, drive
        Lift: power, innov, platform, environ, give, leverag, solut
        Score: power, innov, platform, environ, give, leverag, solut
```

PLOT FOR TOP 15 TOPICS OF MISSION STATEMENT

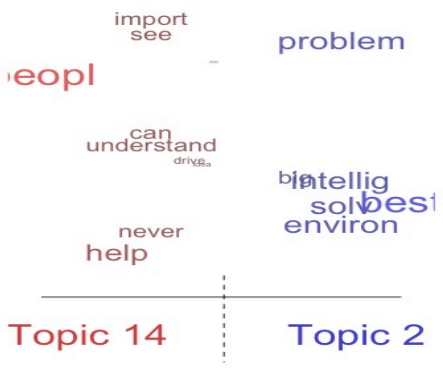
```
> plot.STM(top15, type="summary")
```




```
> plot.STM(top15, type="labels", topics=c(14,6,2))
```

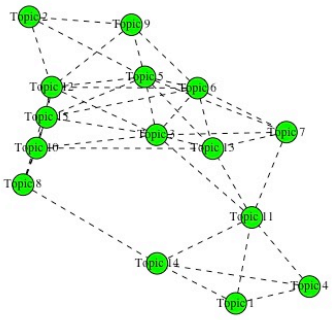



```
> plot.STM(top15, type="perspectives", topics = c(14,2))
```



CORRELATION BETWEEN TOP 15 TOPICS OF MISSION STATEMENT

```
> # to aid on assigment of labels & intepretation of topics
>
> mod.out.corr <- topicCorr(top15)
> plot.topicCorr(mod.out.corr)
```



CREATE A CORPUS FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> head(valuescorpus)

InsightSquared
                                                                    "Problem Solvers; Taking Ownership; Team Driven; Posit
ive; Nimble, Work Hard"

Trifacta
                                                                    "Passionate; Commited; Honest; Innovators; Pr
oviding Best Solutions"

Cloudera
"Open to new ideas; Believe in transperency with customer;Creating growth opportunities for employees; optimizing globally;
Innovator; Courageous"

Birst
                                                                    "Unique perspective;
Innovative; Efficient"
```

CLEANING CORPUS OF CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> # Cleaning corpus of Core Values : removes punctuation, digits, converts to lower case
>
> valuescorpus<- toLower(valuescorpus, keepAcronyms = FALSE)
> cleanvaluescorpus <- tokenize(valuescorpus,
+                               removeNumbers=TRUE,
+                               removePunct = TRUE,
+                               removeSeparators=TRUE,
+                               removeTwitter=FALSE,
+                               verbose=TRUE)
Starting tokenization...
  ...preserving Twitter characters (#, @)...total elapsed: 0 seconds.
  ...tokenizing texts...total elapsed:  0.00199999999999534 seconds.
  ...replacing Twitter characters (#, @)...total elapsed: 0 seconds.
  ...replacing names...total elapsed:  0 seconds.
Finished tokenizing and cleaning 30 texts.
> head(cleanvaluescorpus)
$InsightSquared
 [1] "problem"   "solvers"   "taking"    "ownership" "team"      "driven"    "positive"  "nimble"    "work"
[10] "hard"

$Trifacta
[1] "passionate" "commited"   "honest"     "innovators" "providing"  "best"       "solutions"

$Cloudera
 [1] "open"          "to"            "new"           "ideas"         "believe"       "in"            "transperency"
 [8] "with"          "customer"      "creating"      "growth"        "opportunities" "for"           "employees"
[15] "optimizing"    "globally"      "innovator"     "courageous"

$Birst
[1] "unique"      "perspective" "innovative"  "efficient"

$Gainsight
[1] "the"         "golden"       "rule"         "success"      "of"           "all"          "stakeholders"
[8] "child-like"  "joy"

$Google
 [1] "focus"       "on"          "the"         "user"        "and"         "all"         "else"        "will"
 [9] "follow"      "it"          "is"          "best"        "to"          "do"          "one"         "thing"
[17] "really"      "really"      "well"        "fast"        "is"          "better"      "than"        "slow"
[25] "democracy"   "on"          "the"         "web"         "works"       "you"         "don't"       "need"
[33] "to"          "be"          "at"          "your"        "desk"        "to"          "need"        "an"
[41] "answer"      "you"         "can"         "make"        "money"       "without"     "doing"       "evil"
[49] "there's"     "always"      "more"        "information" "out"         "there"       "the"         "need"
[57] "for"         "information" "crosses"     "all"         "border"      "you"         "can"         "be"
[65] "serious"     "without"     "a"           "suit"        "great"       "just"        "isn't"       "good"
[73] "enough"
```

CREATE A DOCUMENT FREQUENCY MATRIX FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #create document feature matrix for core values
>
> dfm.simple1<- dfm(cleanvaluescorpus,
+                   toLower = TRUE,
+                   ignoredFeatures = stopwords("english"),
+                   stem=TRUE,
+                   verbose=FALSE)
> head(dfm.simple1)
Document-feature matrix of: 30 documents, 182 features.
(showing first 6 documents and first 6 features)
                features
docs             problem solver take ownership team driven
  InsightSquared       1      1    1         1    1      1
  Trifacta             0      0    0         0    0      0
  Cloudera             0      0    0         0    0      0
  Birst                0      0    0         0    0      0
  Gainsight            0      0    0         0    0      0
  Google               0      0    0         0    0      0
```

CREATE CUSTOMIZED STOP WORDS FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #to create a custom dictionary list of stop words
>
> swlist = c("s", "make", "use", "see", "big", "insan", "bring", "type", "mean", "ve", "everyth", "even", "ipaa", "line", "open"
, "need", "take", "new", "without", "integr",
+           "inform", "end", "els", "will", "one", "thing", "well", "slow", "alway", "evil", "cross", "run", "move", "back")
>
> dfm.stem1<- dfm(cleanvaluescorpus, toLower = TRUE,
+                 ignoredFeatures = c(swlist, stopwords("english")),
+                 verbose=TRUE,
+                  stem=TRUE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 30 documents
   ... indexing features: 232 feature types
   ... removed 48 features, from 208 supplied (glob) feature types
   ... stemming features (English), trimmed 17 feature variants
   ... created a 30 x 167 sparse dfm
   ... complete.
Elapsed time: 0.163 seconds.
```

GENERATE TOP 50 FREQUENT TERMS OF CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> topfeatures.stem1<-topfeatures(dfm.stem1, n=50)
> topfeatures.stem1
      innov      custom      commit     success      integr       solut       dedic     reliabl    teamwork
         17          10           7           5           4           3           3           3           3
    respons     problem   ownership       posit        work     passion     provid        best       creat
          3           2           2           2           2           2           2           2           2
     growth    opportun     employe      global       uniqu     realli        fast      better         can
          2           2           2           2           2           2           2           2           2
     inform         win        busi    qualiti        valu  technolog     simplic     respect       speed
          2           2           2           2           2           2           2           2           2
      excel       maxim    challeng       agil trustworthi      divers    transpar      result         fun
          2           2           2           2           2           2           2           2           2
     togeth  leadership      solver        take        team
          2           2           1           1           1
```

GENERATING WORDCLOUD FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Anal
> #Generating wordcloud for core values
>
> library(wordcloud)
Loading required package: RColorBrewer
> set.seed(200)
> dark2 <- brewer.pal(8, "Set1")
> freq<-topfeatures(dfm.stem1, n=500)
>
>
> wordcloud(names(freq),
+           freq, max.words=400,
+           scale=c(3, .1),
+           colors=brewer.pal(8, "Set1"))
> #dfm with bigrams for core values
>
```

DFM FOR BIGRAMS FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #dfm with bigrams for core values
>
> cleanvaluescorpusbigram <- tokenize(valuescorpus,
+                                     removeNumbers=TRUE,
+                                     removePunct = TRUE,
+                                     removeSeparators=TRUE,
+                                     removeTwitter=FALSE,
+                                     ngrams=2, verbose=TRUE)
Starting tokenization...
  ...preserving Twitter characters (#, @)...total elapsed: 0 seconds.
  ...tokenizing texts...total elapsed:  0.0010000000000477 seconds.
  ...replacing Twitter characters (#, @)...total elapsed: 0.0010000000000477 seconds.
  ...creating ngrams...total elapsed: 0.00600000000000023 seconds.
  ...replacing names...total elapsed:  0 seconds.
Finished tokenizing and cleaning 30 texts.
>
> dfm.bigram<- dfm(cleanvaluescorpusbigram, toLower = TRUE,
+                  ignoredFeatures = c(swlist, stopwords("english")),
+                  verbose=TRUE,
+                  stem=FALSE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 30 documents
   ... indexing features: 320 feature types
   ... removed 168 features, from 208 supplied (glob) feature types
   ... created a 30 x 152 sparse dfm
   ... complete.
Elapsed time: 0.024 seconds.
```

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> topfeatures.bigram1<-topfeatures(dfm.bigram, n=50)
> topfeatures.bigram1
 innovation_customer      customer_success       problem_solvers       solvers_taking
                   2                     2                     1                    1
    taking_ownership       ownership_team           team_driven       driven_positive
                   1                     1                     1                    1
     positive_nimble         nimble_work           work_hard    passionate_commited
                   1                     1                     1                    1
     commited_honest     honest_innovators  innovators_providing       providing_best
                   1                     1                     1                    1
      best_solutions       ideas_believe      customer_creating      creating_growth
                   1                     1                     1                    1
 growth_opportunities  employees_optimizing    optimizing_globally     globally_innovator
                   1                     1                     1                    1
 innovator_courageous  unique_perspective  perspective_innovative  innovative_efficient
                   1                     1                     1                    1
        golden_rule        rule_success stakeholders_child-like        child-like_joy
                   1                     1                     1                    1
      really_really          web_works     information_crosses          suit_great
                   1                     1                     1                    1
        great_just         good_enough     dedicated_innovator     innovator_commited
                   1                     1                     1                    1
team-players_work-hard  work-hard_play-hard  play-hard_philosophy     philosophy_fight
                   1                     1                     1                    1
          win_find      existing_business      business_problems      problems_deliver
                   1                     1                     1                    1
 deliver_unparallaled  unparallaled_quality
                   1                     1
```

CORRELATION BETWEEN MOST FREQUENT WORDS OF CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #specifying a correlation limit of 0.5
> dfm.tm<-convert(dfm.stem1, to="tm")
> findAssocs(dfm.tm,
+            c("reliabl", "commit", "integr"),
+            corlimit=0.5)
$reliabl
enterprise-readi             time            scale            ration             singl           databas
            0.56             0.56             0.56             0.56              0.56              0.56
        platform       cost-effect            sacal  customer-servic
            0.56             0.56             0.56             0.56

$commit
  sustain communiti     divers
     0.83       0.83       0.55

$integr
teamwork
    0.52
```

PROCESS THE DATA FOR CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #Process the data of core values for analysis.
>
> temp1<-textProcessor(documents=orgcorpus$Core.Values, metadata = orgcorpus)
Building corpus...
Converting to Lower Case...
Removing stopwords...
Removing numbers...
Removing punctuation...
Stemming...
Creating Output...
> names(temp1)  # produces:  "documents", "vocab", "meta", "docs.removed"
[1] "documents"    "vocab"        "meta"         "docs.removed"
> meta<-temp1$meta
> vocab<-temp1$vocab
> docs<-temp1$documents
> out <- prepDocuments(docs, vocab, meta)
Removing 134 of 181 terms (134 of 264 tokens) due to frequency
Your corpus now has 30 documents, 47 terms and 130 tokens.
> docs<-out$documents
> vocab<-out$vocab
> meta <-out$meta
> #running stm for top 15 topics
> top15values <-stm(docs , vocab ,
+            K=15,
+            verbose=TRUE,
+            data=meta,
+            max.em.its=25)
```

EXPLORING TOP 15 TOPICS OF CORE VALUES

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> #Process the data of core values for analysis.
>
> temp1<-textProcessor(documents=orgcorpus$Core.Values, metadata = orgcorpus)
Building corpus...
Converting to Lower Case...
Removing stopwords...
Removing numbers...
Removing punctuation...
Stemming...
Creating Output...
> names(temp1)  # produces:  "documents", "vocab", "meta", "docs.removed"
[1] "documents"    "vocab"        "meta"         "docs.removed"
> meta<-temp1$meta
> vocab<-temp1$vocab
> docs<-temp1$documents
> out <- prepDocuments(docs, vocab, meta)
Removing 134 of 181 terms (134 of 264 tokens) due to frequency
Your corpus now has 30 documents, 47 terms and 130 tokens.
> docs<-out$documents
> vocab<-out$vocab
> meta <-out$meta
> #running stm for top 15 topics
> top15values <-stm(docs , vocab ,
+            K=15,
+            verbose=TRUE,
+            data=meta,
+            max.em.its=25)
```

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
Beginning Initialization.
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 1 (approx. per word bound = -3.997)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 2 (approx. per word bound = -3.833, relative change = 4.098e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 3 (approx. per word bound = -3.746, relative change = 2.263e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 4 (approx. per word bound = -3.693, relative change = 1.421e-02)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 5 (approx. per word bound = -3.646, relative change = 1.258e-02)
Topic 1: reliabl, valu, passion, simplic, new
 Topic 2: success, innov, teamwork, trustworthi, respons
 Topic 3: solut, challeng, result, problen, win
 Topic 4: commit, integr, employe, global, growth
 Topic 5: take, busi, win, problem, work
 Topic 6: respect, divers, simplic, integr, innov
 Topic 7: opportun, growth, work, new, maxim
 Topic 8: speed, qualiti, maxim, win, agil
 Topic 9: posit, provid, technolog, problem, passion
 Topic 10: dedic, open, ownership, respons, innov
 Topic 11: innov, passion, integr, win, open
 Topic 12: respons, fun, agil, win, well
 Topic 13: togeth, better, win, success, make
 Topic 14: custom, excel, transpar, maxim, ownership
 Topic 15: best, uniqu, fast, make, well
```

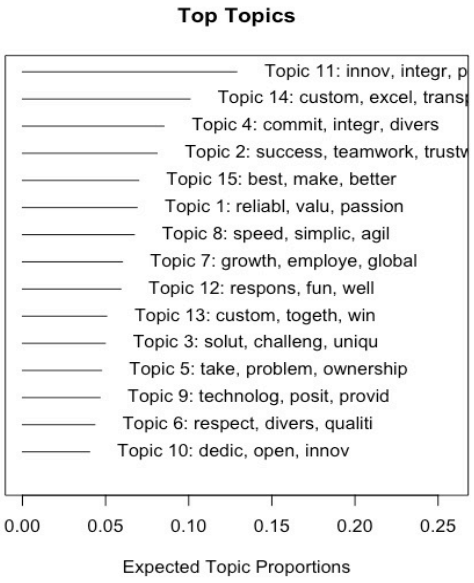
```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
Completing Iteration 20 (approx. per word bound = -3.317, relative change = 3.755e-03)
Topic 1: reliabl, valu, passion, new, simplic
 Topic 2: success, teamwork, trustworthi, innov, respons
 Topic 3: solut, challeng, result, busi, uniqu
 Topic 4: commit, integr, employe, global, divers
 Topic 5: take, problem, busi, ownership, work
 Topic 6: respect, divers, qualiti, integr, simplic
 Topic 7: growth, employe, global, new, open
 Topic 8: speed, simplic, qualiti, agil, maxim
 Topic 9: technolog, posit, provid, solut, passion
 Topic 10: dedic, open, ownership, innov, respons
 Topic 11: innov, integr, passion, custom, open
 Topic 12: respons, fun, well, agil, innov
 Topic 13: togeth, win, custom, result, integr
 Topic 14: custom, excel, transpar, maxim, ownership
 Topic 15: best, make, fast, better, uniqu
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 21 (approx. per word bound = -3.305, relative change = 3.528e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 22 (approx. per word bound = -3.295, relative change = 2.960e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 23 (approx. per word bound = -3.287, relative change = 2.429e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 24 (approx. per word bound = -3.281, relative change = 1.965e-03)
.....
Completed E-Step (0 seconds).
Completed M-Step.
Model Terminated Before Convergence Reached
```

TOP 15 TOPICS OF CORE VALUES

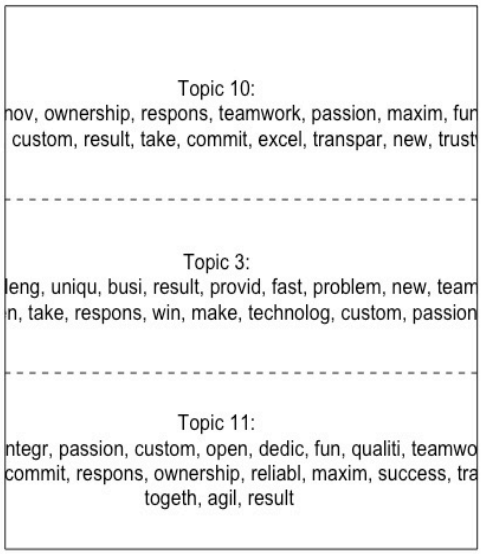
```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 1/
> topics1 <-labelTopics(top15values , topics=c(1:15))
> topics1
Topic 1 Top Words:
        Highest Prob: reliabl, valu, passion, new, simplic, qualiti, innov
        FREX: reliabl, valu, passion, new, qualiti, open, simplic
        Lift: reliabl, valu, passion, new, simplic, qualiti, speed
        Score: reliabl, valu, passion, new, simplic, speed, qualiti
Topic 2 Top Words:
        Highest Prob: success, teamwork, trustworthi, innov, respons, dedic, maxim
        FREX: success, teamwork, trustworthi, maxim, innov, well, agil
        Lift: success, trustworthi, teamwork, innov, respons, dedic, maxim
        Score: success, trustworthi, teamwork, innov, respons, dedic, maxim
Topic 3 Top Words:
        Highest Prob: solut, challeng, uniqu, busi, result, provid, fast
        FREX: challeng, solut, provid, uniqu, busi, result, new
        Lift: challeng, solut, uniqu, busi, result, provid, fast
        Score: solut, challeng, busi, uniqu, result, provid, fast
Topic 4 Top Words:
        Highest Prob: commit, integr, divers, innov, custom, dedic, maxim
        FREX: commit, integr, maxim, new, custom, qualiti, innov
        Lift: commit, integr, divers, innov, dedic, custom, maxim
        Score: commit, integr, divers, dedic, innov, custom, maxim
Topic 5 Top Words:
        Highest Prob: take, problem, ownership, busi, work, new, qualiti
        FREX: take, problem, ownership, busi, work, new, uniqu
        Lift: take, problem, ownership, busi, work, new, qualiti
        Score: take, problem, work, busi, ownership, new, posit
Topic 6 Top Words:
        Highest Prob: respect, divers, qualiti, integr, teamwork, innov, trustworthi
        FREX: respect, divers, qualiti, integr, maxim, teamwork, custom
        Lift: respect, divers, qualiti, integr, teamwork, innov, trustworthi
        Score: divers, respect, qualiti, integr, teamwork, innov, trustworthi
Topic 7 Top Words:
        Highest Prob: growth, employe, global, open, new, opportun, innov
        FREX: growth, employe, global, open, new, opportun, maxim
        Lift: growth, employe, global, new, opportun, open, innov
        Score: growth, employe, global, opportun, new, open, innov
Topic 8 Top Words:
        Highest Prob: speed, simplic, agil, maxim, opportun, qualiti, valu
        FREX: speed, simplic, agil, qualiti, maxim, opportun, new
        Lift: speed, simplic, agil, maxim, opportun, qualiti, valu
        Score: speed, simplic, agil, opportun, maxim, qualiti, valu
Topic 9 Top Words:
        Highest Prob: technolog, posit, provid, solut, new, passion, qualiti
        FREX: technolog, posit, provid, new, uniqu, busi, work
        Lift: technolog, posit, provid, solut, new, passion, qualiti
        Score: posit, technolog, provid, solut, new, passion, problem
Topic 10 Top Words:
        Highest Prob: dedic, open, innov, ownership, respons, teamwork, passion
        FREX: dedic, open, maxim, ownership, result, innov, integr
        Lift: dedic, open, ownership, innov, respons, teamwork, passion
        Score: dedic, open, ownership, innov, respons, teamwork, passion
Topic 11 Top Words:
        Highest Prob: innov, integr, passion, custom, open, dedic, fun
        FREX: innov, open, qualiti, integr, maxim, ownership, custom
        Lift: innov, integr, passion, custom, open, dedic, fun
        Score: innov, integr, passion, dedic, fun, open, trustworthi
Topic 12 Top Words:
        Highest Prob: respons, fun, well, agil, innov, trustworthi, open
        FREX: fun, respons, agil, well, open, maxim, innov
        Lift: fun, respons, well, agil, trustworthi, innov, open
        Score: respons, fun, well, agil, innov, trustworthi, open
Topic 13 Top Words:
        Highest Prob: custom, togeth, win, result, integr, new, better
        FREX: togeth, custom, win, integr, result, new, busi
        Lift: togeth, win, result, custom, integr, new, better
        Score: togeth, win, result, custom, integr, new, better
Topic 14 Top Words:
        Highest Prob: custom, excel, transpar, ownership, maxim, integr, innov
        FREX: excel, custom, transpar, ownership, maxim, well, integr
        Lift: excel, transpar, custom, ownership, maxim, integr, innov
        Score: custom, excel, transpar, ownership, maxim, integr, innov
Topic 15 Top Words:
        Highest Prob: best, make, better, fast, work, well, uniqu
        FREX: best, make, better, fast, well, work, uniqu
        Lift: best, make, better, fast, work, well, uniqu
        Score: better, best, make, fast, work, well, uniqu
```

PLOT TOP 15 TOPICS OF CORE VALUES

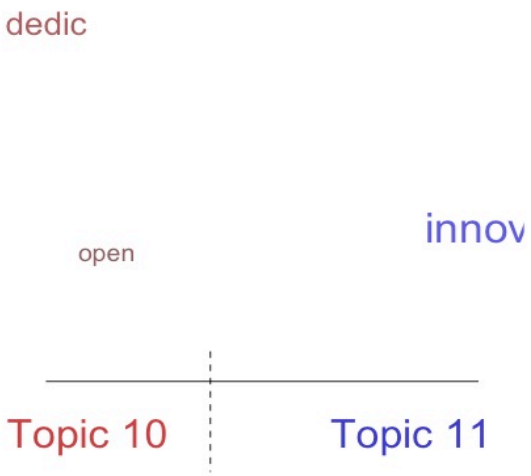
```
> plot.STM(top15values, type="summary")
```




```
> plot.STM(top15values, type="labels", topics=c(10,3,11))
```

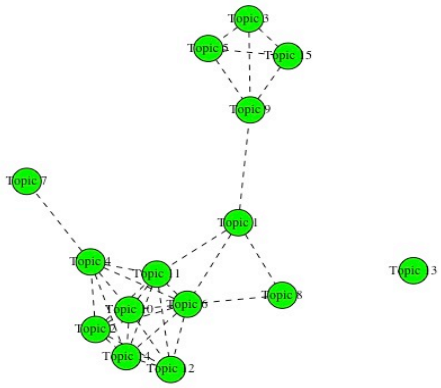


```
> plot.STM(top15values, type="perspectives", topics = c(10,11))
```



CORRELATION PLOT FOR TOP 15 TOPICS

```
> mod.out.corr <- topicCorr(top15values)
> plot.topicCorr(mod.out.corr)
```



INSIGHT FOR HOW TO STRUCTURE A FIRM FOR DATA-ANALYSIS READINESS

From the document frequency matrix analysis of mission statements and core values, it can be seen that most of the data-driven companies focus on helping customers in making appropriate data related decisions. They rely on their core capabilities, which are innovation, commitment, integrity and reliability. So, these are the few characteristics that are should be an essential part of a data-driven company.

RECOMMENDED DATA DRIVEN APPROACH TO THE CEO

In my opinion, the performance of the company largely depends on the way it uses or perceives data. How well data is understood, plays an important aspect in providing any data related solution to the people and can help in changing the world.

PART 2

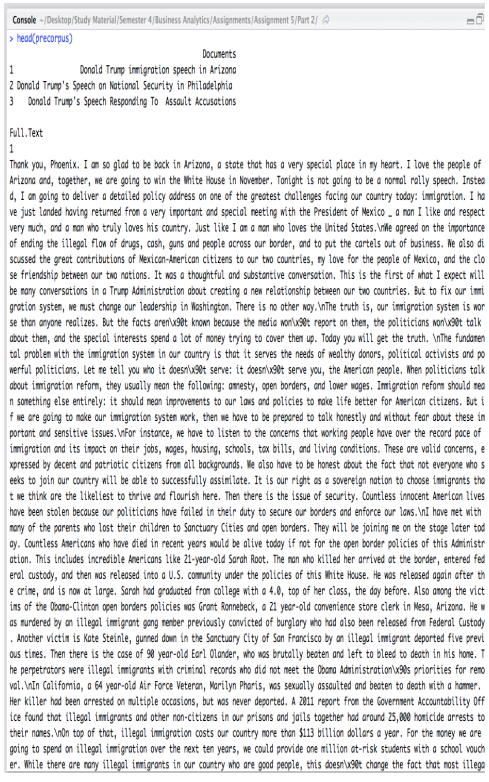
Case: This video discusses Donald Trump's linguistic style. Sometimes when he speaks he seems erratic and unfocused, yet many people like him and connect with him. Assume you are providing "intelligence" to Trump's campaign manager on what is making this candidate effective with people

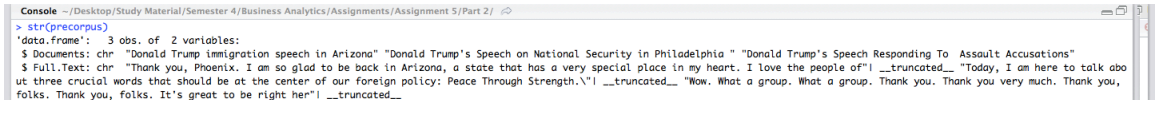
Steps: Search for "Donald Trump speech transcript" and select 3 speeches of your choice; Create a corpus for the speeches; Complete a frequency analysis of word usage; Complete a sentiment analysis; What are the common topics in the corpus; Write a memo style report summarizing Trump's linguistic effectiveness.

The three speeches that I considered are:

- *Donald Trump Immigration Speech in Arizona*
- *Donald Trump's Speech on National Security on Philadelphia*
- *Donald Trump's Speech Responding to Assault Accusations*

LOADING PACKAGES AND DONALDTRUMPSPEECH.CSV





CREATE CORPUS OF DONALD TRUMP SPEECH

```
> # Creating a corpus for speech
>
> require(quanteda)
>
> speechcorpus<- corpus(precorpus$Full.Text,
>                       docnames=transcriptcorpus$Documents)
> #explore the corpus of speech
>
> names(speechcorpus)
[1] "documents" "metadata"  "settings"  "tokens"
> summary(speechcorpus)
Corpus consisting of 3 documents.

  Text Types Tokens Sentences
 text1  1256   4611       197
 text2   959   2757       138
 text3  1186   5489       321

Source:  /Users/anushiarora/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/* on x86_64 by anushiarora
Created: Fri Nov 18 14:26:26 2016
Notes:

> head(speechcorpus)

text1

"Thank you, Phoenix. I am so glad to be back in Arizona, a state that has a very special place in my heart. I love the people of Arizona and, together, we are going to win the White House in No
vember. Tonight is not going to be a normal rally speech. Instead, I am going to deliver a detailed policy address on one of the greatest challenges facing our country today: immigration. I hav
e just landed having returned from a very important and special meeting with the President of Mexico _ a man I like and respect very much, and a man who truly loves his country. Just like I am
a man who loves the United States.\nWe agreed on the importance of ending the illegal flow of drugs, cash, guns and people across our border, and to put the cartels out of business. We also dis
cussed the great contributions of Mexican-American citizens to our two countries, my love for the people of Mexico, and the close friendship between our two nations. It was a thoughtful and sub
stantive conversation. This is the first of what I expect will be many conversations in a Trump Administration about creating a new relationship between our two countries. But to fix our immigr
ation system, we must change our leadership in Washington. There is no other way.\nThe truth is, our immigration system is worse than anyone realizes. But the facts aren't known because the med
ia won't report on them, the politicians won't talk about them, and the special interests spend a lot of money trying to cover them up. Today you will get the truth. \nThe fundamental problem
with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful politicians. Let me tell you who it doesn't serve: it doesn't serve y
ou, the American people. When politicians talk about immigration reform, they usually mean the following: amnesty, open borders, and lower wages. Immigration reform should mean something else e
ntirely: it should mean improvements to our laws and policies to make life better for American citizens. But if we are going to make our immigration system work, then we have to be prepared to
talk honestly and without fear about these important and sensitive issues.\nFor instance, we have to listen to the concerns that working people have over the record pace of immigration and its
impact on their jobs, wages, housing, schools, tax bills, and living conditions. These are valid concerns, expressed by decent and patriotic citizens from all backgrounds. We also have to be ho
nest about the fact that not everyone who seeks to join our country will be able to successfully assimilate. It is our right as a sovereign nation to choose immigrants that we think are the lik
eliest to thrive and flourish here. Then there is the issue of security. Countless innocent American lives have been stolen because our politicians have failed in their duty to secure our borde
```

CREATE A DOCUMENT FREQUENCY MATRIX FOR DONALD TRUMP SPEECH

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> #Generate DFM
> corpus<- toLower(speechcorpus, keepAcronyms = FALSE)
> cleancorpus <- tokenize(corpus,
+                         removeNumbers=TRUE,
+                         removePunct = TRUE,
+                         removeSeparators=TRUE,
+                         removeTwitter=FALSE,
+                         verbose=TRUE)
Starting tokenization...
  ...preserving Twitter characters (#, @)...total elapsed: 0 seconds.
  ...tokenizing texts...total elapsed:  0.00600000000008549 seconds.
  ...replacing Twitter characters (#, @)...total elapsed: 0.000999999999976353 seconds.
  ...replacing names...total elapsed:  0 seconds.
Finished tokenizing and cleaning 3 texts.
> stop_words <- c("re", "net", "six", "room", "g", "gut", "oliv", "tripi","physic", "craft", "fair", "second",
+                 "may", "touch", "don", "voucher", "draw", "aren", "oh", "hello", "lo", "gotten", "glass","whose",
+                 "__they'v", "__so", "__it", "__for", "per", "novemb", "averag", "chao", "materi", "tool", "seven",
+                 "vet", "howev", "without", "lot", "wit", "line", "nov", "didn", "set", "abl", "would'v", "__we",
+                 "one", "year", "s", "t", "know", "also", "just", "like", "can", "need", "number", "say", "includ",
+                 "new", "go","now", "look", "back", "take", "thing", "even", "ask", "seen", "said", "put", "day",
+                 "anoth", "come", "use", "total", "happen", "place", "thank", "ve", "get", "much")
> stop_words <- tolower(stop_words)
> dfm<- dfm(cleancorpus, toLower = TRUE,
+                ignoredFeatures = c(stop_words, stopwords("english")),
+                verbose=TRUE,
+                stem=TRUE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 3 documents
   ... indexing features: 2,188 feature types
   ... removed 200 features, from 257 supplied (glob) feature types
   ... stemming features (English), trimmed 414 feature variants
   ... created a 3 x 1574 sparse dfm
   ... complete.
Elapsed time: 0.078 seconds.
```

TOP 100 FREQUENT TERMS IN DONALD TRUMP SPEECH

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> # Reviewing top features
>
> topfeatures(dfm, 100)
     will    immigr   countri     peopl   clinton  american     illeg     state   hillari     stori       law     great
      170        68        64        62        54        41        39        35        32        32        31        29
       go      year      time    border    nation      mani     never       job    system    crimin    polici     offic
       28        28        28        27        25        25        25        23        22        22        21        21
   defens      work    enforc administr     media       let      make     secur    includ   america      unit     power
       21        20        20        19        19        19        19        19        19        19        18        18
   polit       want    believ establish   corrupt      love     obama      visa     world      need      citi    govern
       17        17        17        17        17        16        16        16        16        15        15        15
   today      talk  interest     feder   control  militari   special     anoth   billion   million    worker      plan
       14        14        14        14        14        14        13        13        13        13        13        13
   elect       lie    presid      issu    releas     remov      part     futur        us   destroy       end     first
       13        13        12        12        12        12        12        12        12        12        11        11
  amnesti      live     refus   protect      last      stop    togeth       two     spend      life     crime    deport
       11        11        11        11        11        11        10        10        10        10        10        10
   allow  congress     radic      base  campaign      york    import     trump      must      fact    reform      mean
       10        10        10        10        10        10         9         9         9         9         9         9
      u.      educ      ever      noth
       9         9         9         9
```

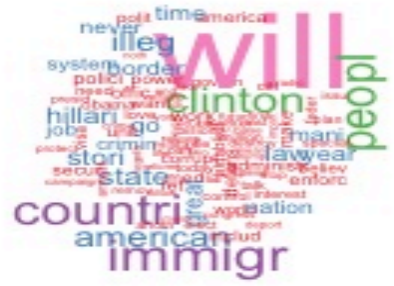
TRIGRAM FOR DONALD TRUMP SPEECH

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> #dfm with trigrams
>
> cleancorpus1 <- tokenize(corpus,
+                          removeNumbers=TRUE,
+                          removePunct = TRUE,
+                          removeSeparators=TRUE,
+                          removeTwitter=FALSE,
+                          ngrams=3, verbose=TRUE)
Starting tokenization...
  ...preserving Twitter characters (#, @)...total elapsed: 0 seconds.
  ...tokenizing texts...total elapsed:  0.00599999999985812 seconds.
  ...replacing Twitter characters (#, @)...total elapsed: 0.00100000000020373 seconds.
  ...creating ngrams...total elapsed: 0.307000000000016 seconds.
  ...replacing names...total elapsed:  0.000999999999976353 seconds.
Finished tokenizing and cleaning 3 texts.
>
> dfm.trigram<- dfm(cleancorpus1, toLower = TRUE,
+                   ignoredFeatures = c(stop_words, stopwords("english")),
+                   verbose=TRUE,
+                   stem=FALSE)
Creating a dfm from a tokenizedTexts object ...
   ... lowercasing
   ... indexing documents: 3 documents
   ... indexing features: 10,291 feature types
   ... removed 9,586 features, from 257 supplied (glob) feature types
   ... created a 3 x 705 sparse dfm
   ... complete.
Elapsed time: 0.209 seconds.
```

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> topfeatures.trigram<-topfeatures(dfm.trigram, n=50)
> topfeatures.trigram
              will_make_america                   make_america_great          radical_islamic_terrorism
                              5                                    4                                  4
    million_illegal_immigrants               law_enforcement_officers corrupt_political_establishment
                              3                                    3                                  3
       crooked_hillary_clinton                  hillary_clinton_meets          criminal_illegal_immigrants
                              3                                    2                                  2
        border_patrol_officers                biometric_entry-exit_visa               entry-exit_visa_tracking
                              2                                    2                                  2
         visa_tracking_system                 legal_immigration_system                defeating_radical_islamic
                              2                                    2                                  2
      ballistic_missile_defense            missile_defense_capability            offensive_cyber_capabilities
                              2                                    2                                  2
     global_special_interests                    will_determine_whether                made_inappropriate_advances
                              2                                    2                                  2
           normal_rally_speech                     rally_speech_instead                detailed_policy_address
                              1                                    1                                  1
   greatest_challenges_facing               country_today_immigration                        drugs_cash_guns
                              1                                    1                                  1
       special_interests_spend                 wealthy_donors_political               donors_political_activists
                              1                                    1                                  1
      powerful_politicians_let                  following_amnesty_open                    amnesty_open_borders
                              1                                    1                                  1
       lower_wages_immigration                 wages_immigration_reform                     mean_something_else
                              1                                    1                                  1
       something_else_entirely                        make_life_better                 immigration_system_work
                              1                                    1                                  1
             jobs_wages_housing                  wages_housing_schools                    housing_schools_tax
                              1                                    1                                  1
              schools_tax_bills                  valid_concerns_expressed              security_countless_innocent
                              1                                    1                                  1
    countless_innocent_american                   innocent_american_lives                     stage_later_today
                              1                                    1                                  1
         later_today_countless                  today_countless_americans
                              1                                    1
```

WORDCLOUD FOR DONALD TRUMP SPEECH

```
Console ~/Desktop/Study Material/Semester 4/Business Analyt
> # Wordcloud for Speech
>
> library(wordcloud)
> set.seed(142)   #keeps cloud' shape fixed
> dark2 <- brewer.pal(8, "Set1")
> freq<-topfeatures(dfm, n=100)
>
>
> wordcloud(names(freq),
+           freq, max.words=200,
+           scale=c(3, .1),
+           colors=brewer.pal(8, "Set1"))
> #Sentiment Analysis
>
```



SENTIMENT ANALYSIS OF THE SPEECH

```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> #Sentiment Analysis
>
>
> mydict <- dictionary(list(positive = c("win", "love", "respect", "prestige", "power", "protect", "struggle",
+                                        "enrich", "good", "survival", "morally", "movement", "strongly",
+                                        "heaven", "highly-successful", "bright", "hope", "fix", "happy",
+                                        "thrilled", "safety", "prosperity", "peace", "rise", "glorious", "great", "stable"),
+                           negative = c("failed", "corrupt", "lie", "threat", "disastrous", "illegal", "dry",
+                                        "robbed", "raided", "locked", "crooked", " illusion", "rigged", "deformed",
+                                        "attack", "destroy", "destruction", "slander", "concerted", "vicious",
+                                        "exposing", "corruption", "misrepresented", "horrible", "ISIS", "deport",
+                                        "incompetent", "worse", "sickness", "immorality", "guilty", "warned", "debt",
+                                        "terrorist", "crime", "crushed", "poverty", "war", "conflict", "destroy", "defeat")))
> dfm.sentiment <- dfm(speechcorpus, dictionary = mydict)
Creating a dfm from a corpus ...
   ... lowercasing
   ... tokenizing
   ... indexing documents: 3 documents
   ... indexing features: 2,188 feature types
   ... applying a dictionary consisting of 2 keys
   ... created a 3 x 2 sparse dfm
   ... complete.
Elapsed time: 0.057 seconds.
> topfeatures(dfm.sentiment)
negative positive
     144      128
```

```
View(dfm.sentiment)
```

	positive	negative
text1	26	39
text2	14	23
text3	88	82

EXPLORING TOPICS OF DONALD TRUMP SPEECH

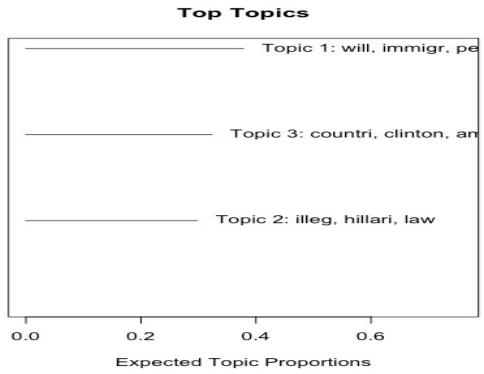
```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/Part 2/
> #running topics
> temp<-textProcessor(documents=precorpus$Full.Text, metadata = precorpus)
Building corpus...
Converting to Lower Case...
Removing stopwords...
Removing numbers...
Removing punctuation...
Stemming...
Creating Output...
> names(temp)  # produces:  "documents", "vocab", "meta", "docs.removed"
[1] "documents"    "vocab"        "meta"         "docs.removed"
> meta<-temp$meta
> vocab<-temp$vocab
> docs<-temp$documents
> out <- prepDocuments(docs, vocab, meta)
Removing 1111 of 1608 terms (1111 of 2295 tokens) due to frequency
Your corpus now has 3 documents, 497 terms and 1184 tokens.
> docs<-out$documents
> vocab<-out$vocab
> meta <-out$meta
> prevfit <-stm(docs , vocab ,
+               K=3,
+               verbose=TRUE,
+               data=meta,
+               max.em.its=25)
Beginning Initialization.
...
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 1 (approx. per word bound = -5.744)
...
Completed E-Step (0 seconds).
Completed M-Step.
Completing Iteration 2 (approx. per word bound = -5.744, relative change = 1.173e-05)
...
Completed E-Step (0 seconds).
Completed M-Step.
Model Converged
>
```

THREE TOPICS OF DONALD TRUMP SPEECH

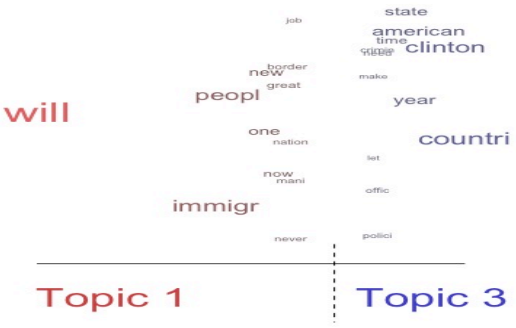
```
Console ~/Desktop/Study Material/Semester 4/Business Analytics/Assignments/Assignment 5/
> topics <-labelTopics(prevfit , topics=c(1:3))
> topics
Topic 1 Top Words:
         Highest Prob: will, immigr, peopl, one, new, now, great
         FREX: will, immigr, peopl, one, new, now, great
         Lift: will, peopl, immigr, one, new, great, now
         Score: will, immigr, peopl, one, new, now, great
Topic 2 Top Words:
         Highest Prob: illeg, hillari, law, just, like, system, enforc
         FREX: illeg, hillari, law, just, system, enforc, like
         Lift: illeg, law, hillari, system, enforc, just, number
         Score: illeg, hillari, law, just, like, system, enforc
Topic 3 Top Words:
         Highest Prob: countri, clinton, american, year, state, time, need
         FREX: countri, clinton, american, year, state, time, need
         Lift: countri, clinton, american, year, state, time, need
         Score: countri, clinton, american, year, state, time, need
```

PLOT OF THE TOPICS OF DONALD TRUMP SPEECH

```
> plot.STM(prevfit, type="summary")
```



```
> plot.STM(prevfit, type="perspectives", topics = c(1,3))
```

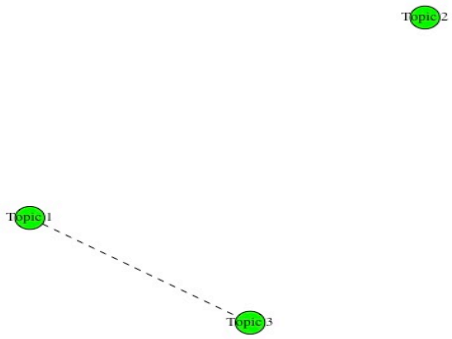



```
> plot.STM(prevfit, type="perspectives", topics = c(1,2))
```

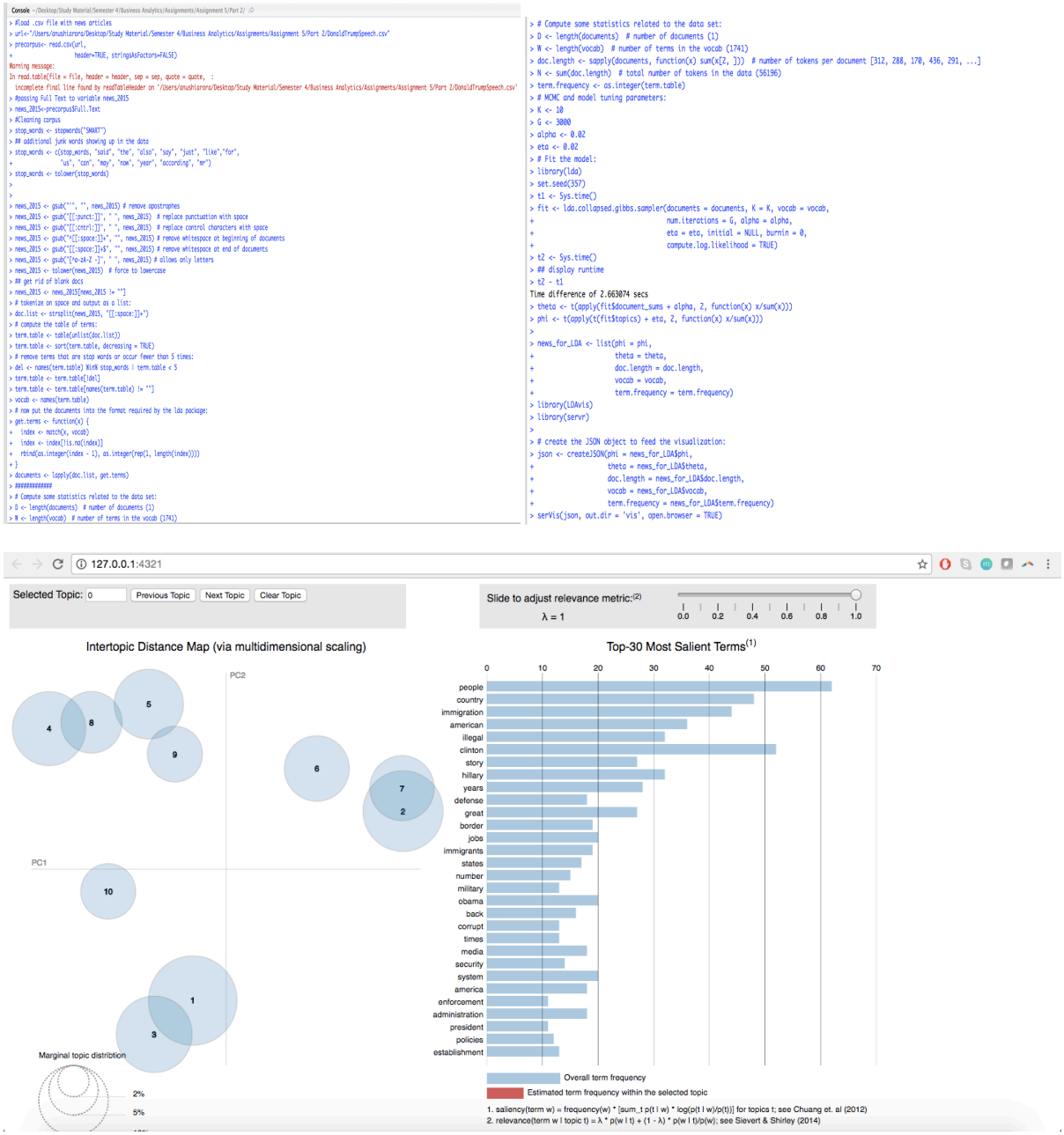

```
> plot.STM(prevfit, type="perspectives", topics = c(2,3))
```

CORRELATION PLOT OF TOPICS FOR DONALD TRUMP SPEECH

```
> mod.out.corr <- topicCorr(prevfit)  #Estimates a graph of topic correlations
> plot.topicCorr(mod.out.corr)
```



TOPIC MODELLING OF DONALD TRUMP SPEECH



GITHUB LINK: https://github.com/Anushi0701/Assignment-5

COMMON TOPICS:

The topics being discussed in all the three speeches were related to the following topics: immigration, war, terrorist, Hillary Clinton, American, Country and Security. There were many other things discussed, but these terms were frequently used by Donald trump in his three different speeches.

FINDING MEMO

Memo to: Donald Trump's Campaign Manager

From: Anushi Arora

Date: November 18th, 2016

Subject: Donald Trump's Linguistic Style

This memo focuses on How Donald trump uses words in his speech, including a sentiment analysis of his speech and deep interpretations from the analysis of his three speeches.

For analyzing the content of his speech, I performed the frequency analysis, sentiment analysis and topic modeling on his speech. The words like "will", "country", "Clinton", "immigration", "American" and "illegal" are frequently being used in his speech. The sentiment analysis shows that Trump focuses on negative word more than the positive words. There is a correlation between the positive and negative words being used in the three speeches.

Trump linguistic style definitely has helped him in attracting a lot of audience. He does not complete the sentences and try referring to each problem with very powerful examples; such as the one he mentioned in his immigration speech – "This includes incredible Americans like 21-year-old Sarah Root. The man who killed her arrived at the border, entered federal custody, and then was released into a U.S. community under the policies of this White House. He was released again after the crime, and is now at large. " Trump focuses on making people remember about the pain caused by the terrorist activities in the country and blame his opponents for those activities. Trump tries to convince people that he can make the country safe again and will always give priority to Americans.

In all the three speeches, Trump have used various powerful words and tried expressing high negative sentiments in his thoughts. Due to his repetitive use of strong words and his focus on topics such as National Security, Immigration, Jobs etc. have helped him in attracting interest of lot of American population.

References:

Donald Trump Immigration Speech in Arizona
http://www.politico.com/story/2016/08/donald-trump-immigration-address-transcript-227614

Donald Trump's Speech on National Security on Philadelphia
http://thehill.com/blogs/pundits-blog/campaign/294817-transcript-of-donald-trumps-speech-on-national-security-in

Donald Trump's Speech Responding to Assault Accusations
http://www.npr.org/2016/10/13/497857068/transcript-donald-trumps-speech-responding-to-assault-accusations