Homework Part 1

Q1 Create a corpus of their mission statements

1. Create corpus

```
newscorpus1 <- corpus(Mission$Mission_Statement,
                      docnames=Mission$Company_Name,
                      docvar=data.frame(Categries=Mission$Categories))
```

explore the corpus:

	Text	Types	Tokens	Sentences	Categries
InsightSquared		27	31	1	Big Data
Trifacta		24	27	1	Big Data
Cloudera		18	24	1	Big Data
Sumo Logic		38	48	1	Big Data
Google		25	28	2	Big Data
Visier		27	29	1	Big Data
Domo		24	27	1	Big Data
Tableau Software		24	34	3	Big Data
Hortonworks		11	12	1	Big Data
Informatica		18	18	1	Big Data
Talend		31	37	1	Big Data
Platfora		17	19	1	Big Data
SAP		35	37	2	Big Data
SAS institute		11	11	1	Big Data
Microsoft		15	17	1	Big Data
EMC		24	35	1	Big Data
Palantir Technologies		20	22	1	Big Data
GoodData		12	12	1	Big Data
MongoDB		12	12	1	Big Data
Predixion		10	10	1	Big Data
Qlik		25	27	1	Big Data
Salesforce		29	35	1	Big Data
DataStaks		22	23	1	Big Data
Neo Technology		24	26	1	Big Data
Teradata		20	21	1	Big Data
Dell		33	38	2	Big Data
1010data		26	29	1	Big Data
Hewlett-packard		19	21	2	Big Data
Alteryx		10	10	1	Big Data
Information Builders		19	21	1	Big Data

2. Tokenizing the coupus

```
newscorpus1<- toLower(newscorpus1, keepAcronyms = FALSE)
cleancorpus1 <- tokenize(newscorpus1, removeNumbers=TRUE, removePunct =
TRUE, removeSeparators=TRUE, removeTwitter=FALSE, verbose=TRUE)
```

Explore the clean corpus:

```
> summary(cleancorpus1)
                      Length Class  Mode
InsightSquared        26     -none- character
Trifacta              23     -none- character
Cloudera              18     -none- character
Sumo Logic            36     -none- character
Google                26     -none- character
Visier                26     -none- character
Domo                  26     -none- character
Tableau Software      28     -none- character
Hortonworks            9     -none- character
Informatica           16     -none- character
Talend                27     -none- character
Platfora              14     -none- character
SAP                   32     -none- character
SAS institute         10     -none- character
Microsoft             16     -none- character
EMC                   22     -none- character
Palantir Technologies 17     -none- character
GoodData              11     -none- character
MongoDB               11     -none- character
Predixion              9     -none- character
Qlik                  21     -none- character
Salesforce            28     -none- character
DataStaks             22     -none- character
Neo Technology        21     -none- character
Teradata              20     -none- character
Dell                  34     -none- character
1010data              25     -none- character
Hewlett-packard       19     -none- character
Alteryx                9     -none- character
Information Builders  19     -none- character
```

3. Cleaning stop words, stemming and creating Document Feature Matrix

```
dfm.simple1 <- dfm(cleancorpus1,
                   toLower = TRUE,
                   ignoredFeatures =stopwords("english"),
                   verbose=TRUE,
                   stem=TRUE)
topfeatures1 <-topfeatures(dfm.simple1, n=50)
view(dfm.simple1)
```

	insightsquar	found	mission	chang	way	small	mid-siz	busi	run	provid	visual	action	afford	insight	trifacta
InsightSquared	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Trifacta	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
Cloudera	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Sumo Logic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Google	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Visier	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
Domo	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0
Tableau Software	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Hortonworks	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Informatica	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
Talend	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
Platfora	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
SAP	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0

```
topfeatures1
```

```
> topfeatures1
   mission     custom       data       busi      deliv          s   intellig      peopl        way     provid  transform
        17         15         12         11          7          6          6          6          5          5          5
   compani      world       help      innov    insight      oper     servic       make     empow     better  technolog
         5          5          5          5          4          4          4          4          4          4          4
     solut      chang     analyt  enterpris        can      everi       work   qualiti      found       enabl      allow
         4          3          3          3          3          3          3          3          2          2          2
     today    million      power      organ     inform        see understand      decis      manag      drive       valu
         2          2          2          2          2          2          2          2          2          2          2
       one    connect        new     market     improv    perform
         2          2          2          2          2          2
```

Frequency Analysis of DFM:

According to the DFM, the root words show that the 30 leading data-driven companies are focusing on: “data”, “customer”, ”business”, intelligence”, “insights”, “value”, “decision”, ”transform and change”, “better technologic solution”, “innovation and new”. From these key points, we can conclude that theses companies are pursuing to transform the traditional decision-making approach to an innovative data-driven method to find insights and deliver more values to businesses and individuals.

4. Exploration in context

```
kwic(cleancorpus1, "data", 2)
kwic(cleancorpus1, "analytics", window = 3)
```

```
                              contextPre keyword           contextPost
          [Trifacta, 16]    transform big [    data ] from a
          [Cloudera, 12]        all their [    data ] where all
         [Sumo Logic, 11]     petabytes of [    data ] more than
         [Sumo Logic, 31] powerful machine [    data ] analytics services
    [Tableau Software, 9]   and understand [    data ] is one
   [Tableau Software, 27]        mantle of [    data ]
         [Hortonworks, 9]          world s [    data ]
             [MongoDB, 8]     software and [    data ] for innovators
            [Teradata, 3]         the best [    data ] warehouse foundation
           [1010data, 14]         of their [    data ] whether they
           [1010data, 19]            are a [    data ] scientist or
[Information Builders, 7]   deliver better [    data ] and better
> kwic(cleancorpus1, "analytics", window = 3)
                                  contextPre   keyword       contextPost
         [Sumo Logic, 32] powerful machine data [ analytics ] services in the
[Information Builders, 10]       data and better [ analytics ] to more people
```

Analysis of "Data Analytics" in context:

In the context, we can see that the word "data" shows up in context which is in relates to "big", "powerful", "analytics", "better", "scientist" and "services". The word "Analytics" shows up in the context that relates to "data", "machine", "services", "better" and "people". The conclusion is similar to which is generated from the DFM: data-driven companies need data scientists to make data analytics a powerful tool to help decision makers and deliver better services.

5. Explore bigrams

```
cleancorpus1 <- tokenize(newscorpus1,
                         removeNumbers=TRUE,
                         removePunct = TRUE,
                         removeSeparators=TRUE,
                         removeTwitter=FALSE,
                         ngrams=2, verbose=TRUE)

dfm.bigram1<- dfm(cleancorpus1, toLower = TRUE,
                  ignoredFeatures = c(swlist1, stopwords("english")),
                  verbose=TRUE,
                  stem=FALSE)
topfeatures.bigram1<-topfeatures(dfm.bigram1, n=50)
topfeatures.bigram1
```

```
business_intelligence                 s_mission           highest_quality      mid-sized_businesses
                    3                         2                         2                         1
    visual_actionable       affordable_insights                trifacta_s          quickly_transform
                    1                         1                         1                         1
       transform_big                  big_data          strategic_asset        cloudera's_mission
                    1                         1                         1                         1
      allow_companies        operative_word purpose-built_cloud-native       cloud-native_service
                    1                         1                         1                         1
     service_analyzes         million_searches              delivers_10s          insights_daily
                    1                         1                         1                         1
    daily_positioning         positioning_sumo               sumo_among         powerful_machine
                    1                         1                         1                         1
        machine_data           data_analytics        analytics_services         google's_mission
                    1                         1                         1                         1
  world's_information    universally_accessible         relatively_young               young_life
                    1                         1                         1                         1
         name_visier          empower_leaders                better_see          see_understand
                    1                         1                         1                         1
      make_decisions          domo_transforms           way_executives        executives_manage
                    1                         1                         1                         1
        drives_value      traditional_business      intelligence_systems          believe_helping
                    1                         1                         1                         1
      helping_people          understand_data         important_missions             21st_century
                    1                         1                         1                         1
         proudly_wear                 data_geek
                    1                         1
```

Analysis of bigram:

From the bigram, we can see how single word relates to each other. For example, "mid-sized businesses" tell us that companies focus on different sizes of businesses and services.

6. Correlation analysis

```
dfm.tm1<-convert(dfm.stem1, to="tm")

findAssocs(dfm.tm1,
           c("data", "analytics", "big"),
           corlimit=0.5)
findAssocs(dfm.tm1,
           c("business","predictive", "data" ),
           corlimit=0.5)
```

```
$analytics
purpose-built  cloud-native      analyzes     petabytes       million      searches          10s      millions         daily
         0.69          0.69          0.69          0.69          0.69          0.69          0.69          0.69          0.69
  positioning          sumo          among       powerful       machine      services        better
         0.69          0.69          0.69          0.69          0.69          0.69          0.65

$big
 trifacta   quickly    burden strategic     asset    enable    change
     1.00      1.00      1.00      1.00      1.00      0.69      0.56

$analytics
purpose-built  cloud-native      analyzes     petabytes       million      searches          10s      millions         daily
         0.69          0.69          0.69          0.69          0.69          0.69          0.69          0.69          0.69
  positioning          sumo          among       powerful       machine      services        better
         0.69          0.69          0.69          0.69          0.69          0.69          0.65

$big
 trifacta   quickly    burden strategic     asset    enable    change
     1.00      1.00      1.00      1.00      1.00      0.69      0.56
```

Analysis:

The figures which larger than 0.8 indicates that the two words are highly correlated, the figures between 0.5 and 0.8 indicates that the two words are moderate correlated.

Q2 Create a corpus of their core values

1. Creating corpus and tokenizing

```
newscorpus_V <- corpus(Value$Core.Values,
                       docnames=Value$Company_Name,
                       docvar=NULL)
newscorpus_V<- toLower(newscorpus_V, keepAcronyms = FALSE)
cleancorpus_V <- tokenize(newscorpus_V,
                          removeNumbers=TRUE,
                          removePunct = TRUE,
                          removeSeparators=TRUE,
                          removeTwitter=FALSE,
                          verbose=TRUE)
```

Explore the corpus:

	Text	Types	Tokens	Sentences
1	InsightSquared	26	32	1
2	Trifacta	19	24	1
3	Cloudera	18	24	1
4	Sumo Logic	7	10	1
5	Google	154	291	1
6	Domo	24	31	1
7	Tableau Software	25	34	1
8	Hortonworks	38	50	1
9	Informatica	16	18	1
10	Talend	31	37	1
11	Platfora	21	25	1
12	SAP	20	23	1
13	Microsoft	12	17	1
14	Palantir Technologies	28	30	1
15	MongoDB	19	22	1
16	Qlik	13	16	1
17	Salesforce	14	20	1
18	Neo Technology	27	32	1
19	Teradata	9	15	1
20	Dell	8	11	1
21	1010data	10	11	1
22	Hewlett-packard	19	27	1
23	Information Builders	19	20	1
24	Ayasdi	49	60	1
25	Actifio	42	56	1
26	Alpine data labs	25	25	1
27	Dataguise	46	69	1
28	Oracle	14	22	1
29	Splunk	6	9	1
30	EMC	76	110	1

2. Cleaning stop words, stemming and creating Document Feature Matrix

```
dfm_V<- dfm(cleancorpus_V,
            toLower = TRUE,
            ignoredFeatures =stopwords("english"),
            verbose=TRUE,
            stem=TRUE)
# Reviewing top features
view(dfm_V)
```

	proving	points	data	free	lunch	seriously	every	day	high	growth	software	companies	brainy	alumni
InsightSquared	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Trifacta	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Cloudera	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Sumo Logic	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Google	0	0	1	0	0	0	2	1	0	1	0	0	0	0
Domo	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Tableau Software	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hortonworks	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Informatica	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Talend	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Platfora	0	0	1	0	0	0	0	0	0	0	0	0	0	0

```
topfeatures(dfm_V, 50)
```

```
> topfeatures(dfm_V, 60)
     data     peopl    custom      work   respect   success     innov      busi       big     great      take     integr
       15        13        13        10         9         8         7         6         6         6         6         6
 challeng      open      valu      make     thing communiti   passion    result     everi    compani    provid     world
        5         5         5         5         5         5         5         5         4         4         4         4
      can     googl    commit     trust     think   respons  teamwork    better     human   problem    expect     creat
        4         4         4         4         4         4         4         3         3         3         3         3
  environ technolog     build   product       fun    inform     right   honesti    servic      care   support        _s
        3         3         3         3         3         3         3         3         3         3         3         3
  insight   account   qualiti     power    believ    cultur   serious       day    growth   softwar    market     decis
        3         3         3         3         3         3         2         2         2         2         2         2
```

Term frequency analysis and Top Features:

The most frequent word roots shown in the DFM indicate that the data-driven companies focus on the following core values: “data”, ”people”, ”customers”, “innovation”, “business”, “integrity”, “challenge”, “value”, “communication”, “teamwork”, “technology”, “service”, “insight”, “power”, “culture”, “decision”, etc. Data-driven approach should have those features: innovative, integrate, value-created, team-oriented, powerful, better services for customers, etc.

3. Analysis of bigrams

```
cleancorpus_Vbi <- tokenize(newscorpus_V,
                            removeNumbers=TRUE,
                            removePunct = TRUE,
                            removeSeparators=TRUE,
                            removeTwitter=FALSE,
                            ngrams=2, verbose=TRUE)
dfm.bigram_V<- dfm(cleancorpus_Vbi, toLower = TRUE,
                   ignoredFeatures = c(stopwords("english")),
                   verbose=TRUE,
                   stem=FALSE)
topfeatures.bigram_V<-topfeatures(dfm.bigram_V, n=50)
topfeatures.bigram_V
```

```
> topfeatures.bigram_V
          big_data          every_day        secure_data       great_people      customers_can
                 4                  2                  2                  2                  2
       take_care       proving_points          data_free         free_lunch     lunch_seriously
                 2                  1                  1                  1                  1
seriously_every          day_high        high_growth     growth_software  software_companies
                 1                  1                  1                  1                  1
companies_brainy      brainy_alumni successful_startups      startups_sales      sales_marketing
                 1                  1                  1                  1                  1
marketing_alignment alignment_kegerators      faster_better   better_decisions decisions_connecting
                 1                  1                  1                  1                  1
connecting_human   human_intuition intuition_visualizations   human_attention  attention_intuitive
                 1                  1                  1                  1                  1
 intuitive_agile         agile_new            new_way            world_s            s_fastest
                 1                  1                  1                  1                  1
 fastest_easiest     data_platform      platform_solve demanding_business business_challenges
                 1                  1                  1                  1                  1
 problem_solvers       solvers_big      data_pioneers     pioneers_data         data_nerds
                 1                  1                  1                  1                  1
      hire_great        people_can       can_flourish       treat_people      other's_ideas
                 1                  1                  1                  1                  1
```

Analysis of bigram:

In the bigrams, we can see more detailed information. For example, "secure_data" indicates that companies should consider the venerability of exploring data; "growth_software" shows that the data analytics tools should be well updated to ensure the effectiveness; "human_intuition" points out that data analysis can support decision making but should not be the complementary to human decisions. "Hire-great" suggests that those successful firms also need talented people to support the data-driven approach.

Q3 Analyze the corpus and provide insight on how to structure a firm for data-analysis readiness

I would advise that leaders to clarify the importance of data analytics first. In addition, learning from the benchmarking can direct companies to a right way to create values for their companies, other businesses and individuals. When these happen, leaders should hire more talented people outside to train the current employees and make data-driven approach on a right track. Furthermore, teamwork and collaboration should be well considered when collecting valuable data and finding insights. Therefore, everyone in the company would feel like it is their changes that are being implemented.

****Please also find detailed analysis (in red) in the former two questions.**

Q4 Are there any other data-driven approaches you would recommend the CEO to implement?

Data-driven approaches are complicated, which can be applied to many industries. It can be used to analyze market, financial risks, public safety, education, etc. In addition, unstructured data, such as text is difficult do to analysis. Therefore, I would suggest CEOs to build their companies' core responsibility, clarify their missions and values of doing data analytics. Moreover, CEOs should focus on building powerful infrastructures, such as upgrading the functionalities of data analysis and visualization software, communicating with customers to deliver better services and focusing on innovation. Most importantly, a good data-driven company should find insights from data and provide a valuable solution for decision makers.

Homework part 2

Q1 Create a Corpus for the Speeches

a. Create Corpus

```
newscorpus2 <- corpus(DT$Full_Text,
docnames=DT$Speech_Topic, docvar=NULL)
#clean corpus
newcorpus2<- toLower(newscorpus2, keepAcronyms = FALSE)
cleancorpus2 <- tokenize(newscorpus2,
                         removeNumbers=TRUE,
                         removePunct = TRUE,
                         removeSeparators=TRUE,
                         removeTwitter=FALSE,
                         verbose=TRUE)
```

b. Explore the Corpus:

```
> summary(newscorpus2)  #summary of corpus
Corpus consisting of 4 documents.

    Text Types Tokens Sentences
 Speech1   536   1910       170
 Speech2  1469   6426       453
 Speech3  1129   8613       639
 Speech4   959   2783       138

Source:  /Users/apple/* on x86_64 by apple
Created: Mon Nov 14 22:01:04 2016
```

Q2 Complete a frequency analysis of word usage

a. Generate (DFM):

to create a custom dictionary list of stop words:

```
swlist2 = c("thank", "much", "can","will","just","trump")

dfm2<- dfm(cleancorpus2,
           toLower = TRUE,
           ignoredFeatures = c(swlist2, stopwords("english")),
           verbose=TRUE,
           stem=TRUE)
```

Reviewing top features:

```
view(dfm2)
```

	everybody	sorry	keep	waiting	complicated	business	received	call	secretary	clinton	congratulated	us	victory	family	hard-fought	campaign	mean	fought	hard	hillary
Speech1	1	1	1	1	2	2	1	4	1	1	2	7	2	5	1	5	2	1	3	1
Speech2	3	0	4	1	1	1	0	4	2	14	0	6	0	0	0	0	3	0	5	15
Speech3	10	1	2	0	0	1	0	5	1	1	0	27	1	7	0	1	1	0	1	0
Speech4	0	0	0	0	0	0	1	1	3	13	0	5	0	0	0	2	0	0	0	11

```
topfeatures(dfm2, 100)        # displays 100 features
```

```
> topfeatures(dfm2, 100)     # displays 100 features
      people         going         great          know            re       country           now           one
         121           102            73            72            72            70            54            50
        want           don           say            us          said           get     immigration         right
          50            50            46            45            42            40            36            35
        like          back       clinton          take          even        states          make        number
          33            32            29            29            29            29            29            29
        need          many        really           new            ll       hillary         world       illegal
          29            28            28            28            28            27            27            27
        jobs          time        united          also       billion      military      american           big
          27            26            26            26            25            25            24            24
       years         china     president       believe           got         never          good          work
          24            24            23            23            22            22            22            21
      happen          tell         think          love           way        mexico        border         money
          21            21            21            20            20            20            20            20
    audience          come administration    immigrants       defense       america           law        member
          20            19            19            19            19            18            18            18
     million            ve          plan            go          ever        office         bring         build
          18            18            17            17            17            17            17            17
      things         first     countries        system         obama           day           put       nothing
          16            16            16            16            16            16            15            15
  thousands          iraq     everybody          call           job          look           lot         place
          15            15            14            14            14            14            14            14
     needs      criminal          nice       support         folks           end           two         state
          14            14            14            13            13            13            13            13
     today    politicians       leaders        family
          13            13            13            12
```

Analysis of DFM:

First, because the texts are speeches, so I choose to create new dictionary to screen out some common words in speeches, such as “can”, “thank”, “will”, “much” and “just”. In the DFM, we can see that Mr. Trump’s most frequent words are “people”, “going”, “great”, “know”, “re-”,” country”, “now”, “one”, “want”. These words are all simple monosyllables. In addition, besides the top 10 frequent words, we can also see some common topics, such as “immigration”, “military”, “jobs”, “Mexico”, “China”, “criminal”. Furthermore, he also mentioned his competitors Hillary a lot in his speeches.

b. Find Root words:

```
dfm.stem2 <- dfm(cleancorpus2, toLower = TRUE,
                 ignoredFeatures = c(swlist2, stopwords("english")),
                 verbose=TRUE,
                 stem=TRUE)
topfeatures.stem2 <-topfeatures(dfm.stem2, n=50) #fifty common words
topfeatures.stem2
```

```
> topfeatures.stem2
   peopl       go  countri     know    great       re     want   immigr      now      say      one      don      get
     121      119       87       75       74       72       59       58       54       54       54       50       48
    need       us     said    state      job     take     make    right     work     like     year  clinton american
      46       45       42       42       41       38       38       35       34       34       34       32       32
    back    illeg   number     come     even    build   happen      law     mani   realli      new       ll  hillari
      32       31       31       30       30       30       29       29       28       28       28       28       27
    time    world     unit    thing     also     tell   border    think  billion militari     call
      27       27       26       26       26       26       26       26       25       25       24
```

Analysis of Root Words:

The root words can be used to justify the conclusion showed in the “Analysis of DFM”, which indicate the most frequent root words and the most likely topics in Mr. Trump’s speeches.

c. Analyzing DFM with bigrams:

```
cleancorpus_bi <- tokenize(newscorpus2,
                           removeNumbers=TRUE,
                           removePunct = TRUE,
                           removeSeparators=TRUE,
                           removeTwitter=FALSE,
                           ngrams=2, verbose=TRUE)

dfm.bigram2 <- dfm(cleancorpus_bi, toLower = TRUE,
                   ignoredFeatures = c(swlist2, stopwords("english")),
                   verbose=TRUE,
                   stem=FALSE)
topfeatures.bigram2 <-topfeatures(dfm.bigram2, n=50)
topfeatures.bigram2
```

```
> topfeatures.bigram2
      united_states      hillary_clinton      audience_member   illegal_immigrants         middle_east
                 25                   24                   17                   13                   11
          re_going      law_enforcement      criminal_aliens     president_obama          bring_back
                 11                    9                    9                    7                    7
            ll_say           net_worth      missile_defense          take_care            right_now
                  6                    6                    6                    5                    5
        great_job   immigration_system    special_interests        open_borders          number_one
                  5                    5                    5                    5                    5
      right_people    islamic_terrorism             web_site        make_america          member_yes
                  5                    5                    5                    5                    5
      saudi_arabia         nice_person      foreign_policy        great_people      american_people
                  5                    5                    5                    4                    4
         many_many     sanctuary_cities            air_force      billion_dollars           one_thing
                  4                    4                    4                    4                    4
           day_one           even_know       common_sense       need_somebody            re_gonna
                  4                    4                    4                    4                    4
       common_core     radical_islamic tremendous_potential        inner_cities         truly_great
                  4                    4                    3                    3                    3
  talented_people           new_york          four_years         white_house illegal_immigration
                  3                    3                    3                    3                    3
```

Analysis of bigrams:

The bigrams provide more detailed information. Mr. Trump's speeches focus on making United States a great country, his most frequent topic would be "illegal immigrants" and "jobs".

Q3 Complete a sentiment analysis

```
mydict2<- dictionary(list(negative = c("detriment*", "bad*", "awful*", "terrib*",
"horribl*","stupid","weak", "loser","tough","dangerous","zeor","hate","worse"),

postive = c("fantastic","classy","good", "great", "super*", "excellent",
"yay","win","smart","amazing","terrific")))      ###create your own dictionary

dfm.sentiment2 <- dfm(cleancorpus2, dictionary = mydict2)
topfeatures(dfm.sentiment2)
View(dfm.sentiment2)
```

	negative	postive
Speech1	6	34
Speech2	20	28
Speech3	24	53
Speech4	2	3

	negative	positive
Speech1	15%	85%
Speech2	41.67%	58.33%
Speech3	31.17%	68.83%
Speech4	40%	60%

Sentiment analysis:

The outcomes indicate that speech 1 (victory speech) used more positive words; speech 2 used more positive words but the difference between negative words and positive words is small; speech 3 use more positive words; speech 4 is more neutral because there are few emotional words.

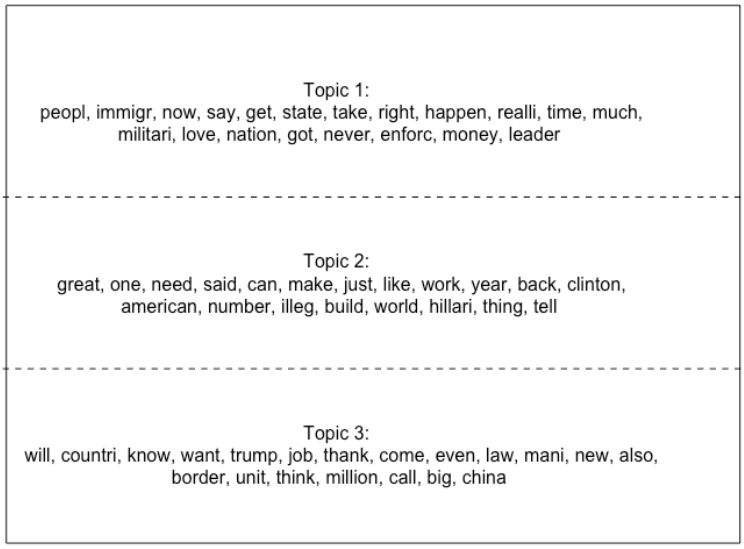
Q4 What are the common topics in the corpus

```
prevfit2 <-stm(docs , vocab ,
               K=3,
               verbose=TRUE,
               data=meta,
               max.em.its=10)

topics <-labelTopics(prevfit2 , topics=c(1:3))
topics      #shows topics with highest probability words

#explore the topics in context.    Provides an example of the text
help("findThoughts")
findThoughts(prevfit2, texts = DT,    topics = 1,    n = 1)

help("plot.STM")
plot.STM(prevfit2, type="summary")
plot.STM(prevfit2, type="labels", topics=c(1,2,3))
plot.STM(prevfit2, type="perspectives", topics = c(1,2))
```



Common topic analysis:

According to the two charts shown above, the common topics are:

1. Immigration Issue
2. How to make America to be great
3. Trump will bring jobs to the country

Context analysis (the screen shots below are just two sample words):

kwic(cleancorpus2, "believe", 5)

kwic(cleancorpus2, "great", window = 3)

```
                                 contextPre keyword                    contextPost
[Speech1, 1163]          A very special person who [ believe ] me I read reports that
[Speech2, 1791]               are we doing Hard to [ believe ] Hard to believe Now that
[Speech2, 1794]            Hard to believe Hard to [ believe ] Now that you've heard about
[Speech2, 2089]              work with us I really [ believe ] it Mexico will work with
[Speech2, 2098]           work with us I absolutely [ believe ] it And especially after meeting
[Speech2, 2112] wonderful president today I really [ believe ] they want to solve this
[Speech2, 2258]               so great It's hard to [ believe ] people don't even talk about
[Speech2, 2847]               And they will go face [ believe ] me They're going to go
[Speech2, 4080]          the right people doing it [ believe ] me very very few will
[Speech2, 4362]             take them back Hard to [ believe ] with the power we have
[Speech2, 4370]              power we have Hard to [ believe ] We're like the big bully
[Speech2, 4735]         If people around the world [ believe ] they can just come on
 [Speech3, 242]            They are not our friend [ believe ] me But they re killing
 [Speech3, 457]             hotel in Syria Can you [ believe ] this They built a hotel
 [Speech3, 794]              from to percent Don t [ believe ] the Don t believe it
 [Speech3, 798]                t believe the Don t [ believe ] it That s right A
[Speech3, 1138]             They will not bring us [ believe ] me to the promised land
[Speech3, 1274]            level that you wouldn t [ believe ] It makes it impossible for
[Speech3, 3190]           my opinion the new China [ believe ] it or not in terms
[Speech3, 3460]          them one for each country [ Believe ] me folks We will do
[Speech3, 4236]            there except for us And [ believe ] me you look at the
[Speech3, 5671]                    I don t have to [ believe ] it or not I m
[Speech3, 5894]        builds walls better than me [ believe ] me and I ll build
```

	contextPre	keyword	contextPost
[Speech1, 196]	and unify our [	great]	country As I've
[Speech1, 214]	an incredible and [	great]	movement made up
[Speech1, 421]	care of our [	great]	veterans who have
[Speech1, 505]	We have a [	great]	economic plan We
[Speech1, 546]	We will have [	great]	relationships We expect
[Speech1, 552]	expect to have [	great]	great relationships No
[Speech1, 553]	to have great [	great]	relationships No dream
[Speech1, 564]	challenge is too [	great]	Nothing we want
[Speech1, 709]	me right now [	Great]	people I've learned
[Speech1, 724]	every regard Truly [	great]	parents I also
[Speech1, 757]	brother Robert my [	great]	friend Where is
[Speech1, 779]	that's okay They're [	great]	And also my
[Speech1, 786]	late brother Fred [	great]	guy Fantastic guy
[Speech1, 796]	was very lucky [	Great]	brothers sisters great
[Speech1, 799]	Great brothers sisters [	great]	unbelievable parents To
[Speech1, 876]	much What a [	great]	group You've all
[Speech1, 1034]	is Jeff A [	great]	man Another great
[Speech1, 1037]	great man Another [	great]	man very tough
[Speech1, 1432]	to do a [	great]	job and I
[Speech1, 1449]	will do a [	great]	job We will
[Speech1, 1455]	will do a [	great]	job I look
[Speech2, 209]	also discussed the [	great]	contributions of Mexican-American
[Speech2, 679]	many of the [	great]	parents who lost
[Speech2, 998]	Force veteran a [	great]	woman according to
[Speech2, 1536]	our country with [	great]	dignity So important
[Speech2, 1911]	Immigration offices very [	great]	people Among the
[Speech2, 1976]	will build a [	great]	wall along the
[Speech2, 2006]	it And they're [	great]	people and great
[Speech2, 2009]	great people and [	great]	leaders but they're

The outcomes indicate that Mr. Trump applied catchphrases, which are the speech styles that salesmen use, such as “believe me”, “many people are saying” and “great”.

Advanced topic modeling visualization:
http://127.0.0.1:4321/#topic=0&lambda=1&term=

Q5 Write a memo style report summarizing insting on Trump’s linguistic effectiveness

MEMORANDUM
To: All BA members
From: Yuchen Liu
Date: Nov 16, 2016
Subject: Trump’s Linguistic Effectiveness Report

Mr. Trump usually applies simple monosyllables in his speeches; he basically just uses casual speech in a public setting. In addition, Mr. Trump’s speeches are filled with sentiments when the speech topics are relating immigration and jobs, which indicates that he is adept in connecting audiences on an emotional level. He often uses catchphrases, which are actually versions of speech mechanisms that salesmen use. Furthermore, the topics he often targets in his speeches are “immigration” and “jobs”.

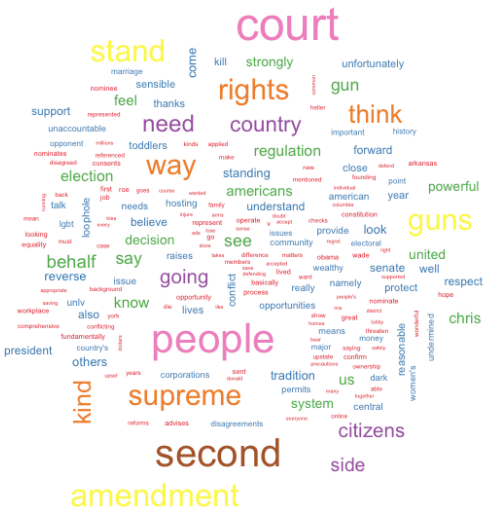
I will be glad to discuss these conclusions and follow through on how to compute statistical significance between Trump’s and other presidents’ speech styles.

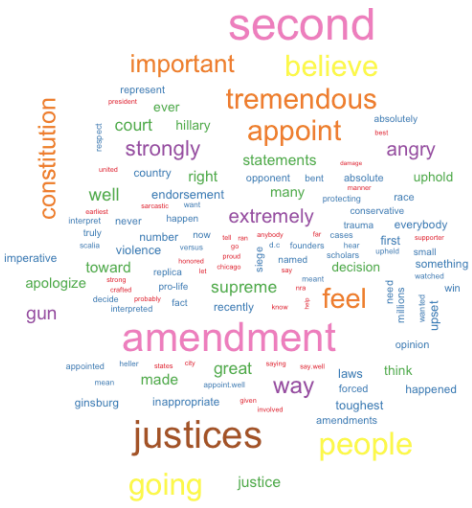
Thank you.

Extra Research

World Cloud of the Third Debate between Hillary Clinton and Donald Trump

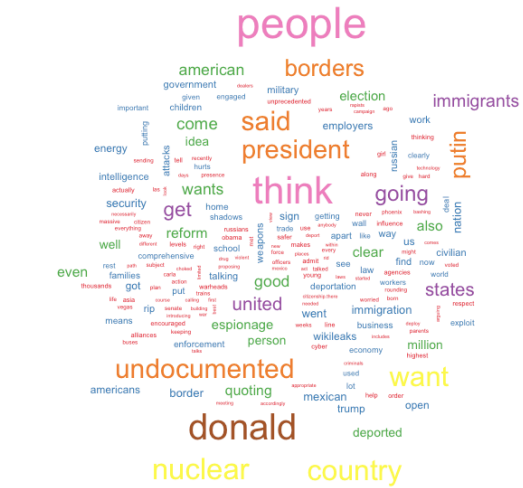
Topic: Supreme Court

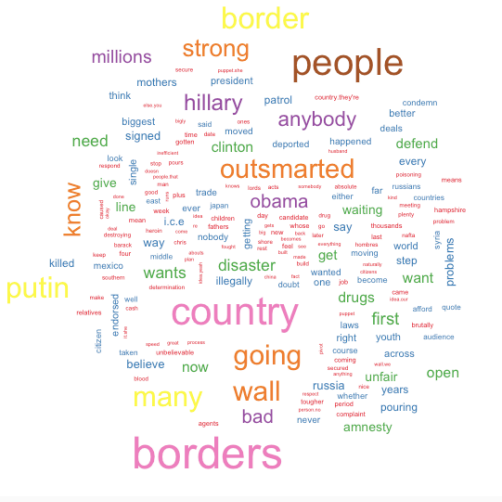




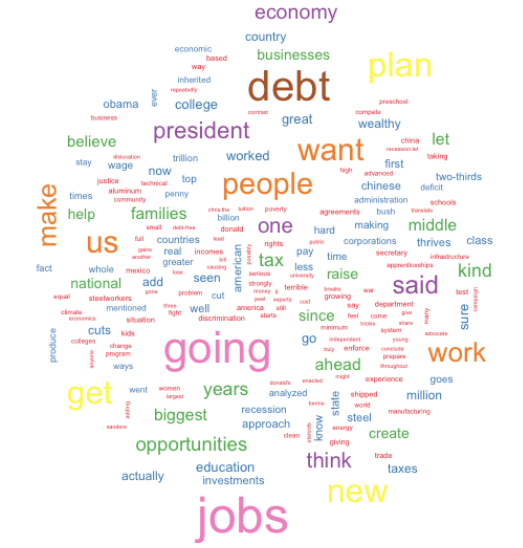
Topic: Abortion

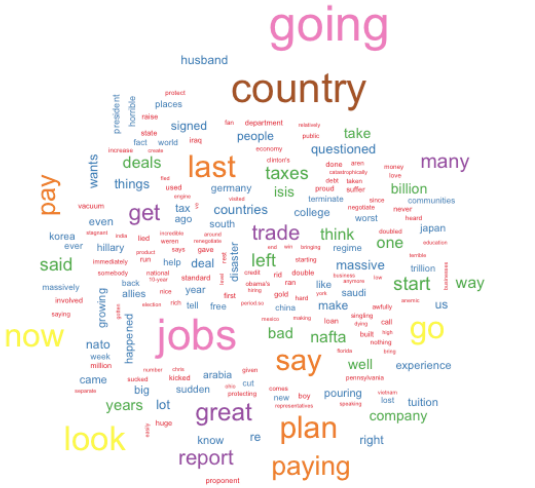
Topic: Immigration



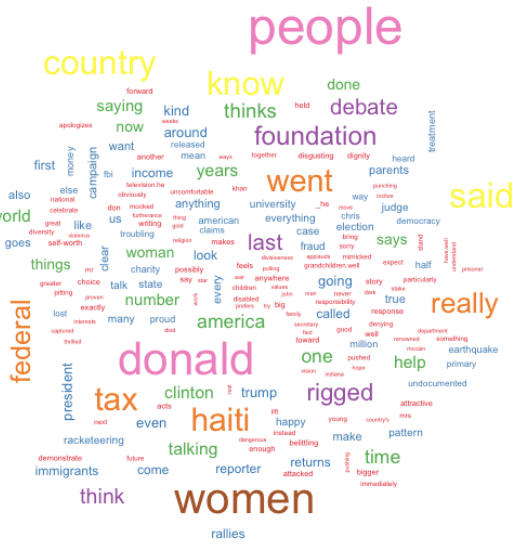


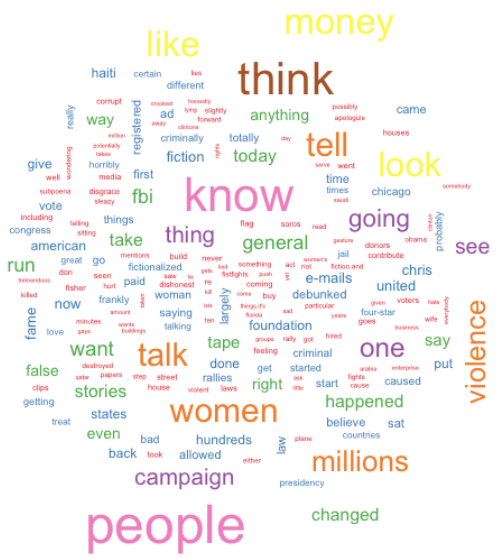
Topic: Economy



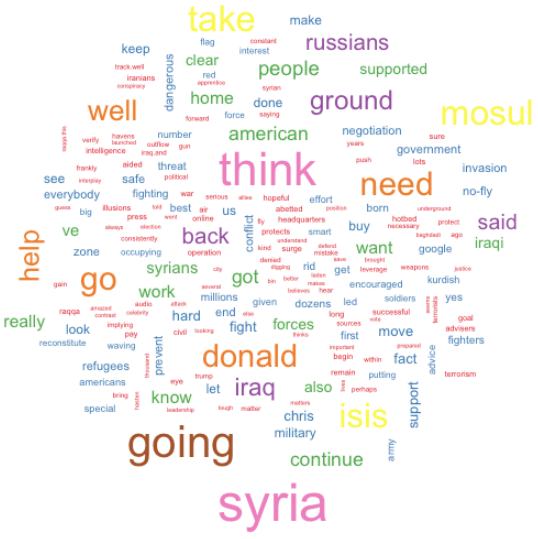


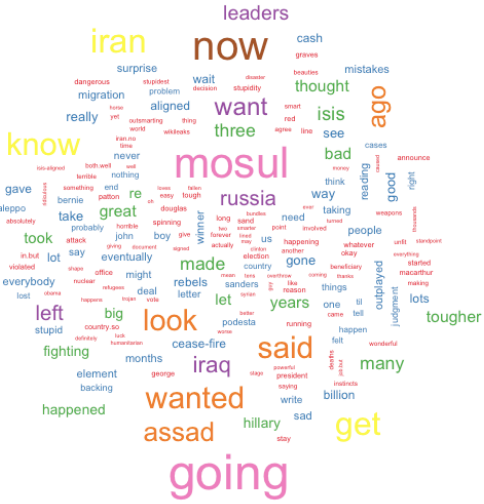
Topic: Fitness to be president of the United States



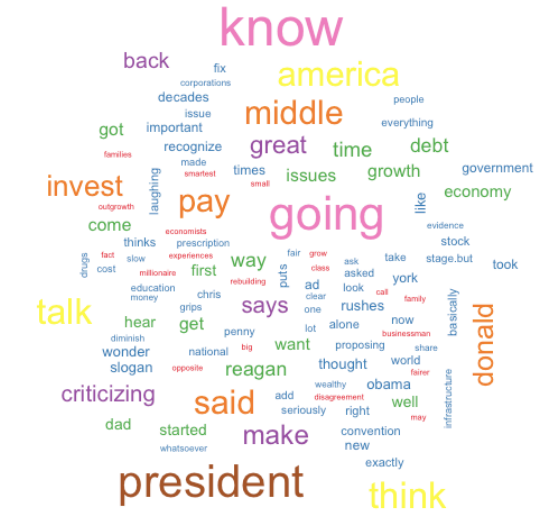


Topic: Foreign hotspots

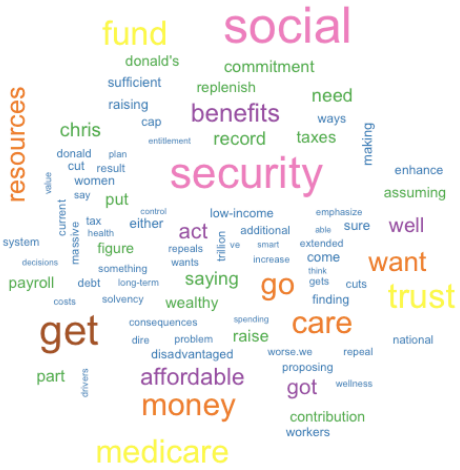


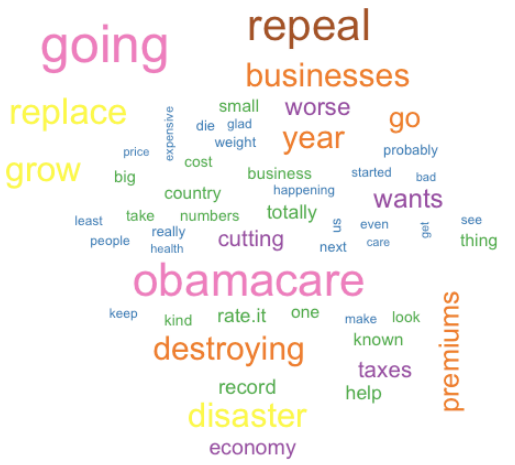


Topic: National debt

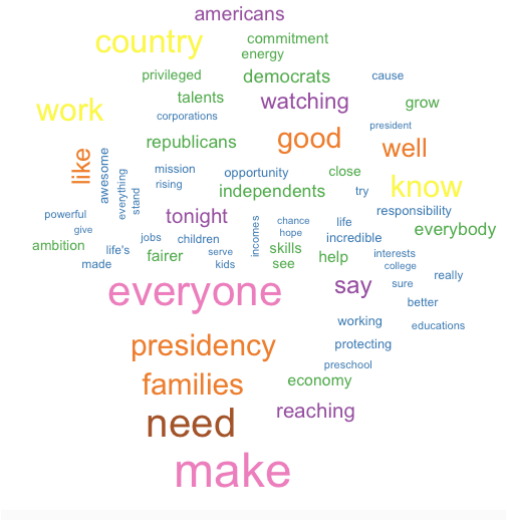


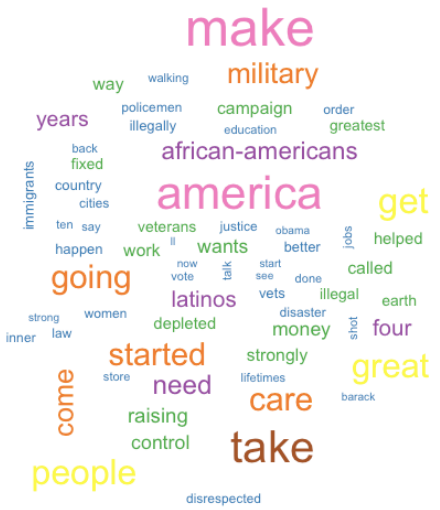
Topic: 60% of all federal spending





Topic: Why they should elect you





Different from Donald Trump, Hillary Clinton say "look" a lot, show a little bit of emotion – it would be highly effective.