Text Analytics of Open-Ended Survey Data: Techniques and Applications

Xiaohui Zheng
Institutional Research and Academic Planning
University of California, Office of the President

Presentation at CAIR Conference
San Diego, CA
November 20th, 2014

Presentation Outline

Four Parts

- Introduction
- Data Processing
- Analytic Techniques
- Application

Introduction

Text Analytics

- Automated processing of texts
- Derive information from unstructured texts
- Statistical and machine learning techniques
- Highly interdisciplinary: statistics, computer science and linguistics

Introduction

Text Mining vs. Data Mining

- A variation of data mining
- Data mining: structured data
- Text Mining: unstructured/semi-structured data
- Examples: Open responses, full-text docs, html files, Tweets

Introduction

Values

- Processing texts in large volumes at high speed
- Academic and business intelligence applications
- Survey research: Open-response analysis
- Business applications: detect spams, classify news, marketing research

Introduction

Limitations

- Complexity in natural language
  - Spelling variations
  - Linguistic patterns
  - Contextual meaning
  - Semantic ambiguity

Data Processing

Process

- Text Import
- Text Transformation
- Document-by-Term Matrix

Data Processing

Text Import

- Corpus: a collection of texts
- Texts from different formats
- Supported formats: Txt, Word, Excel, Pdf, Html, etc.

Data Processing

Text Transformations

- Remove special characters
- Convert upper cases to lower cases
- Remove numbers, punctuations, whitespace
- Remove stopwords with no substantial meaning
- Replace synonyms (e.g. ‘pay’ with ‘salary’)

Data Processing

Document-by-term Matrix

- A structured representation
- Describes the frequency of terms
- Documents as rows, terms as columns and counts as cells
- Fundamental unit where we perform text analytics

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1

Techniques

- Frequent Terms
- Weighted Frequencies
- Word Cloud
- Associations
- Concept Linking
- Text Clustering

Techniques

Frequent Terms

- obtain term frequencies by summing the column counts
- Identify most/least frequent terms

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2
Sum:	9	5	14	14	2	12	2	8

Techniques

Weighted Frequencies

- Term frequency-inverse document frequency (TF-IDF)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

TF= # times term appears/ # of terms in the doc

How frequently a term occurs in a document

IDF=log (total # of docs/# of docs with the term)

Weighs down the frequent terms and scales up rare ones

Techniques

Word Cloud

- Visualize relative frequencies of words
- Term size proportional to its frequency
- Add colors to words at different frequency levels

Techniques

Term Association

- Correlation between a given term and all the other terms
- Correlation indicates how closely related two terms are
- Concept linking graph: Visually display correlations among frequent terms

Techniques

Text Clustering

- Group documents with similar contents
- Unsupervised learning: clustered developed on the fly
- No single algorithm that works best in all situations:
  - Hierarchical clustering
  - K-means clustering

Application

Survey Question

8. In the past year, have you seriously considered leaving UCOP?
   - ❍ No [Go to Question 11]
   - ❍ Yes

9. If you wish to elaborate on why you seriously considered leaving, please do so here.

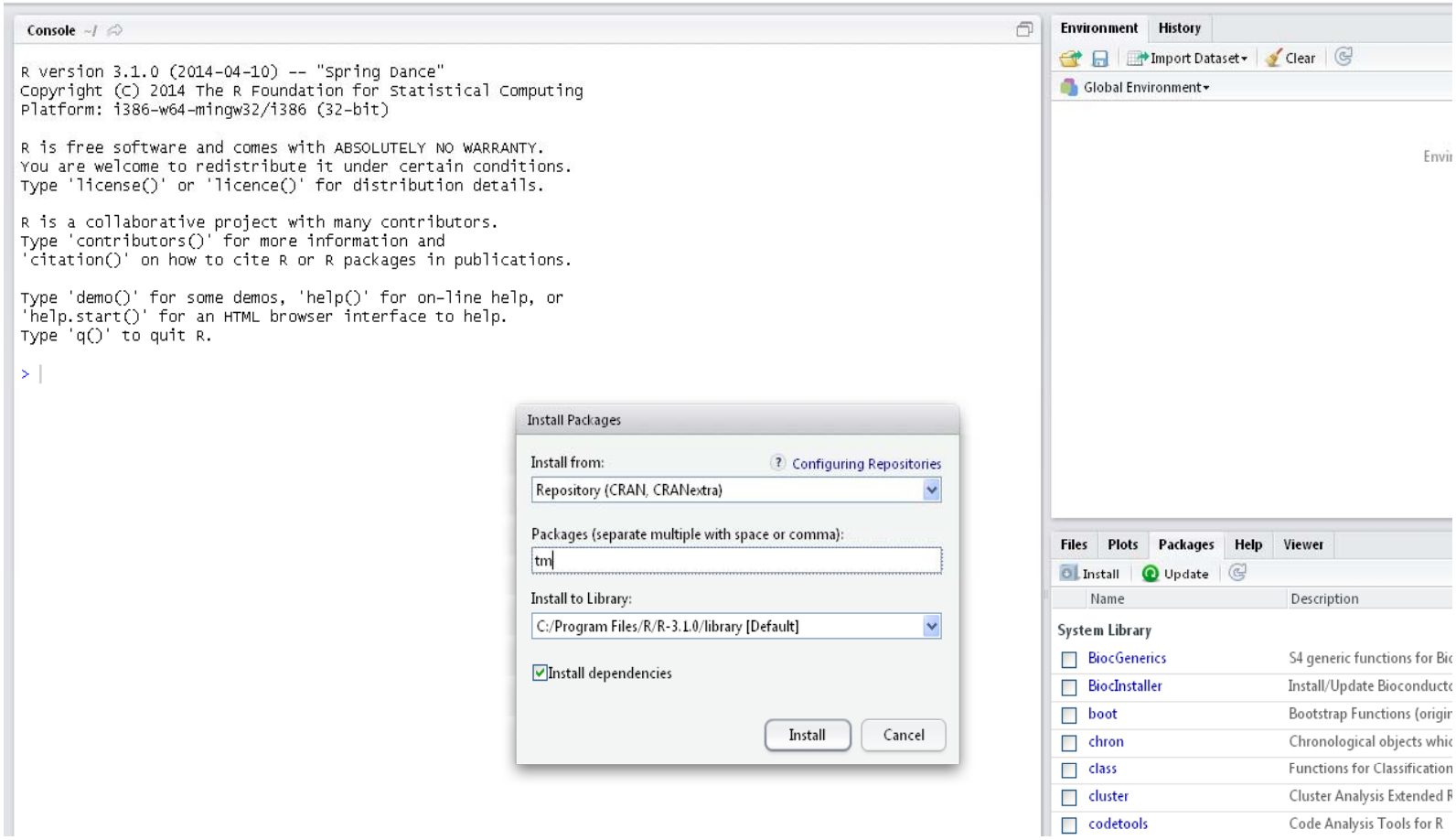
_____

_____

_____

Application

Installing ‘tm’ package



Application

Text Import

```
> x <- read.csv("X:\\Climate Survey\\opentext.csv", header = F)
> q9<- Corpus(DataframeSource(x))

> x <-file.path("X:/Climate Survey/OpenText/files")
> q9<-Corpus(DirSource(x))

> inspect(q9)
<<VCorpus (documents: 332, metadata (corpus/indexed): 0/0)>>
[[1]]
<<PlainTextDocument (metadata: 7)>>
```

Application

Text Transformations

- Default functions

```
> q9 <- tm_map(q9, tolower)
> q9 <- tm_map(q9, removePunctuation)
> q9 <- tm_map(q9, removeNumbers)
> q9 <- tm_map(q9, stripWhitespace)
```

- Customized functions

```
> removeslash <- function(x) gsub("/", " ", x)
> q9<- tm_map(q9, removeslash)
> myStopwords <- c(stopwords('english'), "ucop", "within", "etc", "feel", "like", “get“)
> q9 <- tm_map(q9, removeWords, myStopwords)
```

Application

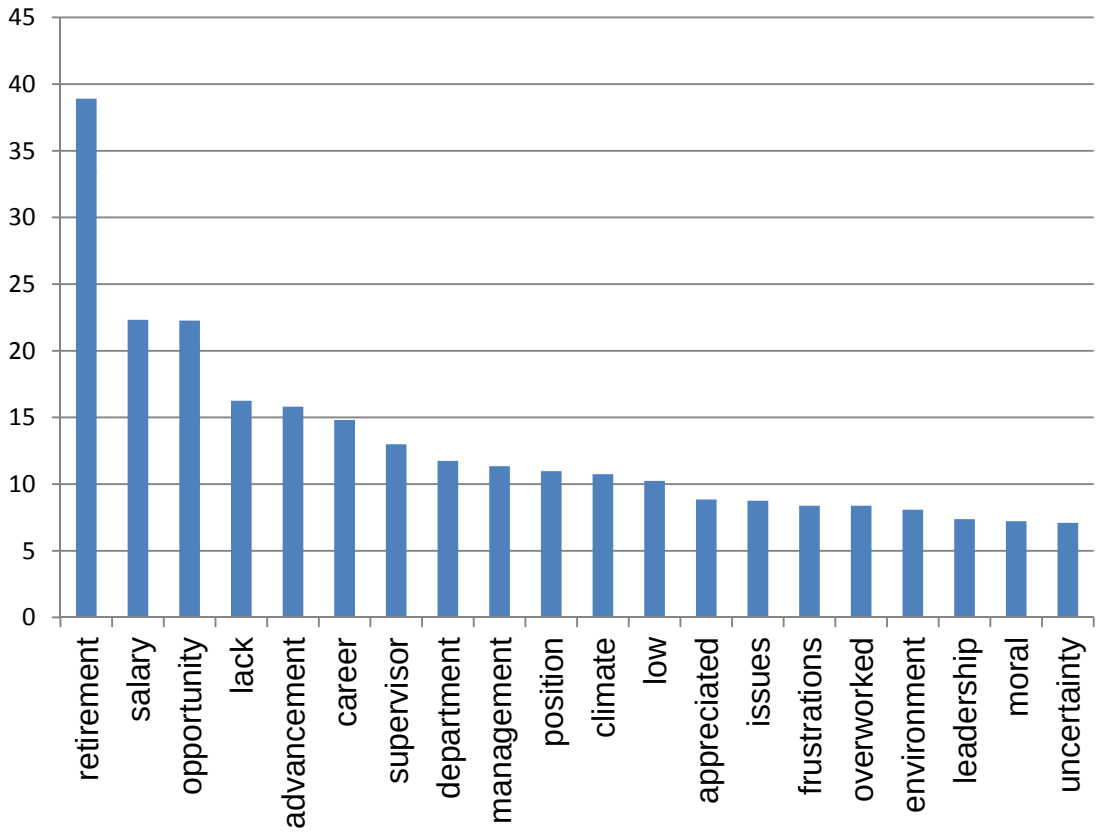
Document-by-Term Matrix

```
# Constructs the matrix based on term frequencies
> dtm<-DocumentTermMatrix(q9)

# Constructs the matrix based on weighted frequencies
> dtm2<-DocumentTermMatrix(q9, control=list(weighting = weightTfIdf))

# Adds up the counts by column
>freq<-colSums(as.matrix(dtm2))

# Outputs the summed frequencies into Excel
> write.csv(as.matrix(freq), file="dtm2.csv")
```

Application

Frequent Terms

Rank	Term	TF-IDF
1	retirement	38.9
2	salary	22.3
3	opportunity	22.3
4	lack	16.3
5	advancement	15.8
6	career	14.8
7	supervisor	13.0
8	department	11.7
9	management	11.4
10	position	11.0
11	climate	10.7
12	low	10.2
13	appreciated	8.9
14	issues	8.8
15	frustrations	8.4
16	overworked	8.4
17	environment	8.1
18	leadership	7.4
19	moral	7.2
20	uncertainty	7.1



Application

Word Cloud ('wordcloud' package)

Application

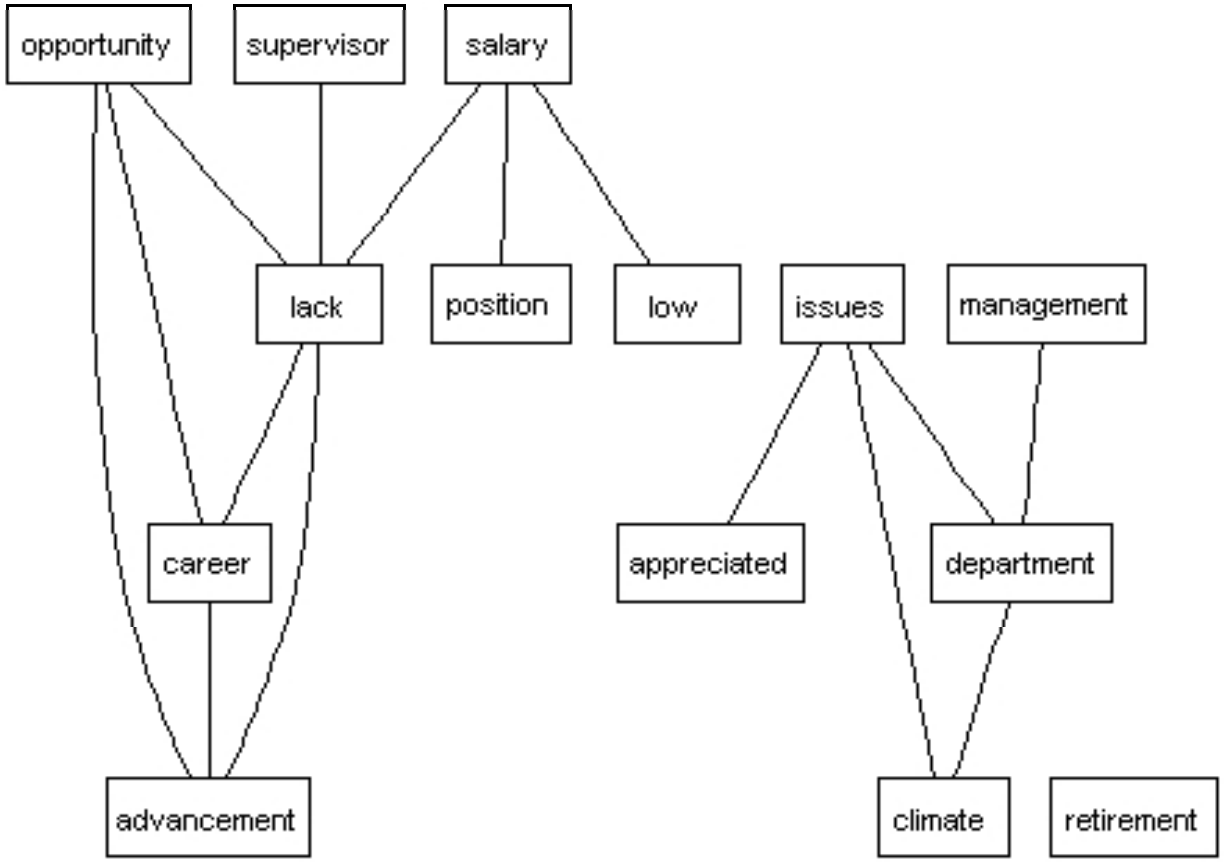
Term Associations

- Show all terms associated with ‘lack’ at corr>= 0.2

```
> findAssocs(dtm2, "lack", corlimit=0.2)
             lack
organization 0.37
increases    0.35
raises       0.25
trust        0.24
vision       0.22
```

Application

Correlation Plot ('Rgraphviz' package)



Application

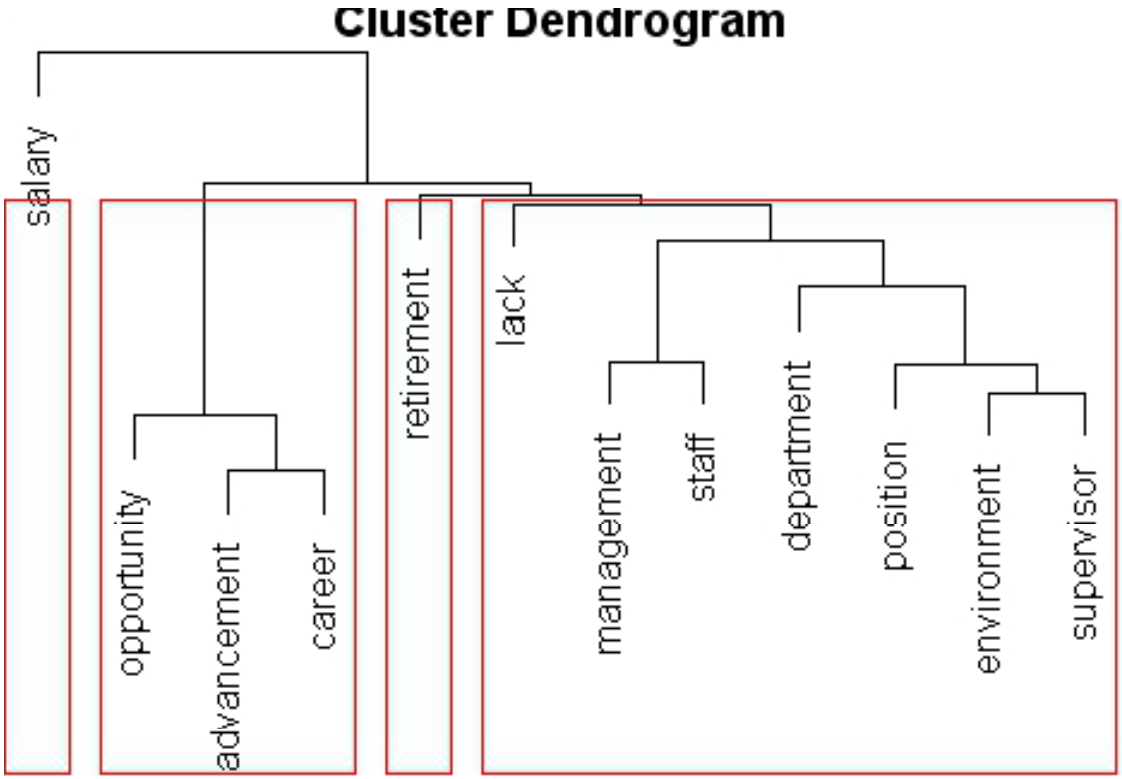
Cluster Analysis

```
# Hierarchical cluster analysis using Ward's minimum variance criteria
> distancematrix<-dist(matrix, method="euclidian")
> model<- hclust(distancematrix, method="ward")

# Display dendrogram
> plot(model)

# Add rectangles around the branches of the dendrogram highlighting clusters
> rect.hclust(model, k=4, border="red")
```

Application

Cluster Analysis

Terms	Cluster
advancement	1
career	1
opportunity	1
department	2
environment	2
lack	2
management	2
position	2
staff	2
supervisor	2
retirement	3
salary	4



The End