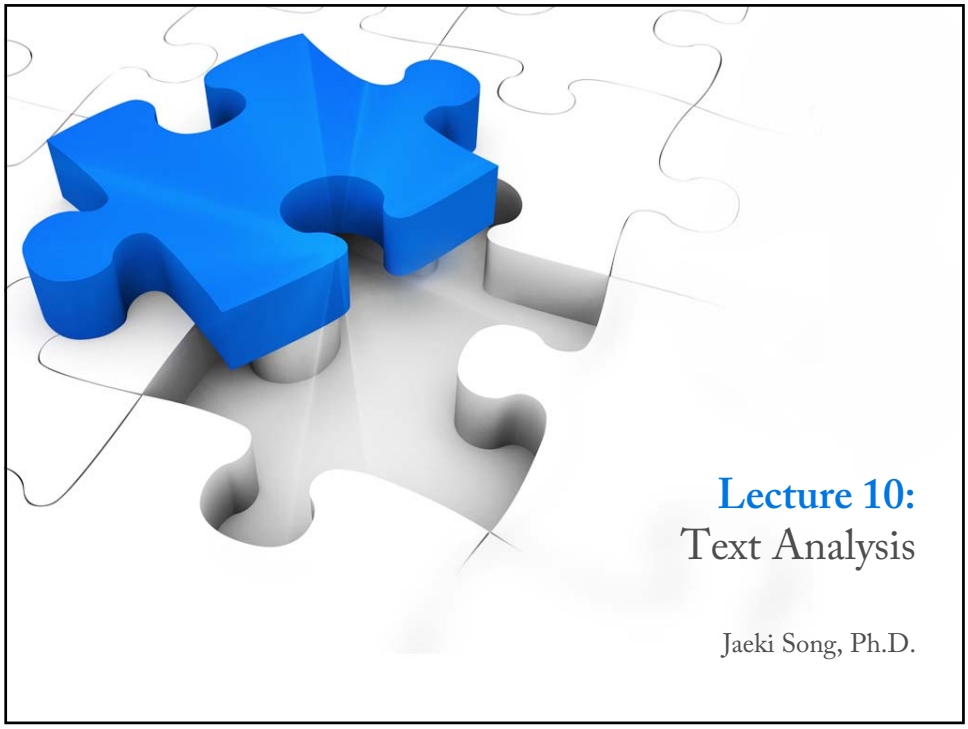


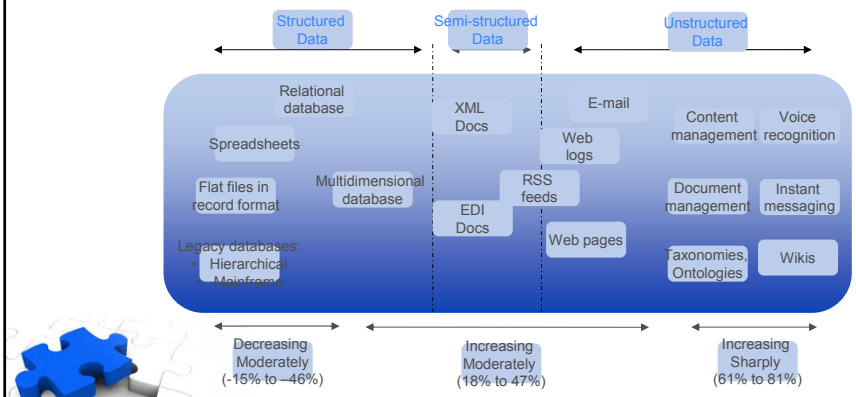
Lecture 10:
Text Analysis

Jaeki Song, Ph.D.

Text Analysis

Why text? - Explosions of Unstructured Data

- 800%+ growth in data volume within next 5 years
- Amount of unstructured data is growing 62% faster
- 80% of data will be unstructured data in 2019



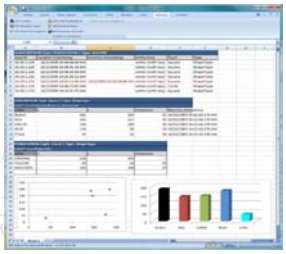
Three Major Areas within Data Continuum

Source: TDWI Philip Russom 2007

Why text? - Explosions of Unstructured Data

Unstructured data:

- Diverse types, many inputs
- Text, audio, image, video, metadata, health records, etc.
- Need to be able to search, compare, understand, and prediction



Structured data:

- Well-studied
- Columnar + Relational
- Interval/categorical/ordinal

What is text mining?

- Use of *computational techniques* to extract high quality information from text
- Extract and discover knowledge hidden in text *automatically*
- Discovery by computer of new previously unknown information, by automatically extracting information from a usually *large amount* of different *unstructured* textual resources

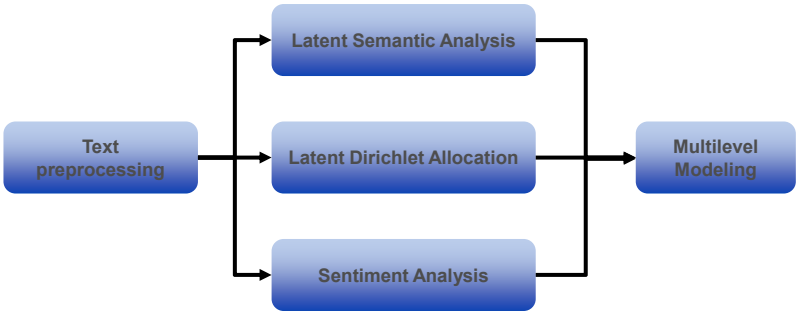
Why text mining?

- Leveraging text should improve decisions and predictions
- Text mining is gaining momentum
  - Sentiment analysis (twitter, Facebook)
    - Predicting stock market
    - Predicting churn
    - Customer influence

Basic Text Mining Tasks

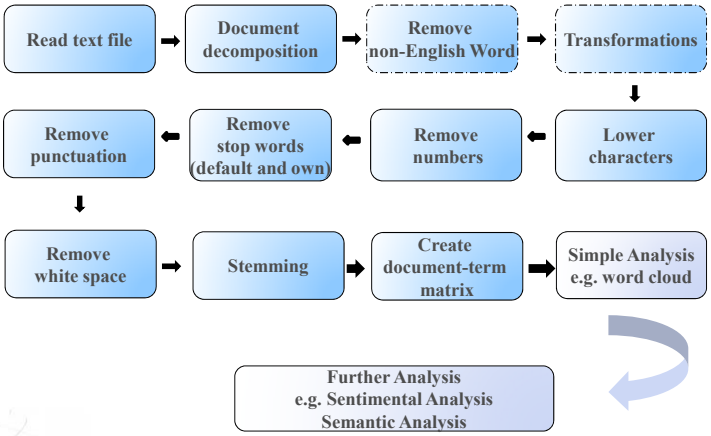
- Document classification (categorization)
- Information Retrieval
- Clustering/Organization of Documents
- Information Extraction

Text Analysis Process



Text Preprocessing

Text Mining – General Procedure



Text Information

Online Customer Review for "HBO Now" Mobile Application

It successfully charged me $14.99 for subscription. It doesn't let me watch anything by asking me to renew subscription. In billing, it says I am not subscribed. I try to play a video and when renew subscription dialog pops up, I pressed the Renew Button. It gets stuck at Validating Account. Their phone support is unavailable 24x7 and gives a busy signal instead of telling us about their work time or that it is closed currently. I used email support and got someone to respond within a couple of hours. I sent the necessary information - an attachment of the subscription confirmation from play store. After that, no response at all. Today I called phone support. I have been on the wait queue for 20 minutes. I think I just lost my 15 bucks and ruined my Sunday night and wasted my time dealing with their incompetence. It would have been nice if they at least acknowledged the issue and sent out an email with an ETA of the fix to everyone (looking at reviews, I am not the only one facing this issue) affected by the issue and maybe a free month or 2 to compensate for the loss of time and happiness.

Cost

Functionality

Customer Service

Subscription

Text Mining – Lower Case and Remove Numbers

14.99

24x7

it successfully charged me $ for subscription. it doesn't let me watch anything by asking me to renew subscription. in billing, it says i am not subscribed. i try to play a video and when renew subscription dialog pops up, i pressed the renew button. It gets stuck at validating account. their phone support is unavailable x and gives a busy signal instead of telling us about their work time or that it is closed currently. i used email support and got someone to respond within a couple of hours. i sent them the necessary information - an attachment of the subscription confirmation from play store. after that, no response at all. today i called phone support. i have been on the wait queue for minutes. i think i just lost my bucks and ruined my sunday night and wasted my time dealing with their incompetence. it would have been nice if they at least acknowledged the issue and sent out an email with an eta of the fix to everyone (looking at reviews, i am not the only one facing this issue) affected by the issue and maybe a free month or to compensate for the loss of time and happiness.

20

15

Text Mining – Remove Stop Words

- English Stop Words: I, the, he, they, is, am, don't, does, not, etc.
- Self-defined Stop Words: HBO, app, now

successfully charged $. subscription. let watch anything asking renew subscription. billing, says subscribe. try play video renew subscription dialog pops , pressed renew button. gets stuck validating account. phone support unavailable x gives busy signal instead telling us work time closed currently. used email support got someone respond within couple hours. sent necessary information - attachment subscription confirmation play store. , response . today called phone support. wait queue minutes. think just lost bucks ruined sunday night wasted time dealing incompetence nice least acknowledged issue sent email eta fix everyone (looking reviews, one facing issue) affected issue maybe free month compensate loss time happiness

Text Mining – Remove Punctuation and Space

- Punctuation: colon, comma, period, dash, hyphen, parentheses, question mark, etc.

successfully charged subscription let watch anything asking renew subscription billing says subscribed try play video renew subscription dialog pops pressed renew button gets stuck validating account phone support unavailable x gives busy signal instead telling us work time closed currently used email support got someone respond within couple hours sent necessary information attachment subscription confirmation play store response today called phone support wait queue minutes think just lost bucks ruined sunday night wasted time dealing incompetence nice least acknowledged issue sent email eta fix everyone looking reviews one facing issue affected issue maybe free month compensate loss time happiness

Text Mining – Stemming

success charg subscript let watch anyth ask renew subscript bill say subscrib tri play video renew subscript dialog pop press renew button get stuck valid account phone support unavail x give busi signal instead tell us work time close current use email support got someon respond within coupl hour sent necessari inform attach subscript confirm play store respons today call phone support wait queue minut think just lost buck ruin sunday night wast time deal incompet nice least acknowledg issu sent email eta fix everyon look review one face issu affect issu mayb free month compens loss time happi

Sample Results of LSA

Factor Label	High-loading Terms
Cost	pric, discount, purchas, bui, monei, tim, wast,
Customer Service	support, telephon, servic, call, email,
Functionality	work, watch, plai, enjoi, stuck, show
Connectivity	connect, wifi, speed, rat, brows,
Subscription	subscript, subscrib, renew, updat, trail

Required Package in R

- Library(tm)
  - Data import
  - Corpus handling
  - Preprocessing
  - Term-document matrices creation
- Library(Matrix)
  - Deal with matrix
- Library(SnowballC)
  - Stemming
- Library(wordcloud)
  - Visual overview
- Library(lsa)
  - Concept extraction
- Library(pdftools)
  - Read pdf file
- Library(qdapTools)
  - Read word file
- Library(stringr)
  - Deal with string

Read Text Files

- Read files one by one
  - Text:

```
> filepath <- "~/Desktop/class_code/processtext629/txt"
> setwd(filepath)
> corpus <- Corpus(DirSource(filepath))
```

  - PDF:
  - Word:

```
> filepath <- "~/Desktop/class_code/processtext629/pdf"
> setwd(filepath)
> files <- list.files(pattern = "pdf$")
> text<-lapply(files,pdf_text)
> text<-lapply(text,str_replace_all,"[\n]" , " ")
> corpus <- Corpus(VectorSource(text))
```

- Read lines in one file
  - Open the file and use “read” function

```
> filepath <- "~/Desktop/class_code/processtext629"
> setwd(filepath)
> file<-'HBO_NOW.txt'
> text=file(file,open="r")
> text.decomposition=readLines(text)
> close(text)
> corpus <- Corpus(VectorSource(text.decomposition))
```

Lower Characters & Remove Punctuations, Numbers, and White Space

- All functions are available in library(tm)
  - Lower characters: tolower
  - Remove punctuation: removePuctuatioin
  - Remove numbers: removeNumbers
  - Remove white space: stripWhitespace

```
> corpus <- Corpus(VectorSource(doc.decomposition))
> corpus <- tm_map(corpus,PlainTextDocument)
> corpus <- tm_map(corpus, tolower) # Transfer all words to lower characters
> corpus <- tm_map(corpus, removePunctuation) # Remove puctuation
> corpus <- tm_map(corpus,removeNumbers) # Remove numbers
> corpus <- tm_map(corpus, stripWhitespace) #Remove whitespace
```

Remove Stop Words

- Remove a list of words that we should *ignore* when processing documents, since they give no useful information about content.
  - English stop words: 174 words in R
  - Own defined stop words

```
> stopwords("english") # show stopwords
  [1] "i"          "me"         "my"         "myself"     "we"
  [9] "you"        "your"       "yours"      "yourself"   "yourselves"
 [17] "himself"    "she"        "her"        "hers"       "herself"
 [25] "they"       "them"       "their"      "theirs"     "themselves"
 [33] "whom"       "this"       "that"       "these"      "those"
```

```
> selfstopwords <- c("app","hbo","now") #self-defined stopwords
> corpus <- tm_map(corpus, removeWords, c(stopwords("english"),selfstopwords))
```

Stemming

- Inflectional Stemming
  - Remove plurals
  - Normalize verb tenses
  - Remove other affixes
  - Examples:
    - "walking", "walks", "walked", "walker" ➔ "walk"

```
> corpus <- tm_map(corpus,PlainTextDocument)
> corpus <-tm_map(corpus, stemDocument)
```

- Stemming to root
  - Reduce word to most basic element
  - More aggressive than inflectional
  - Examples
    - "denormalization" ➔ "norm"
    - "apply", "applications", "reapplied" ➔ "apply"