MGS 3701 / MKT 3950: Data Mining

Chapter20: Text Mining

Chungil Chae

温州肯恩大学
WENZHOU-KEAN UNIVERSITY

2022/01/01 (updated: 2022-05-17)

What exactly is text mining?

- Classify (label) thousands of documents
  - Extension of predictive modeling (our focus)
  - Extract meaning from a single document – interpreting it like a human reads language
  - “Natural language processing” (very ambitious, not predictive modeling, not our focus)

Classification (labeling) and clustering

- No attempt to extract overall document meaning from a single document
- Focus is on assigning a label or class to numerous documents
- As with numerical data mining, the goal is to do better than guessing

“Bag-of-words”

- Grammar, syntax, punctuation, word order are ignored
- The document is considered as a “bag of words”
- This approach is, nonetheless, effective when the goal is to decide which category or cluster a document falls in
- A typical application is supervised learning Requires lots of documents (a corpus)*
- Do not need 100% accuracy

> **“Corpus”** often refers to a fixed standard set of documents that many researchers can use to develop and tune text mining algorithms.

The spreadsheet model of text

- Columns are terms
- Rows are documents
- Cells indicate presence/absence (or frequency) of terms in documents
- Consider the two sentences:
  - S1 First we consider the spreadsheet model
  - S2 Then we consider another model

- S1. this is the first sentence.
- S2. this is a second sentence.
- S3. the third sentence is here.

```
## <<DocumentTermMatrix (documents: 3, terms: 7)>>
## Non-/sparse entries: 11/10
## Sparsity           : 48%
## Maximal term length: 8
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs first here second sentence the third this
##    1     1    0      0        1   1     0    1
##    2     0    0      1        1   0     0    1
##    3     0    1      0        1   1     1    0
```

Need to turn text into a matrix

- For the two documents (sentences S1 and S2) that we looked at earlier, the process of producing a matrix is simple
  - Word
  - Spaces
  - Periods
- Each word is preceded or followed by a space or period – a delimiter.
- Real text is more complicated

Lots of things to process besides words...

- Numbers (including dates, percents, monetary amounts), e.g. from Google Annual Report 2014:

We considered the historical trends in currency exchange rates and determined that it was reasonably possible that changes in exchange rates of 20% could be experienced in the near term. If the U.S. dollar weakened by 20% at December 31, 2013 and 2014, the amount recorded in AOCI related to our foreign exchange options before tax effect would have been approximately 4 million and 686 million lower at December 31, 2013 and December 31, 2014, and the total amount of expense recorded as interest and other income, net, would have been approximately 123 million and 90 million higher in the years ended December 31, 2013 and December 31, 2014. If the U.S. dollar strengthened by 20% at December 31, 20013 and December 31, 2014, the amount recorded in accumulated AOCI related to our foreign exchange options before tax effect would have been approximately 1.7 billion and 2.5 billion higher at December 31, 2013 and December 31, 2014, and the total amount of expense recorded as interest and other income, net, would have been approximately 120 million and 164 million higher in the years ended December 31, 2013 and December 31, 2014.

- Email addresses, url's, stray characters introduced by file conversions, ...

Sender: Distribution list for statistical items of interest From: "Massimini, Vince" svm@mitre.org Subject: Comparing the Maximal Procedure to Permuted Blocks Randomization To: Precedence: list List-Help: LISTSERV@LISTS.MITRE.ORG?body=INFO%20WSS-ELECTRONIC-MAIL-LIST For more WSS events, see washstat.org WSS Public Health/Biostatistics Section and NCI Division of Cancer Preventi= on Jointly Sponsored Event: =20 SPEAKER: Vance W. Berger, PhD National Cancer Institute and University of = Maryland Baltimore County and Klejda Bejleri, BS Biometry and Statistics, D= epartment of Biological Statistics and Computational Biology, Cornell Unive= rsity, Ithaca, NY 14853 =20 TITLE: Comparing the Maximal Procedure to Permuted Blocks Randomization TIME AND PLACE: Monday, June 8th NCI Shady Grove, 9609 Medical Center Drive= , Rockville MD Room 5E30/32. Bring photo ID, allow time to get through secu= rity

- Proper nouns & terms specific to a particular field

From Techsmith corporate information: All-In-One Capture, Camtasia, Camtasia Studio, Camtasia Relay, Coach's Eye, DubIt, EnSharpen, Enterprise Wide, Expressshow, Jing, Morae, Rich Recording Technology (RRT), Snagit, Screencast.com, ScreenChomp, Show The World, SmartFocus, TechSmith, TechSmith and T Design logo, TechSmith Fuse, TechSmith Relay, TSCC, and UserVue are marks or registered marks of TechSmith Corporation. From medical journal: Eight hundred elderly women and men from the population- based Framingham Osteoporosis Study had BMD assessed in 1988-1989 and again in 1992-1993. BMD was measured at femoral neck, trochanter, Ward's area, radial shaft, ultradistal radius, and lumbar spine using Lunar densitometers. (Risk Factors for Longitudinal Bone Loss in Elderly Men and Women: The Framingham Osteoporosis Study, Journal of Bone and Mineral Research Volume 15, Issue 4, pages 710–720, April 2000)

Tokenization

- We need to move from a mass of text to useful predictor information
- The first step is to separate out and identify individual terms
- The process by which you identify delimiters and use them to separate terms is called tokenization.
- The resulting terms are also called tokens.

Preprocessing

- Goal: reduction of text (also called vocabulary reduction) without losing meaning or predictive power
- Stemming
  - Reducing multiple variants of a word to a common core
  - Travel, traveling, traveled, etc. -> travel
- Ignore case
- Frequency filters can eliminate terms that
  - Appear in nearly all documents
  - Appear in hardly any documents
- Punctuation characters and extra white space can be removed, and treated as delimiters
- Remove terms that are on a stoplist (stopwords)
  - Typically is done to reduce size and noise by getting rid of very common terms
  - Illustrated with the default “english” stopword list that comes with R’s tm
- Frequency vs. presence/absence
- Normalization, when the presence of a type of term might be important but we don’t need the specific term. For example:
  - Replace john@domain.com with “email token”
  - Replace www.domain.com with “url token”

Post-reduction matrix

- Columns are documents, rows are terms
- Options for cell entries: 0/1 (presence absence)
  - Frequency count
  - TF-IDF (term frequency – inverse document frequency)
- TF = frequency of term
- IDF = log of inverse of the frequency with which documents have that term
- There are varying definitions of both TF and IDF, hence of TF-IDF
  - Bottom line:
    - TF-IDF is high where a rare term is present or frequent in a document
    - TF-IDF is near zero where a term is absent from a document, or abundant across all documents

Using R's tm package with simple example

S1. this is the first sentence. S2. this is a second sentence. S3. the third sentence is here.

enter the text as a corpus:

```
text ← c("this is the first sentence",
     "this is a second sentence",
     "the third sentence is here")
corp ← Corpus(VectorSource(text))
```

Producing the term-document matrix


```
tdm ← TermDocumentMatrix(corp)
inspect(tdm)
```

```
## <<TermDocumentMatrix (terms: 7, documents: 3)>>
## Non-/sparse entries: 11/10
## Sparsity           : 48%
## Maximal term length: 8
## Weighting          : term frequency (tf)
## Sample             :
##           Docs
## Terms      1 2 3
##   first    1 0 0
##   here     0 0 1
##   second   0 1 0
##   sentence 1 1 1
##   the      1 0 1
##   third    0 0 1
##   this     1 1 0
```

From terms to concepts – Latent Semantic Indexing

- The post-reduction term/document matrix is often still huge – too big for easy processing
- Recall how, with principal components, we derived a small set of synthetic predictor variables, each of which was a linear combination of “like-minded” original variables.
- Latent semantic indexing does something similar for text – it maps multiple terms to a small set of concepts.

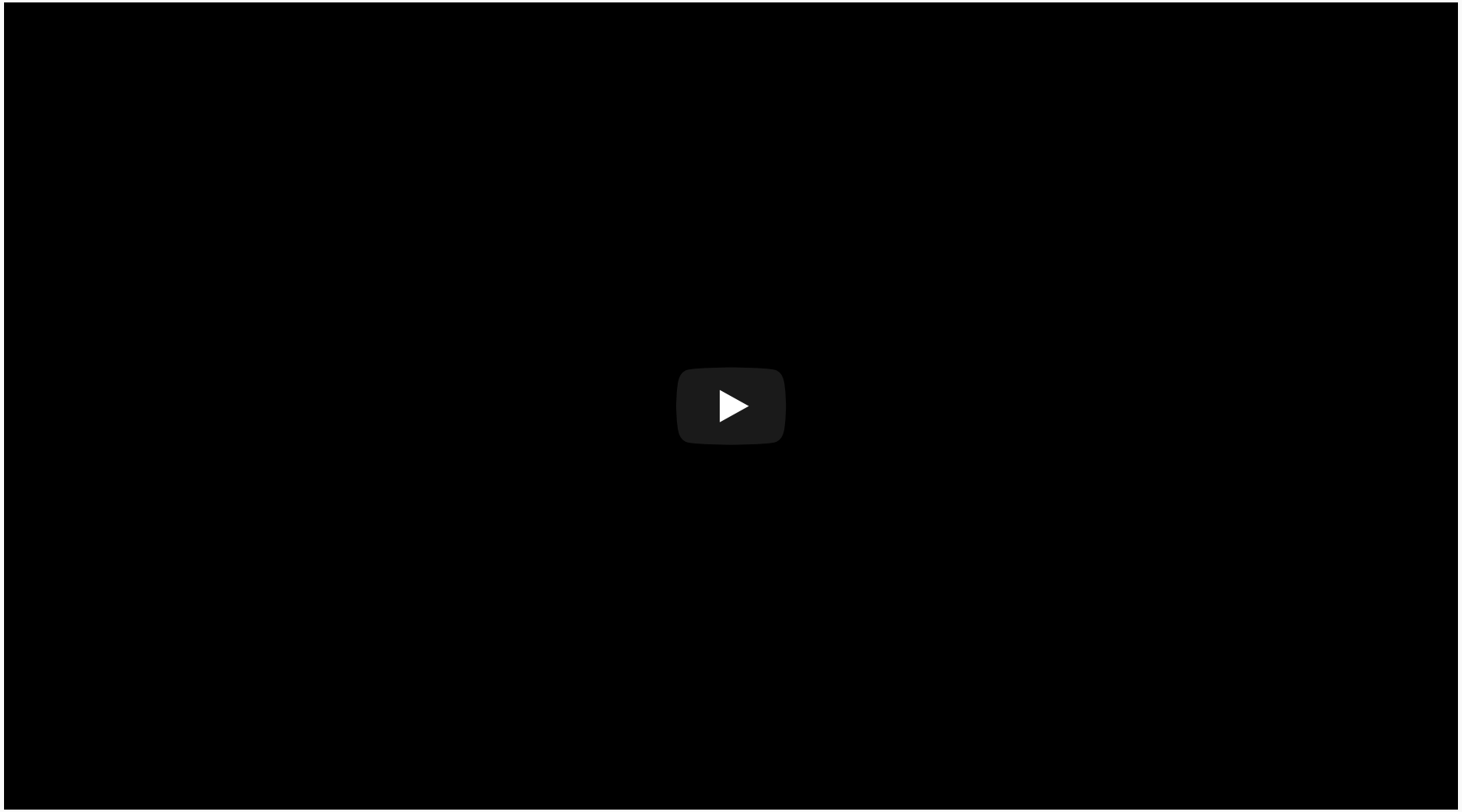
Extracting Meaning?

- It may be possible to use the concepts to identify themes in the document corpus, and clusters of documents sharing those themes.
- Often, however, the concepts do not map in obvious fashion to meaningful themes.
- Their key contribution is simply reducing the vocabulary – instead of a matrix with thousands of columns, we can deal with just a dozen or two.

A predictive model

- Now we have a clean, structured dataset similar to what we have used in our numerical data mining:
  - Class identifications (labels) for training
  - Numerical predictors

Chapter Video



References

Shmueli, G., P. C. Bruce, P. Gedeck, and N. R. Patel (2019). Data mining for business analytics: concepts, techniques and applications in Python. John Wiley & Sons.

Shmueli, G., P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

Assignment

Class Plan & Final Exam

- Class plan
  - May 10 - 13: Chapter6 - Regression
  - May 16 - 20: Chapter20 - Text Mining
- Range: Chapter 5, 6, 20
- May 24 (MKT)
  - (MKT Grouop1), 5:30 - 6:10
  - (MKT Grouop2), 6:15 - 6:45
- May 25 (MGS)
  - (MGS Grouop1), 2:30 - 3:10
  - (MGS Grouop2), 3:15 - 4:45
- 1 learning sheet (2 page, front and back, hand-writing only)
- No class on 26, 27, 31 (Final week)

Final Project (Project Proposal)

- Due date: May 27 (Friday 23:55pm, Beijing Time)
- To do:
  - Individual Assignment
  - 5 page (up to 7 page, no more than 7 page) project plan
  - Double space, 11 point, Times new roman
  - No reference page
  - Project proposal should includes your project plan guide by data mining process (chapter2)
  - PDF only (both R markdown generated or MS WORD)
  - Including your name and student ID and title
  - Submit to LMS (Submit button will be disappear 5 mins before midnight)