TEXT MINING

Peter C. Bruce

Co-author: *Data Mining for Business Intelligence* (2nd ed., 2010)

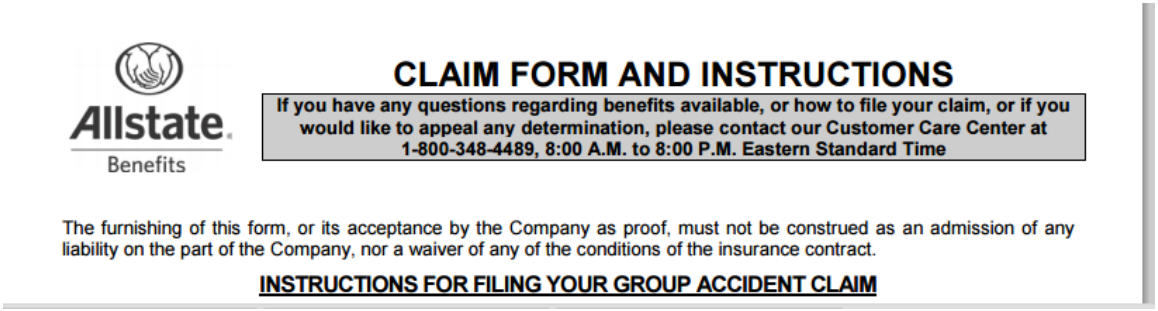
Author: *Introductory Statistics and Analytics* (2014)

Text data

- Up to now we have been dealing with structured quantitative data
  - Numerical
  - Binary (yes/no)
  - Multicategory
- Now we turn to unstructured text

Applications of Text Mining

- Insurance fraud – notes in claim forms can be mined and transformed into predictor variables for a predictive model
- The model is trained on prior claims in two classes – found to be fraudulent, and not found to be fraudulent
- The model is then applied to new claims



Allstate
Benefits

CLAIM FORM AND INSTRUCTIONS

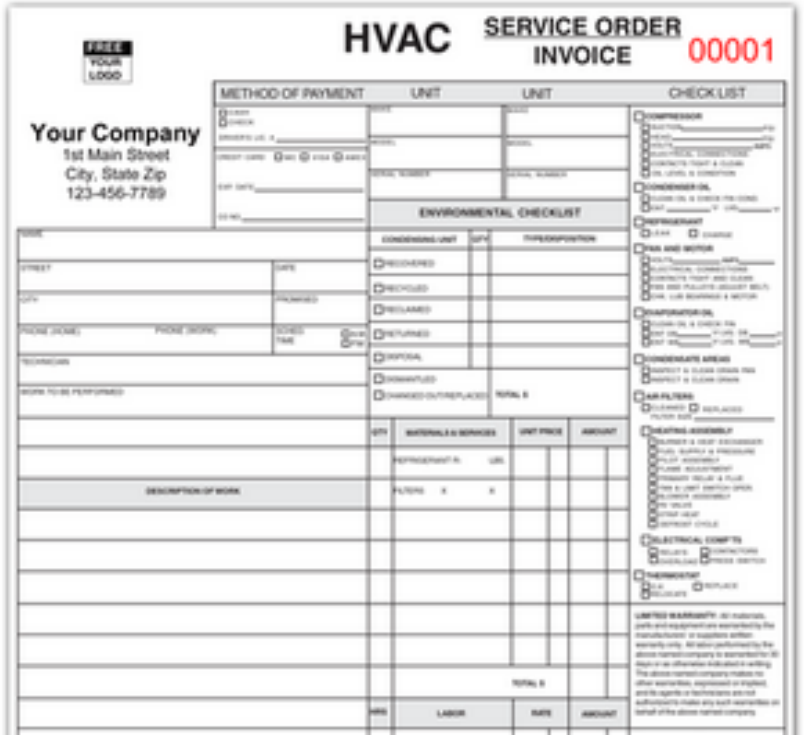
If you have any questions regarding benefits available, or how to file your claim, or if you would like to appeal any determination, please contact our Customer Care Center at 1-800-348-4489, 8:00 A.M. to 8:00 P.M. Eastern Standard Time

The furnishing of this form, or its acceptance by the Company as proof, must not be construed as an admission of any liability on the part of the Company, nor a waiver of any of the conditions of the insurance contract.

INSTRUCTIONS FOR FILING YOUR GROUP ACCIDENT CLAIM

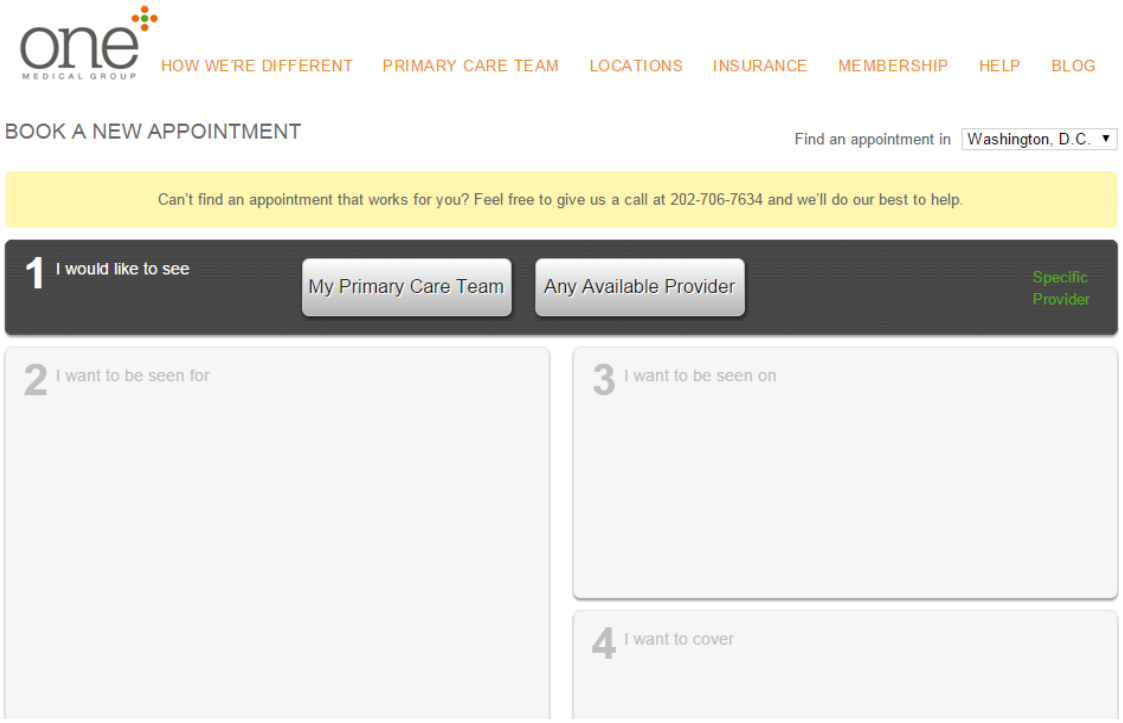
Applications, cont.

- Maintenance or support tickets often contain text fields
- These fields could be mined to classify ticket in several ways:
  - How urgent?
  - How much time to fix?
  - What category of technician is needed to fix?



Applications, cont.

- Medical triage/diagnosis
- Clinics could use patient online appointment request forms to route requests
  - Admin asst.
  - Nurse
  - Doctor



What exactly is text mining?

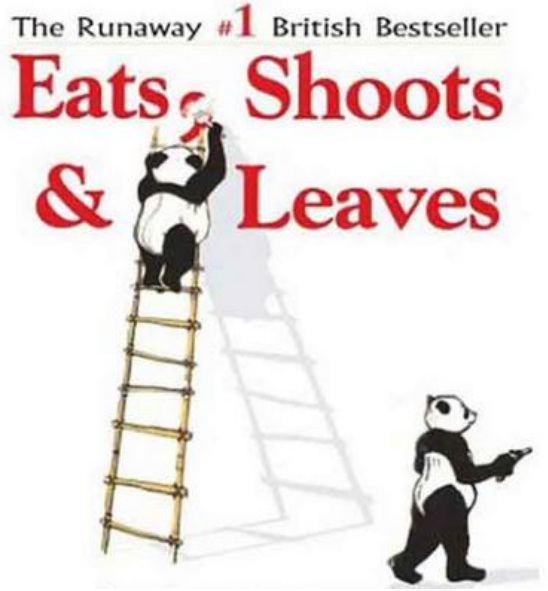
- Extract meaning from a single document – interpreting it like a human reads language
- Extract meaning to describe a set of documents
- Classify (label) thousands of documents

Natural Language Processing (NLP)

- Operates at the level of a single document
- It must master the complex “algorithms” of human language that have evolved over thousands of years
- Grammar, spelling, syntax, punctuation – they all matter
- Getting the meaning of a single document is a major challenge

Example: Adding a comma completely changes the meaning

- Consider the phrase “Eats, shoots and leaves”
- Is it a café robbery (with comma)?
- Or a bear’s eating habits (no comma)?



Or consider this sentence:

“Hitchcock shot The Birds in Bodega Bay”

- To someone indifferent to capitalization, and unfamiliar with movies, it might be about duck hunting.

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

Here is the resulting spreadsheet, using presence/absence:

	first	we	consider	the	spreadsheet	model	then	another
S1	1	1	1	1	1	1	0	0
S2	0	1	1	0	0	1	1	0

Exercise

Create a similar spreadsheet for the following two sentences:

S1: My job interview is Thursday

S2: My boss canceled my Thursday job interview

Create one spreadsheet where the cell entries are binary, indicating presence/absence, and another spreadsheet where the cell entries indicate frequency.