Introduction to Text Mining

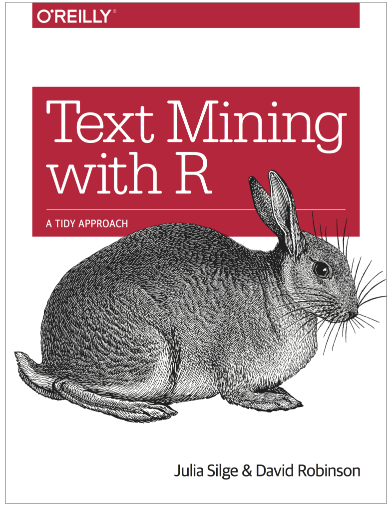
Rei Sanchez-Arias, Ph.D.

Text mining: motivation, applications, and classical concepts

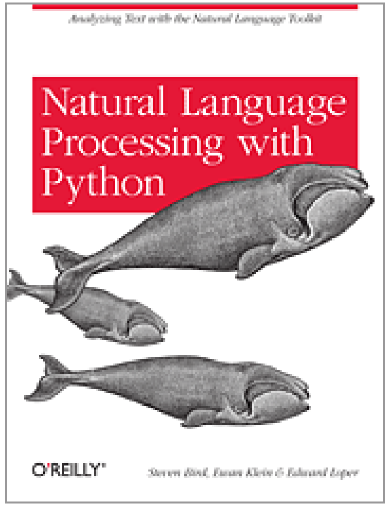
Some Resources and motivation

Some books to check

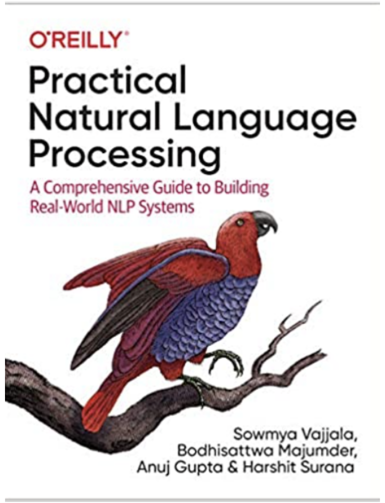
"Text Mining with R: A Tidy Approach" by Julia Silge and David Robinson



"Applied Text Analysis with Python" by Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda



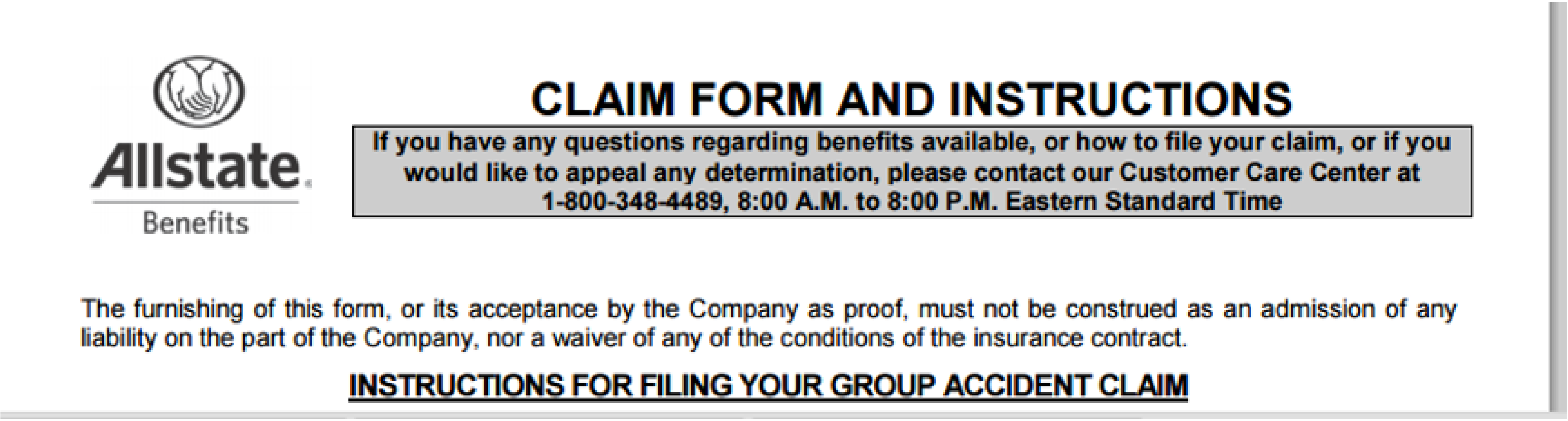
"Practical Natural Language Processing" by Vajjala, Majumder, Gupta, Surana



> Examples and materials in this set of slides are adapted from the books mentioned above.

Some applications of text mining (1)

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

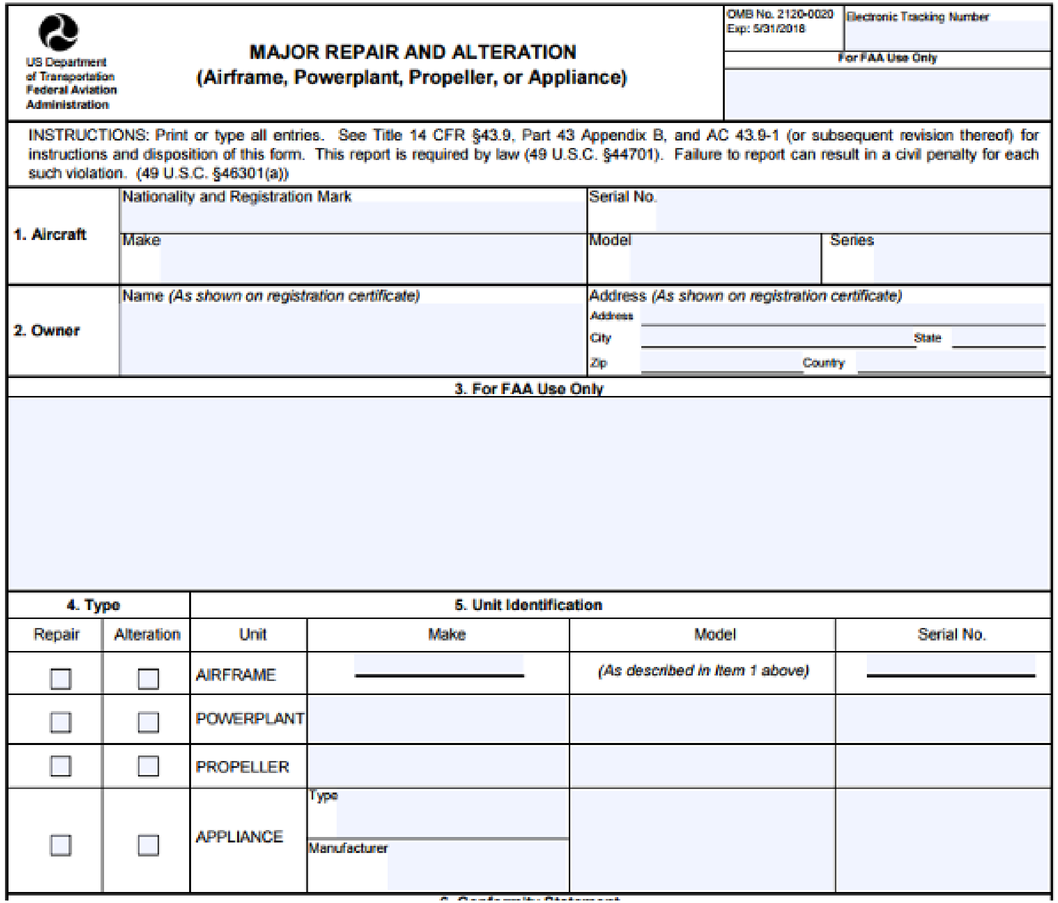
The furnishing of this form, or its acceptance by the Company as proof, must not be construed as an admission of any liability on the part of the Company, nor a waiver of any of the conditions of the insurance contract.

INSTRUCTIONS FOR FILING YOUR GROUP ACCIDENT CLAIM

Some applications of text mining (2)

- 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?

See for example: De Weerdt J, Vanden Broucke S, Vanthienen J, Baesens B (2012) Leveraging process discovery with trace clustering and text mining for intelligent analysis of incident management processes. In: Congress on evolutionary computation. IEEE, pp 1–8.


US Department of Transportation Federal Aviation Administration

MAJOR REPAIR AND ALTERATION
(Airframe, Powerplant, Propeller, or Appliance)

OMB No. 2120-0020 Exp: 5/31/2018 | Electronic Tracking Number | For FAA Use Only

INSTRUCTIONS: Print or type all entries. See Title 14 CFR §43.9, Part 43 Appendix B, and AC 43.9-1 (or subsequent revision thereof) for instructions and disposition of this form. This report is required by law (49 U.S.C. §44701). Failure to report can result in a civil penalty for each such violation. (49 U.S.C. §46301(a))

1. Aircraft	Nationality and Registration Mark	Serial No.	
	Make	Model	Series
2. Owner	Name (*As shown on registration certificate*)	Address (*As shown on registration certificate*) Address, City, State, Zip, Country	

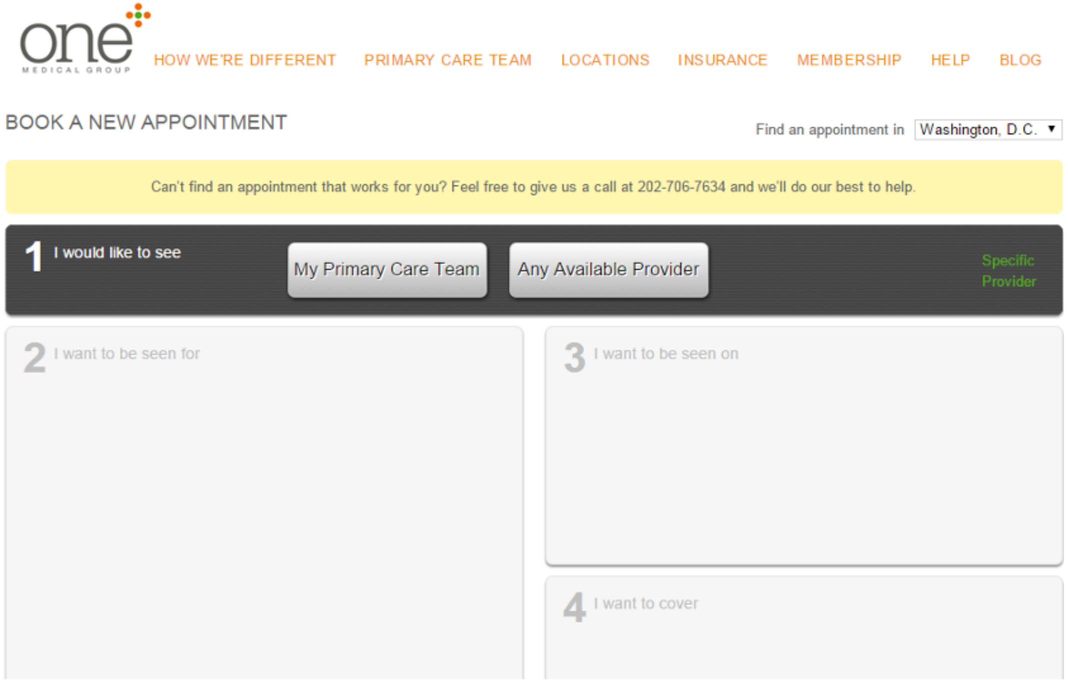
3. For FAA Use Only

4. Type		5. Unit Identification			
Repair	Alteration	Unit	Make	Model	Serial No.
☐	☐	AIRFRAME		(*As described in Item 1 above*)	
☐	☐	POWERPLANT			
☐	☐	PROPELLER			
☐	☐	APPLIANCE	Type / Manufacturer		

Some applications of text mining (3)

- Medical triage/diagnosis
- Clinics could use patient online appointment request forms to route requests
  - Administrative assistance
  - Nurse
  - Doctor

See for example: V. S. Pendyala and S. Figueira, "Automated Medical Diagnosis from Clinical Data," 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), San Francisco, CA, 2017, pp. 185-190.



Some approaches to text analytics and classical concepts

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, that is a **corpus** (often refers to a fixed standard set of documents that many researchers can use to develop and tune text mining algorithms).
- May not need 100% accuracy

A “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	1

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. We had: words, spaces, periods
- Each word is preceded or followed by a space or period – a **delimiter**.
- Real text is evidently more complicated
- There a **lots of things to process besides words**:
    - numbers (including dates symbols, monetary amounts)
    - email addresses, URLs, stray characters introduced by file conversions
    - proper nouns and terms specific to a particular field

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**.

Pre-processing

One of the goals is to reduce text without losing meaning or predictive power

- **Stemming:** reducing multiple variants of a word to a **common core**. For example: switching *travel*ing, *travel*ed to *travel*
- Ignore case

Frequency filters can eliminate terms that may appear in nearly all documents, or 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
- 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 = logarithm of inverse of the frequency with which documents have that term

```
text <- c("this is the first     sentence!!",
          "this is a second Sentence :)",
          "the third sentence, is here",
          "forth of all sentences")
```

```
              Docs
Terms          1 2 3 4
  all          0 0 0 1
  first        1 0 0 0
  forth        0 0 0 1
  here         0 0 1 0
  second       0 1 0 0
  sentence     0 1 0 0
  sentence!!   1 0 0 0
  sentence,    0 0 1 0
  sentences    0 0 0 1
  the          1 0 1 0
  third        0 0 1 0
  this         1 1 0 0
```

TF-IDF

For a given document d and term t, the **term frequency** is the number of times term t appears in document d :

$$TF(d,t) = \#\text{ times } t \text{ appears in document } d$$

To account for terms that appear frequently in the domain of interest, we compute the **Inverse Document Frequency** of term t, calculated over the entire corpus and defined as

$$IDF(t) = \ln\left(\frac{\text{total number of documents}}{\#\text{ documents containing term } t}\right)$$

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

TF-IDF (cont.)

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) = TF(t,d) \times \ln\left(\frac{n_{\text{docs}}}{n_{\text{docs}} \text{ containing term } t}\right)$$

- If term t appears in few documents then $IDF(t,d)$ increases

TF-IDF measures how *important* is a word in a collection of documents

- TF-IDF is large when a rare terms is frequent in a document.
- TF-IDF is close to zero when term is absent from documents, or term is abundant across all documents.