Text Analytics

Robert Stine
Dept of Statistics, Wharton
University of Pennsylvania

Preliminaries

Personal Perspective

Why look at text?

Why are you interested in text analytics?

- Interesting
  - How do they score the written SAT? Diagnose autism?
  - What gives away how a justice on the Supreme Court will vote?
- Prevalent
- Augments classical data in many situations
  - "How can I use these written comments?"

Examples

- Medical data combine lab measurements with clinical evaluations
- Open-ended survey responses (e.g., ANES)
- Scoring written applications
- Ad click prediction based on search text

Statistician's Perspective

Requires large models

Text frequently produces large numbers of predictive features, on the order of thousands to millions

a bit like genetics

High signal to noise ratio

Overwhelms classical statistical methods

- Overfitting
- Diffuse signal spread over many features

Challenge for modern analytic techniques

Recent innovations are designed for "nearly black" models, finding "needle in the haystack" effects from large collection

Text is often "nearly white", with effects spread widely over a large number of features

Biases…

Statistical modeling vs linguistic analysis

I am a statistician

Much more familiar with statistical modeling

Relatively less familiar with linguistic analysis

Convert text into the type of numerical data suited for models

Prediction versus interpretation

Science: Need to be able to reject claims

Interpretation: Beauty in the eye of the beholder

> Tufte: No limits to ability of human mind to explain what has been seen.

Evaluate models with objective measure of performance

Resulting Course

Blend

Fundamentals of NLP: grammar, vocabularies, etc

Techniques that accommodate large amounts of data

Models that mix text with other data

Software that can manipulate text and support analysis

data mining course

Software

R allows you do “do it yourself” or with built-in packages

regular expression text manipulation, singular value decompositions

R includes recent innovations in statistical modeling

lasso, gradient boosting

Materials: www-stat.wharton.upenn.edu/~stine

Course Overview

Syllabus

Its only a plan and likely to change

Key references

Plan

Monday	Foundations of NLP, preparing text data
Tuesday	Sentiment analysis, regression models
Wednesday	Classifiers with text, other modeling techniques
Thursday	Vector space models (embedding)
Friday	Hierarchical models, Bayes, deep learning

Style

Morning lecture, afternoon devoted to hands-on computing

Hands-on examples

Show full analysis of data, from messy initial phases thru various analyses

Objective

After the course,

You feel “comfortable” working with your own text data.

Only going to happen if you

ask questions during lectures

and

engage data analysis

Other Topics in Text

Not doing everything!

- Emphasis on problems that resemble “regression”

Deep grammatical analysis

- Analysis of written style, evolution of style
- Language modeling
  - See some examples of this at end of course

Translation

- From one language to another
- Sequence to sequence modeling

Illustrative Data and Tasks

Wine tasting notes

Can you distinguish a red wine from a white wine using only a brief note that describes its taste and aroma?

classification

Can you recognize the variety of red wine?

Cabernet vs merlot vs pinot vs zinfandel

If not the variety, can you tell the price? Rating points? regression

Prices of homes

What does a written real estate listing reveal about the value of a home that's not conveyed in the usual spreadsheet facts?

Does the interpretation of the listing depend on the location?

Tasting Notes

Running example during lectures

Data

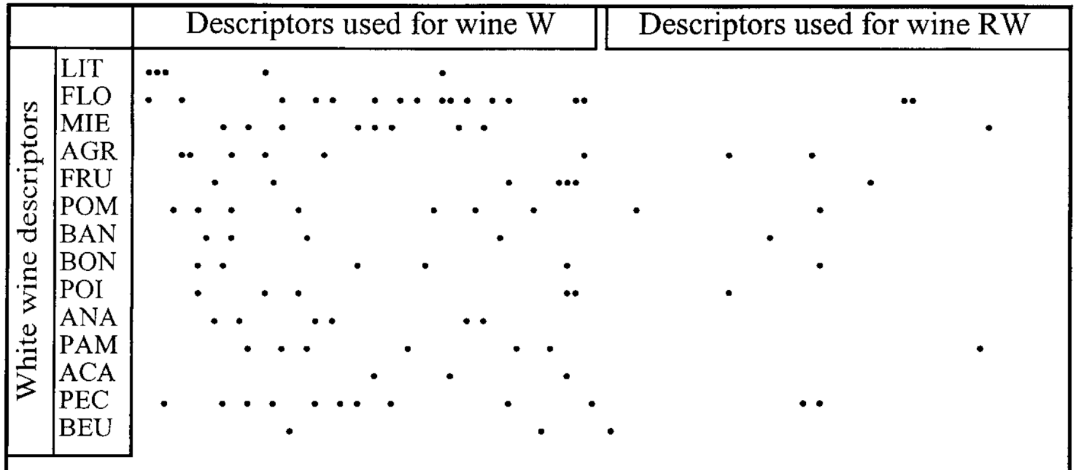
21,000 relatively short notes from Beverage Tasting Institute

> "Earthy, herbal, slightly herbaceous aromas. A medium-bodied palate leads to a short finish that is earthy, tart and has limited fruit."
>
> "Toasty oak, cherry and thyme aromas. A rich entry leads to a full-bodied palate and a well-structured finish with vibrant acidity, refined tannins, and lovely varietal fruit."

What's the color? The variety?

Mark Liberman http://languagelog.ldc.upenn.edu/nll/?p=3887/

Do people describe taste, or do they describe color?



"The color of odors"

Real-Estate Listings

Data

Various locations, written in a an idiosyncratic vernacular.

> "Built in 1893, this bright, spacious home on an oversized lot has east, west and south exposures and is located in the heart of Chicago's famed gold coast. Many beautiful vintage details. 7600 sqft on four levels with additional 1200sf partially finished basement."
>
> "Wow! 4 levels to this well maintained split on quiet st. 2 large eat in kitchens. lr dr and 3 bdrms up all with hardwood floors. Spacious family rm with4th bedroom and walk out to yd with kit."

Challenge

Language in Chicago is different from that in Miami.

Beach vs lake. Palm trees vs park side.

Text or quantitative data?

Other Available Data

Political speech, legal proceedings

- Transcripts from Congressional Record
- Speeches during election campaign
- Favorable testimony and questioning

Social media

- Facebook posts
- Twitter tweets

Internet commerce

- Product ratings (e.g., Amazon stars and Rotten Tomatoes)

Tweets

Not as easy to get as once available

Requires a registration process to obtain authorization keys

Limited number of tweets can be recovered each day

Example: Donald Trump

"Conservatinized: RT @soxsher: MT @SMolloyDVM: #DonaldTrump echoes #TedCruz on: immigration, amnesty, border, military, jobs. http://t.co/W8VN7ohWF2 #CruzCre…"

"luvmom8702: RT @AlterNet: Sorry, GOP! You Created #DonaldTrump, Now He's the Face of Your Party! #uniteblue http://t.co/b9Q9j1CLII http://t.co/TfFSa6ON…"

"FredOrth: RT @Politics_PR: Watch the documentary #DonaldTrump fought decades to kill: http://t.co/csTem3GYPQ #p2 #tcot http://t.co/IujzKMC5uK"

"mesquitepenny: RT @LetsTalkNevada: What's the backstory tying together #immigrants as criminals? A professor explains #history & debunks #donaldtrump . ht…"

Tools

Almost exclusively using R

- tm (text miner)
- Host of supporting packages (openNLP, stringr, topicmodels)

Facilitates using other most techniques

- Principal components
- Multinomial regression
- Regression and classification trees (boosting)

Alternative: NLTK and python

- But then you have to move to R for the analysis

Typical Steps

Prepare data

90% or more

Deciding on role for text

Representative sample… from what?

Editing: removing weird characters (eg html markup)

“Feature engineering” making variables for modeling

Modeling choices, issues

Unsupervised (clustering) vs supervised (regression)

Structural (prob model) vs predictive (conditional mean)

Challenge: richness of the feature space (over-fitting)

Validation and inference

Preparing Text

wine_data.R

can play along now, revisit in afternoon

Preparing Text

Big picture

Extracting text from other formats

- Embedded in HTML documents, XML files
- Mixing text with other information
- Objective: “one observation” per line

Approach

- Depends on nature of the analysis
- Dictionary methods: find embedded patterns/words in text
- Bag-of-words: convert text into dummy variables, counts
  - document-term matrix
- NLP: deeper linguistic analysis
  - identifying named entity, parts of speech, grammatical structure

language specific!

Wine Data

Many possible applications

- Classification and regression problems
  - Clustering
  - Identifying color or type of wine
  - Predicting price or rating of the wine

Initial raw data

- 9 lines per wine
- Lots of embedded goodies, such as vintage, vineyard, variety, price
- Read in 20888, then process
  - Beware encoding issues, such as
  - “smart quote” inserted by software

```
_________
163522
Corey Creek Vineyard 2000 Gewurztraminer, North Fork Long Island.
Brilliant yellow-straw hue. Lemon oil and grapefruit aromas follow
through on a medium-bodied palate with impressive wieght and a dry,
tart finish.
Gewurztraminer, White
12.0%
In Our Chicago Tasting Room
Mar-01-2002
84 points (Recommended)
_________
163528
Kunde 1999 Estate Bottled, Syrah, Sonoma Valley $23.
Brilliant ruby red hue. Bacon fat, black cherry, dill, oak aromas. A
rich entry leads to a moderately full-bodied palate with forward fruit
and a finish that offers sleek tannins and fine acidity. A more subtle
style of California Syrah.
Syrah, Red
13.0%
In Our Chicago Tasting Room
May-01-2002
90 points (Exceptional)
_________
163529
```

Data Frame

Store text as nine-column data frame

Extract and add embedded features to data frame

- Variety, color, points, price, vintage, alcohol
  - Check each using descriptive statistics to identify errors
- Remove part of the description: downstream impact

Regular expressions

- Express patterns
  - Introduction: “egrep for linguists”; R built-in docs

browse

- Iterative approach: start with candidate, try it, refine
  - Look out for serious outliers in this phase
- Price: “[$][[:digit:]]+([.][[:digit:]]+)*
- Vintage: "(1|2)(0|1|9)[0-9]{2}"

value of convenience functions in R

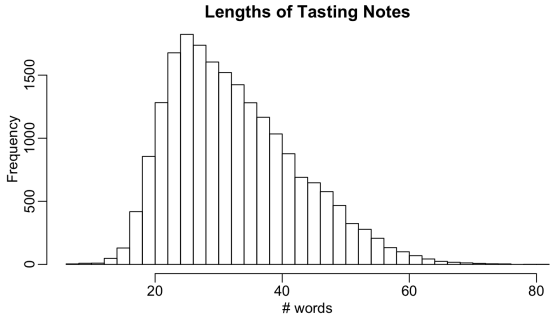
Expanded Data Frame

Data frame is a “list” that masquerades as a matrix

Can mix various types of data to describe each observation

Retains row/column style, but flexible about contents

Key properties of data



10,600 reds and 7,100 whites

About 33 words on average in a description

Bell-shaped distribution for points, skewed for price

differences by color, variety

Transitions in format of reviews: style changes over time

Missing data

602 missing alcohol, remove 379 with no description and 1 “outlier”

essential to recognize outliers in text

Save the data frame!

Document Term Matrix

Sparse matrix of counts of word types

term-document matrix

- Terminlogy: word type vs word token
- Rows are documents, columns are types
- Elements are counts of word tokens of column type in row doc

What's a "word"?

Tokenize documents

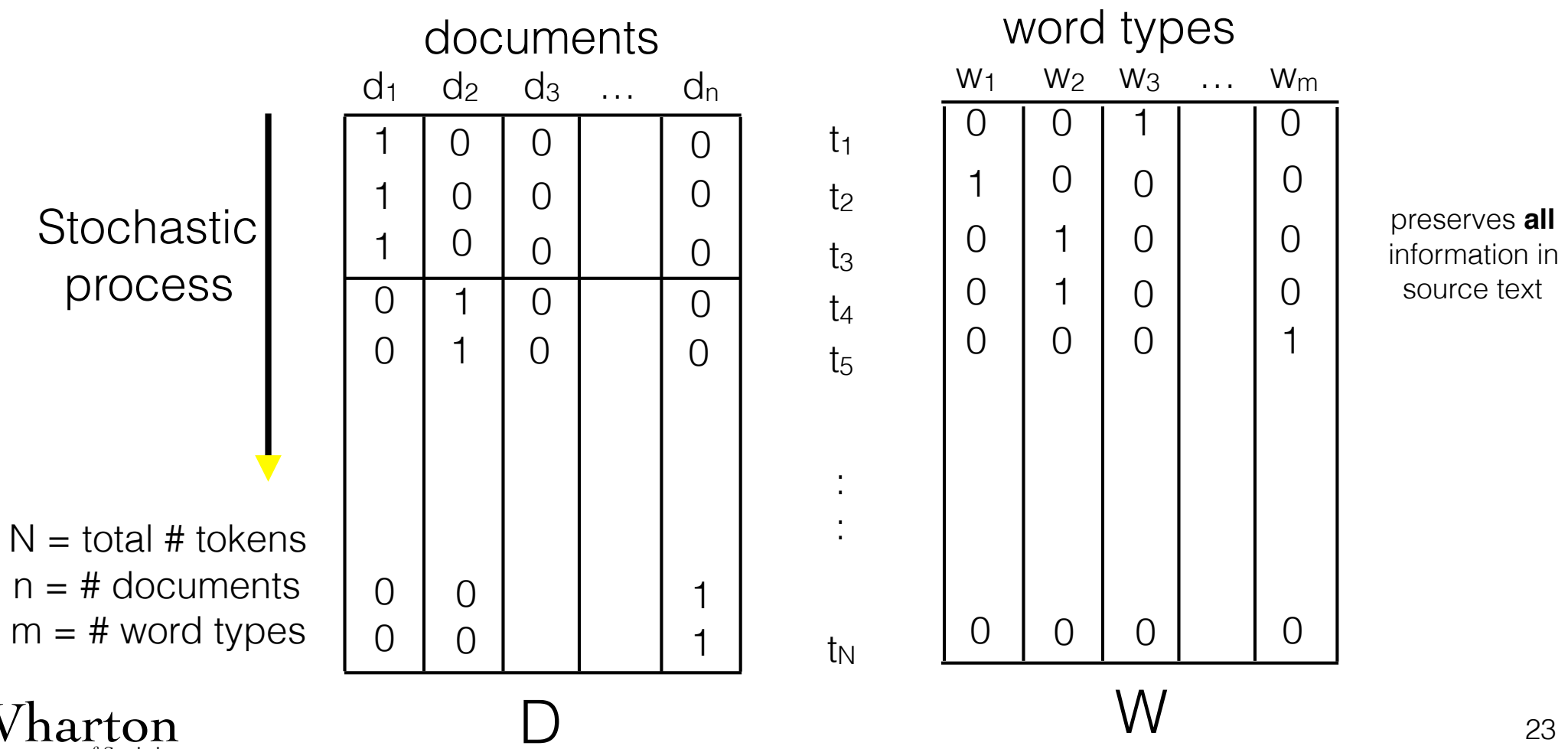
- Distinguish upper from lower case?
- Retain punctuation?
  - Might the presence of an exclamation point be informative?
- Stem words?
  - Distinguish oak from oaky?
  - See Mark Liberman's blog for more discussion: trend to process less

Token Space

Novel perspective on the document-term matrix

Consider two matrices with elements 0 and 1

Total number of rows = total number of word tokens



preserves **all** information in source text

DTM ≈ Covariance

View columns of D and W as indicator vars

Produced by hidden "document generating" stochastic process

Stationary process emits word tokens and document i

Hidden Markov model, changing state with documents.

Counts resemble covariances

Consider nxm matrix $C = D^T W$

Elements of C count the word types in each document

$$C_{ij} = \#\{w_j \text{ in } d_i\}$$

If types are rare, means ≈ 0 and

$$C_{ij} \approx N \, \mathrm{cov}(d_i, w_j)$$

Relevance?

Conceptual basis of vector space models for text.

Building DTM in R

Facilities in tm

Construct corpus from text

Allows using external collection of files for big task

Tokenize

Some steps reduce vocabulary

down case, remove stop words/punctuation, stem words

Others increase vocabulary

tag with parts of speech

Trend is to “do less”

Build document-term matrix

Held as a special “sparse matrix” object to reduce memory

Linguistic Tools

Natural language processing (NLP)

- Available via the NLP and openNLP packages in R
- Include built-in pre-compiled models

Capabilities

- Part-of-speech tagging (Penn Treebank symbols, next slide)
- Sentence chunking (subject-verb-object)
- Named entity recognition

Issues

- Language, corpus specific: standard English grammar, news
- Resource hungry: consider Python (NLTK from Stanford)

Penn Treebank POS Tags

Common symbols for tagged text

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Distribution of Types

Most word types are rare, most tokens are common

Total of 763,013 tokens from 5,486 word types

Zipf distribution for word types

Depends on how text was tokenized

Power law has ideal form…

Frequency of second most common 1/2 frequency of most common

Frequency of third most common 1/3 frequency of most common…

$$p_j = (1/j)\, p_1,\ j = 2,3,4\ldots$$

Highly skewed (plot follows)

Most common are what you'd expect

67587	53906	53299	30836	29046
"_,_"	"and"	"_._"	"_-_"	"with"

Why might you want to keep these words?

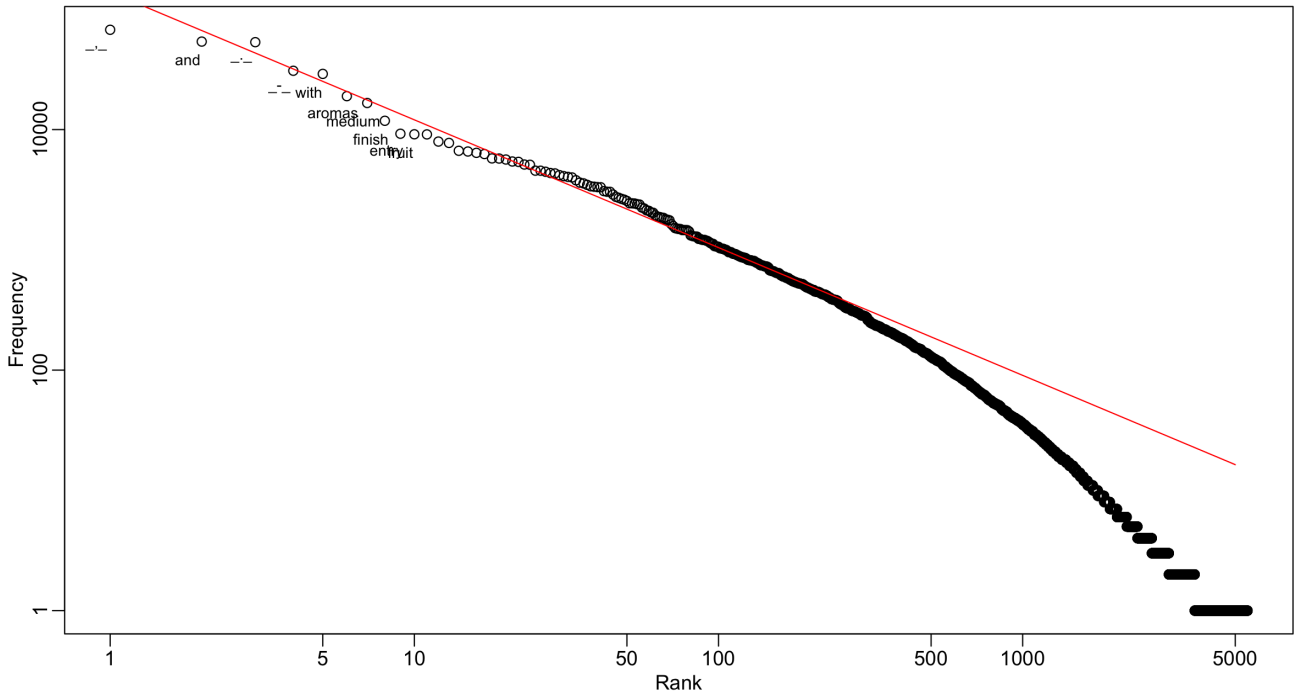
Distribution of Types

Plot log of frequency on log of rank

Sum columns of C, ordered by frequency

Power law would be a line

Most data have a concave shape



slope for first 250 is -1.06

Discussion of DTM

Sensitive to choices of analyst

How was the text tokenized?

Bag-of-words

bag: A collection that allows copies of elements.

A set is a special case of a bag that limits each count to 1.

Each row of C (one document) is a bag.

Sequence order is lost: Random permutations of the tokens produce the same document-term matrix.

Sparse representation is essential

C is 20508 x 5486, having 112,506,888 elements

Common vocabulary might have 50,000 word types

Handling Rare Types

What to do about rare word types?

1809/5486 ≈ 33% of word types appear just once!

Another 653 + 366 = 1019 appear just 2 or 3 times

Anticipate complication

Suppose we use word counts to predict price of wine

Split sample analysis: say, half for modeling, half for testing

Test sample guaranteed to have words we never saw in building our model and possibly omit words in model

Recode as out-of-vocabulary (OOV)

Just one symbol, or distinguish depending on use in context?

Rare Words

Possible ways to reduce number of OOVs

Stem the words: “cigars” found 1 time, “cigar” found 152

But does “fruit” == “fruity”?

Fix spelling errors: “berrry", “ciitrus”

Combine numbers as one type of OOV

Recoding as OOV

Have special OOV for numbers

Part of speech tagging
Special OOV for nouns vs verbs vs places vs things etc

Losing sight of forest for trees?

758,800 tokens seen more than 3 times

4,213 seen 3 or less

Hands-On

Preparing Text

Relevant Aspects of R

Data: Vectors, lists, strings, matrices, data frames

- Indexing

Functions

- First-class objects in the system
  - For example, a function can return a function as its result

Combining functions with data

- Mapping functions (aka, applying functions, eg sapply or tapply)

File input/output

Text resources

- https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

Regular Expressions

Work through “egrep for linguists”

Unix commands may be relevant to some

Patterns

Special characters such as

*	repetition
()	grouping
.	any character
\	special use for the next term
\w	word characters …

Logic expressions

No one I know remembers them for very long!

Examples of Regexs

Finding text near a comma

```
m <- regexpr(".*,", Wine[1:10,"type"])
```

Customized with tagged fields

```
str_replace(c(“xxx&a","bbb&b"),"&(.)","\\1")
```

Extract alternatives

```
str_extract(type, “(Red|White)”)
```

Extract patterns

```
str_extract(rating, “^[[:digit:]]+ “)
```

Exercises

Federalist papers

What distinguishes papers of Hamilton from Madison and Jay?

What about the papers of unknown providence?

Are the new Trump tweets observed in July 2016 different from those sampled in July 2015?

Extracting Data from XML

XML

- Extensible markup language
- Contents delimited by <tag> xxxxxx </tag> pairs
- Recursive

Main example: HTML format for web pages

Common problem

- Build data frame from XML
- Extract document, name variables based on the tags
- Example: Amazon product ratings data

    Blitzer, Dredze, and Pereira 2007