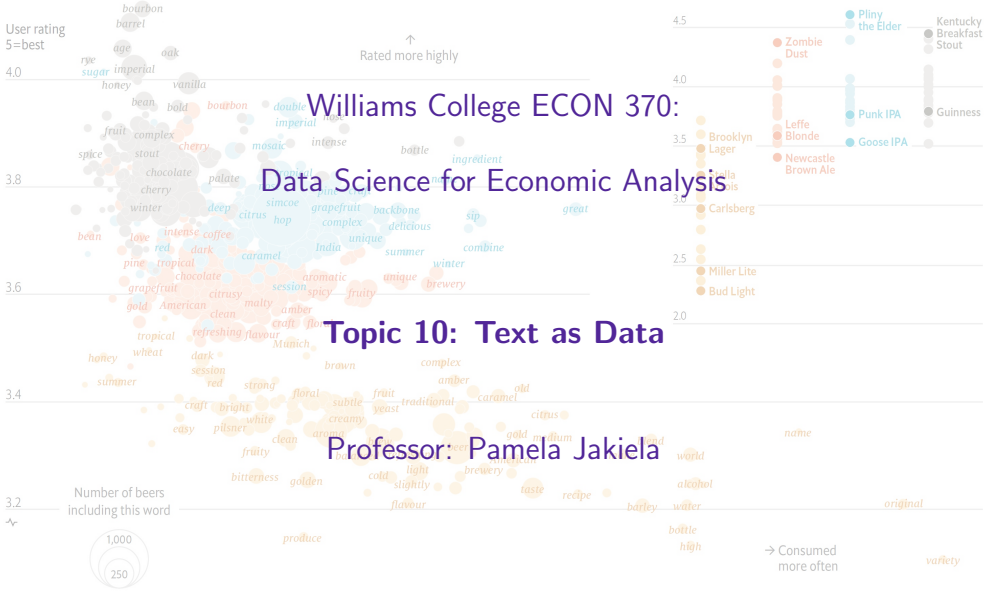


Williams College ECON 370:

Data Science for Economic Analysis

Topic 10: Text as Data

Professor: Pamela Jakiela

Reading in Text as Data

When we load text, it is often in a character vector with lines or paragraphs as observations

Example:

```
1 I'm alone, yeah, I don't know if I can face the night
2 I'm in tears and the cryin' that I do is for you
3 I want your love
4 Let's break the walls between us
5 Don't make it tough
6 I'll put away my pride
7 Enough's enough
8 I've suffered and I've seen the light
9 Baby, you're my angel
...
```

Data is not clean/tidy in the sense of being one observation per row – what is an observation?

Tokenization

To analyze text data, we typically break it into **tokens**, most often single words

	words
1	i'm
2	alone
3	yeah
4	i
5	don't
6	know
7	if
8	i

Tokenization

To analyze text data, we typically break it into **tokens**, most often single words or n-grams

words (1-grams)		bigrams (2-grams)		trigrams (3-grams)	
1	i'm	1	i'm alone	1	i'm alone yeah
2	alone	2	alone yeah	2	alone yeah i
3	yeah	3	yeah i	3	yeah i don't
4	i	4	i don't	4	i don't know
5	don't	5	don't know	5	don't know if
6	know	6	know if	6	know if i
7	if	7	if i	7	if i can
8	i	8	i can	8	i can face

A Simple Text as Data Pipeline

Term Frequencies

words in order			word counts		
1	i'm		1	and	17
2	alone		2	come	14
3	yeah		3	me	12
4	i		4	you're	11
5	don't	⇒	5	save	10
6	know		6	tonight	10
7	if		7	i	9
8	i		8	my	9
9	can		9	angel	8

Stop Words

Stop words are widely used words that are unlikely to distinguish a text/document

- Examples: a, an, the, and, but, or, of, to, from, by, is, are, was, be, he, she, it, they, them
  - Importantly, there are cases where stop words do convey important meaning (gender, author)
- We call them stop words because we (sometimes) filter them out (i.e. we stop them)

There are many lists of stop words, including (different) defaults in R and Python

- Choose a stop word list that makes sense in your specific context
- We often want to add task-specific stop words (e.g. economics, estimate, find, show)

Distinguishing Angel from Angel

Angel by Aerosmith	
and	17
come	14
me	12
you’re	11
save	10
tonight	10
i	9
my	9
angel	8
the	7

Angel by Shaggy	
my	37
you’re	30
angel	20
you	17
girl	14
me	11
when	11
darling	10
i	9
and	8

Distinguishing Angel from Angel: Removing Stop Words

Angel by Aerosmith	
come	0.062
save	0.044
tonight	0.044
angel	0.036
baby	0.022
alright	0.018
make	0.018
reason	0.018
yeah	0.018
love	0.013
without	0.013

Angel by Shaggy	
angel	0.045
girl	0.032
darling	0.023
shorty	0.016
baby	0.011
closer	0.011
friend	0.011
lady	0.011
need	0.011
peeps	0.011
said	0.011

Term Frequency – Inverse Document Frequency (TFIDF)

Some words appear a lot in all (or most) documents within a particular corpus

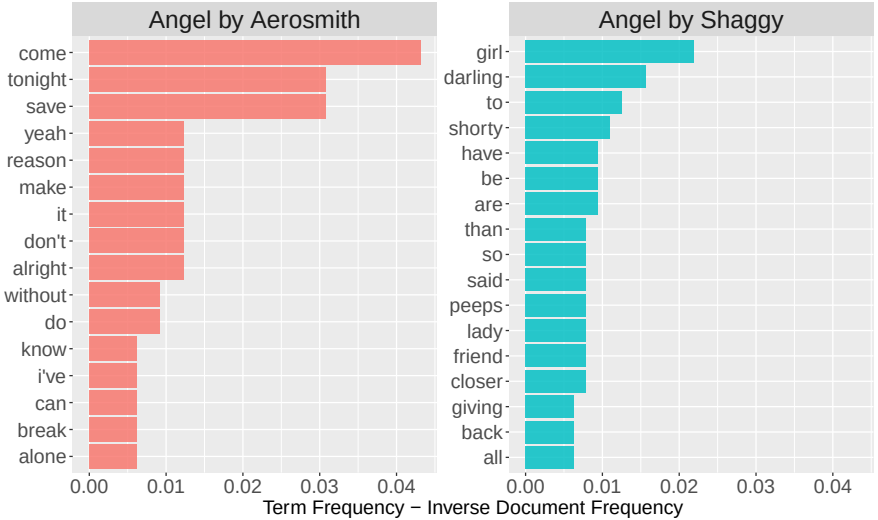
- Example: angel, baby (or, in other contexts, economics or find)
- Such words do not help us distinguish between documents in the corpus

Using **term frequency – inverse document frequency (tfidf)** we can identify words that allow us to better distinguish between documents (or distinct classes/groups of documents)

- **Term frequency (tf)** for word x in document i:
  number of times word x appears in document i / number of words in i in document i
- **Document frequency (df)** for word x:
  number of documents that contain x / number of documents in corpus
- **TFIDF** for word x in document i: term frequency $\times$ ln (1 / document frequency)

We say document frequency, but we might also think of class/group frequency

TFIDF: Aerosmith vs. Shaggy



The Purpose of Text as Data Analysis

Word-level analysis: dictionary methods

- Calculate document-level statistics based on word-level variables (e.g. sentiment)
  - Is Aerosmith's Angel more negative than Shaggy's Angel?
- Construct statistics capturing content of document (e.g. policy uncertainty)

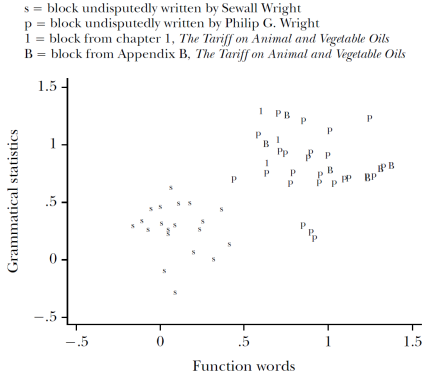
Construct document-level variables based on the frequency with which different words are used

- Can we predict whether a song is by Aerosmith or Shaggy based on the words used?
- Use word frequencies, etc., as variables or use dimension reduction (supervised learning)
- Identify underlying groups in the corpus (unsupervised learning)

Text Analysis and the Mystery of Instrumental Variables

Figure 1

Scatterplot of Predicted Values from Regression on First Four Principal Components: Grammatical Statistics versus Function Words



Source: Stock and Trebbi (2003)

Lab #10

Data set of $N = 219$ National Bureau of Economic Research (NBER) working papers

- From two mutually exclusive NBER programs: Asset Pricing and Children and Families
- Data set contains titles, authors, publication dates, abstracts, keywords

Objective: identify words that distinguish between papers in the two fields/programs