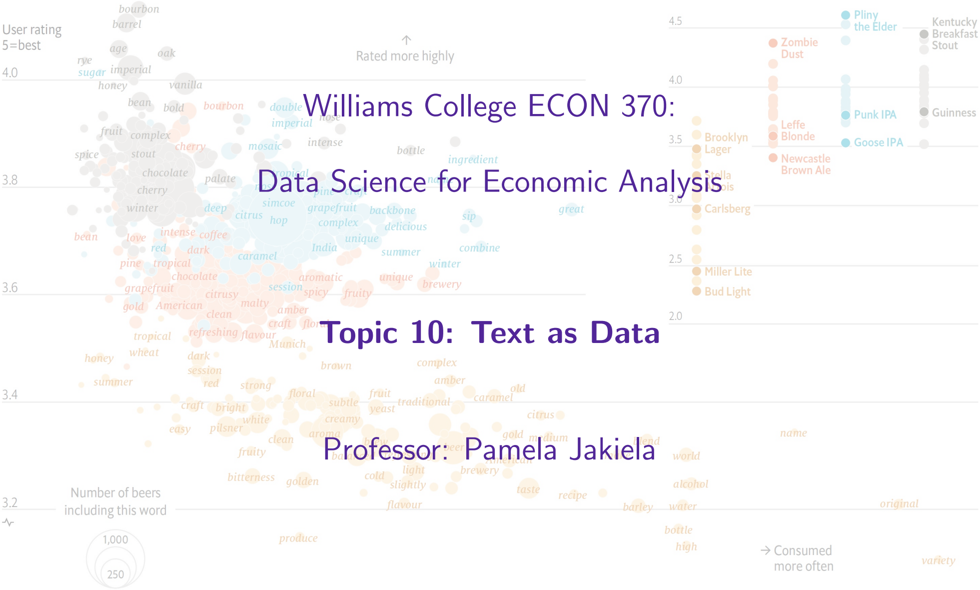



# Williams College ECON 370:

# Data Science for Economic Analysis

## Topic 10: Text as Data

Professor: Pamela Jakiela

# Outline

- Tokens
- Term frequencies
- TF-IDF
- Clustering documents
- Who invented instrumental variables?

# 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 | $\Rightarrow$ | 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

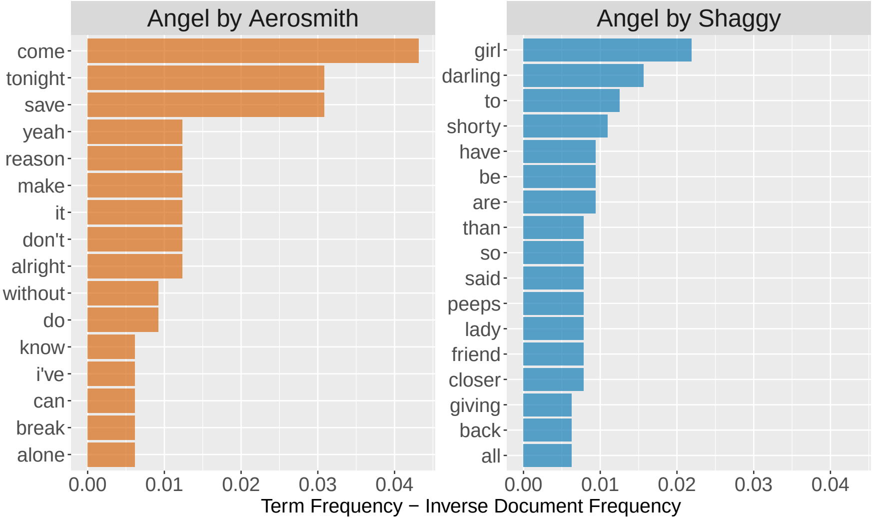
- Example: angel, baby (or, in other contexts, economics or Federal Reserve)
- Such words do not help us distinguish between documents in the corpus

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



# Creating a Document Term Matrix

Text as data methods typically treat the document as the unit of observation

- Each word (included in the analysis) is a variable, so the $X$ matrix may be large

A **document term matrix** is a data frame with documents as rows and words as columns

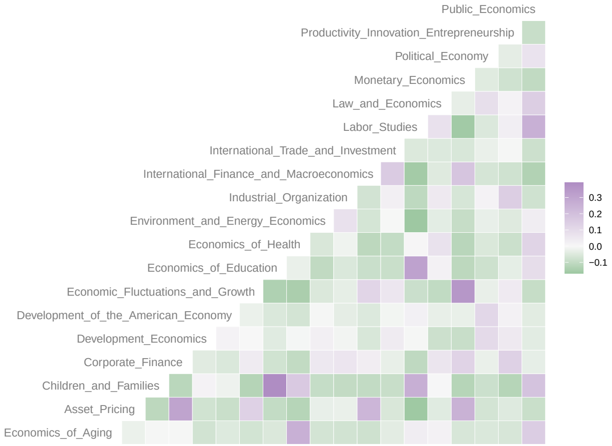
- Values of the word variables typically functions of within-document word counts

Example: a data set of book titles containing the word “sun”

```
["Evil Under the Sun", "Half of a Yellow Sun", "The Sun Also Rises"]
```

| | a | also | evil | half | of | rises | sun | the | under | yellow |
|---|---|---|---|---|---|---|---|---|---|---|
| Evil | 0 | 0 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 |
| Half | 0.2 | 0 | 0 | 0.2 | 0.2 | 0 | 0.2 | 0 | 0 | 0.2 |
| Rises | 0 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0 | 0 |

# Clustering Documents Example: NBER Working Papers




Data source: 1,112 NBER working papers released in 2024

# Example: Asset Pricing and Children's Program Working Papers

$N = 219$ papers from two mutually exclusive programs (Asset Pricing and Children & Families)

- Tokenize abstracts to compare content of the papers in the two programs

$k$-means clustering in R with 50 random starting allocations, no preprocessing

- Chosen partition consistent across seeds, suggesting global maximum
- Cluster sizes are reasonable, no $N = 1$ clusters

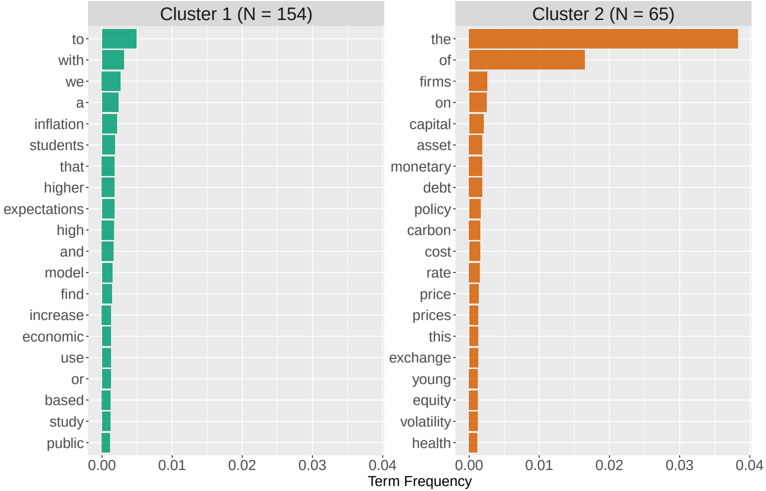
# Example: Breakdown of Clusters by NBER Program

| | Cluster Number | |
|---|---|---|
| | 1 | 2 |
| Asset Pricing | 77 | 39 |
| Children and Families | 77 | 26 |

Most common words in abstracts, by program:

- Asset Pricing: the, of, and, in, to, a, we, that, for, with
- Children: the, of, and, in, to, a, we, that, for, on

# Example: Words Most Associated with Each Cluster



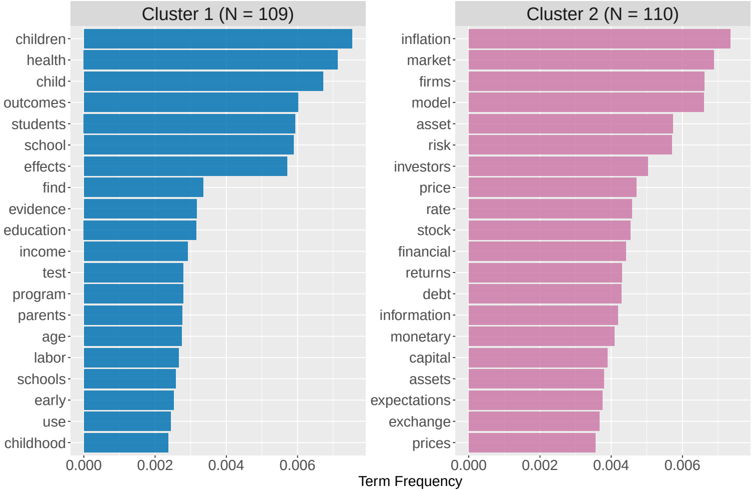
# After Removing Stop Words: Magic Happens

| | Cluster Number | |
|---|---|---|
| | 1 | 2 |
| Asset Pricing | 4 | 110 |
| Children and Families | 105 | 0 |

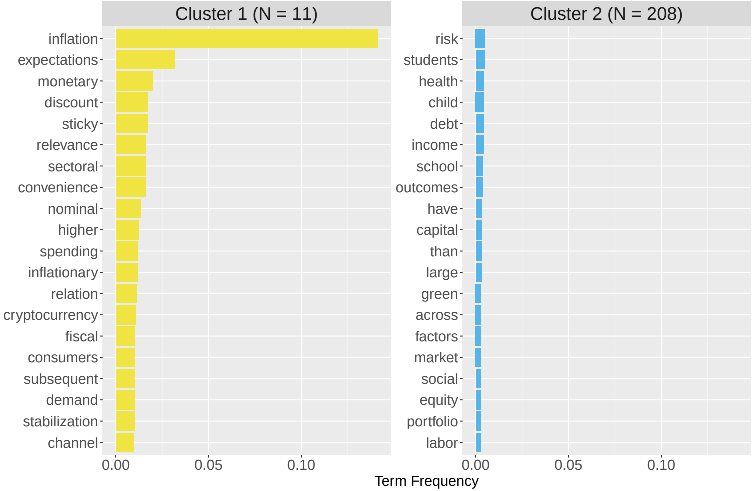
The four Asset Pricing papers included in Cluster 1:

- "On Robust Inference in Time Series Regression"
- "Who Benefits from Retirement Saving Incentives in the U.S.? Evidence on Gaps in Retirement Wealth Accumulation by Race and Parental Income" (also LS, PE)
- "How Do Income-Driven Repayment Plans Benefit Student Debt Borrowers?" (LS, PE)
- "AI and Finance"

# After Removing Stop Words: Magic Happens



# Clustering with TF-IDF



# Text Analysis and the Mystery of Instrumental Variables

*Journal of Economic Perspectives—Volume 17, Number 3—Summer 2003—Pages 177–194*

**Retrospectives**

Who Invented Instrumental Variable Regression?

James H. Stock and Francesco Trebbi

Source: Stock and Trebbi (2003)

# Who Invented Instrumental Variables?

First known discussion of instrumental variables appears in Philip Wright's book *The Tariff on Animal and Vegetable Oils*, published in 1928; IV introduced (out of the blue) in Appendix B

- Philip Wright was an economist and mathematician who liked poetry
- His son Sewall Wright was a statistician and professor at U Chicago
- Appendix B explains how an instrument for price can be used to map the supply/demand curve, which is equivalent to identifying the causal effect of price on supply/demand
- Approach relates to "method of path coefficients" used by Sewall

**Was Appendix B actually written by Sewall and not his father?**

- Generate a data set containing known samples of Philip and Sewall's writing

# Stopwords Can Be Used to Identify Authors

*Table 1*

**Function Words Used in the Stylometric Analysis**

| | | | | | | |
|---|---|---|---|---|---|---|
| a | all | also | an | and | any | are |
| as | at | be | been | but | by | can |
| do | down | even | every | for | from | had |
| has | have | her | his | if | in | into |
| is | it | its | may | more | must | my |
| no | not | now | of | on | one | only |
| or | our | shall | should | so | some | such |
| than | that | the | their | then | there | things[a] |
| this | to | up | upon | was | were | what |
| when | which | who | will | with | would | your |

*Notes:* These are the function words listed in Mosteller and Wallace (1963, Table 2.5).
[a] Dropped from the data set because it occurred only once in the 45 blocks of known authorship.

Source: Stock and Trebbi (2003)

# Grammatical Constructions Can Also Be Used to Identify Authors

*Table 2*
**Grammatical Statistics Used in the Stylometric Analysis**

| |
|---|
| occurrences of Saxon genitives forms 's or s' |
| noun followed by adverb |
| noun followed by auxiliary verb |
| noun followed by coordinating conjunction |
| coordinating conjunction followed by noun |
| coordinating conjunction followed by determiner |
| total occurrences of nouns and pronouns |
| total occurrences of main verbs |
| total occurrences of adjectives |
| total occurrences of adverbs |
| total occurrences of determiners and numerals |
| total occurrences of conjunctions and interrogatives |
| total occurrences of prepositions |
| dogmatic/tentative ratio: assertive elements versus concessive elements |
| relative occurrence of "to be" and "to find" to occurrences of main verbs. |
| relative occurrence of "the" followed by an adjective to occurrences of "the" |
| relative occurrence of "this" and "these" to occurrences of "that" and "those" |
| relative occurrence of "therefore" to occurrences of "thus"; 0 if no occurrences of "thus" |

*Notes:* These grammatical statistics are the subset of those used by Mannion and Dixon (1997) after dropping statistics that overlap with Table 1 or are sequential word counts, which are ambiguous in mathematical texts.

Source: Stock and Trebbi (2003)

# Philip and Sewall Differ in Their Use of Language

*Table 3*

**Summary Statistics for the Six Stylometric Indicators with the Largest *t*-Statistics**

| | *Philip* | | *Sewall* | | | *Appendix B* | |
|---|---|---|---|---|---|---|---|
| | *Mean* | *Standard Deviation* | *Mean* | *Standard Deviation* | *t* | *Mean* | *Standard Deviation* |
| noun followed by coordinating conjunction | 26.8 | 7.0 | 17.3 | 4.6 | 5.55 | 27.0 | 5.0 |
| to | 29.5 | 5.8 | 20.9 | 6.1 | 4.79 | 28.0 | 8.6 |
| now | 1.6 | 1.5 | 0.1 | 0.3 | 4.74 | 1.1 | 1.0 |
| when | 2.4 | 2.1 | 0.3 | 0.7 | 4.72 | 1.8 | 1.2 |
| in | 22.7 | 5.3 | 29.8 | 5.5 | −4.34 | 18.5 | 5.8 |
| so | 2.1 | 1.6 | 0.7 | 0.8 | 3.82 | 2.0 | 1.7 |
| *n* | 25 | | 20 | | | 6 | |

*Notes:* The entries in columns 2 and 3 are the mean and standard deviations of the counts per 1,000 words of the stylometric indicator in column 1 in the 25 blocks undisputedly written by Philip Wright. Columns 4 and 5 contain this information for the 20 blocks undisputedly written by Sewall Wright. The next column contains the two-sample *t*-statistic testing the hypothesis that the mean counts are the same for the two authors. The final two columns contain means and standard deviations for the 6 blocks from Appendix B. Shaded indicators occur in the excerpt in Exhibit 2.

Source: Stock and Trebbi (2003)

# Cross-Validation to Demonstrate Validity of the Approach

*Table 4*

**Cross-Validation Estimates of Accuracy Rates of Assigned Authorship**

| | *Principal Components Regression* | | *Linear Discriminant Analysis* | |
|---|---|---|---|---|
| | *Predicted Author:* | | *Predicted Author:* | |
| *True Author:* | *Sewall* | *Philip* | *Sewall* | *Philip* |
| Sewall | 100% | 0% | 90% | 10% |
| Philip | 0% | 100% | 0% | 100% |

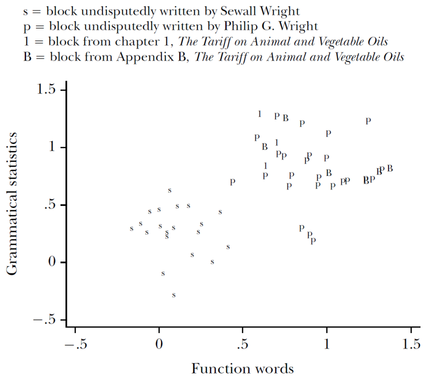
*Notes:* Based on leave-one-out cross-validation analysis of 45 1,000-word blocks of known authorship.

Source: Stock and Trebbi (2003)

# Who Invented 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)