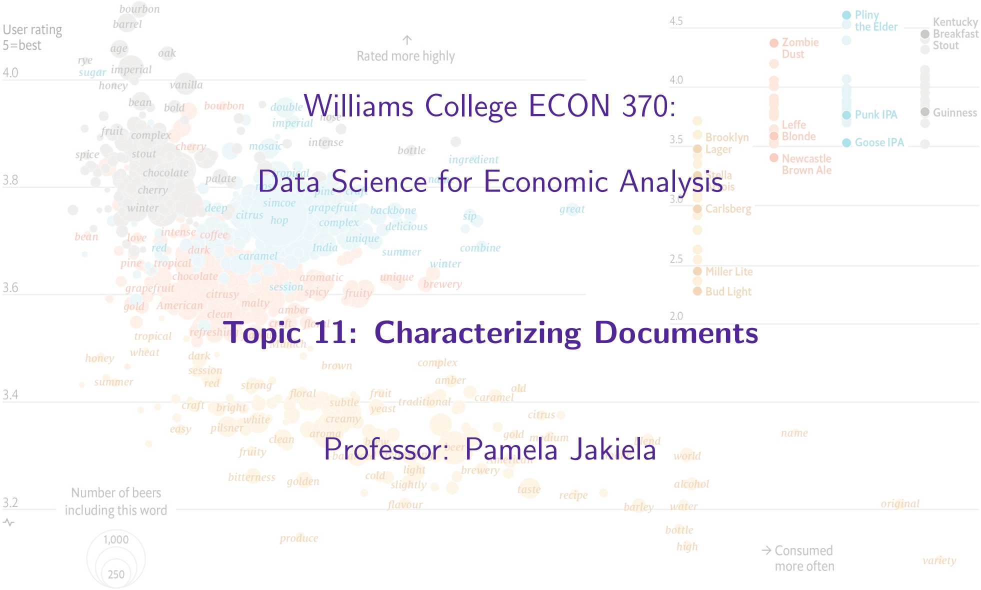


Williams College ECON 370:

Data Science for Economic Analysis

Topic 11: Characterizing Documents

Professor: Pamela Jakiela

Outline

- Sentiment analysis (very briefly)
- The importance of validation
- *k*-means clustering

Sentiment Analysis

In **sentiment analysis**, we assign documents to categories reflecting their emotional content

- The simplest forms of sentiment analysis use a **dictionary-based** methodology, counting the number of (for example) positive and negative words contained in each document
- Dictionary-based methods can also be used to assess political slant or capture topics

The bing sentiments data, for example, classifies almost 7,000 words as positive or negative

- Positive: abound, abounds, abundance, abundant, accessible, acclaim, . . .
- Negative: abnormal, abolish, abominable, abominably, abominate, abomination, . . .

There are many known issues with sentiment analysis (accounting for context, phrases, irony), only some of which can be solved through better validation and context-specific dictionaries

Sentiment Analysis in NBER Working Paper Abstracts

Most negative: Poverty, Hardship, and Government Transfers

We examine how the well-being of those with few resources changed, amidst economic disruption and large, transitory government transfers. We find that in the years leading up to the pandemic and in 2020, the patterns for income and consumption poverty were very similar. In 2021 and 2022, however, changes in income and consumption poverty were quite different–consumption poverty fell less than income poverty in 2021, and then income poverty rose sharply in 2022 while consumption poverty continued to decline. Reports of hardships rose in 2022 for both families with and without children, suggesting increased concern about financial well-being as COVID-era transfer programs expired. A key difference between income and consumption measures appears to be saving during the pandemic followed by dissaving, even among those near the poverty line. This finding indicates that permanent income models can even be relevant when low-income households, that typically have very limited saving, receive very large transitory payments. Unlike past academic studies and numerous politicians and pundits that have attributed most of the decline in income poverty in 2021, and its subsequent rise in 2022, to the Child Tax Credit, we show that expanded Unemployment Insurance and stimulus payments played a larger role.

Sentiment Analysis in NBER Working Paper Abstracts

Most positive: The Essential Role of Altruism in Medical Decision Making

Patients rely on medical care providers to act in their best interests because providers understand disease pathology and appropriate treatment much better than patients. Providers, however, not only give advice (diagnose) but also deliver (sell) treatments based on that advice. This creates a moral hazard dilemma where provider financial interests can diverge from patient interests, especially when providers are motivated more by profits than by altruism. We investigate how profit motivated versus altruistic preferences influence medical care decision making in the context of malaria in Kenya. We measured the appropriateness of care using data from an audit study that employed standardized patients (SP) who were trained to present as real patients the identical clinical case scenario to providers. The SPs were confirmed to be malaria negative before and after field work with a very reliable and sensitive blood test at a high-quality laboratory. We measured provider preferences using a lab in the field, real stakes, modified version of the dictator game. We find that more profit-motivated providers report higher rates of false-positive malaria test results than do more altruistic providers. Specifically, purely profit motivated providers report 30 percentage points more positives than providers who are altruistically motivated, and providers likely knew that the positive results that they reported to their patients were false. We also find that more profit motivated providers sold more unnecessary antimalarial drugs. . .

Validation

In **sentiment analysis**, we assign documents to categories reflecting their emotional content

- Whenever researchers use a method designed to measure a construct (e.g. sentiment), they need to establish that the method captures what they think it does (**validation**)
- Many sentiment dictionaries were developed for purposes other than research (in fact, almost all data science tools were not developed for quantitative social science)

How would you validate a measure of sentiments appropriate for economics?

- Maybe don't look at academic working papers
- Hand code a small sample of relevant documents (e.g. articles about the economy)
- Check that your results make sense (in a subsample), modify dictionary accordingly

Validation is central to all forms of measurement including unsupervised learning

k-Means Clustering

In clustering, we partition the sample into K distinct subsets or **clusters**

- Objective is to identify unobserved groups of units (e.g. documents, authors, consumers)
- Clustering is an **unsupervised learning** technique: no outcome variable to predict
- Groups are chosen by finding observations that look similar in terms of their X variables
- In practice, we must define "similar" in terms of some distance metric (e.g. Euclidean)

Text applications of clustering in economics focus on identifying industries or research fields

- Lubcyzk and Moser (2024) use k-means clustering to define narrow research fields, and instrument for women's entry using the share of (male) scientists who enlisted in WWII

k-Means Clustering Minimizes (?) the Distance Between Observations

- Distance between $x_a = (x_{a1}, \ldots x_{ap})$ and $x_b' = (x_{b1}, \ldots x_{bp})$: $d_{ab} = \sqrt{\sum_{j=1}^{p}(x_{aj} - x_{bj})^2}$
- Variation within cluster C_k if C_k contains N_{C_k} observations:

$$\sum_{a,b \in C} \sum_{j=1}^{p} (x_{aj} - x_{by})^2 / N_{C_k} = \sum_{a,b \in C} d_{ab}^2$$

  In (other) words, k-means clustering minimizes the sum of squared Euclidean distances across all pairs of points within a cluster, so that variance is between (not within) clusters

- In text analysis, the x variables are based on the text, typically constructed from words
  - Word counts or term frequencies, typically after some amount of preprocessing
  - Term frequency-inverse document frequency (tf-idf)

Creating a Document Term Matrix

Clustering algorithms use data structured so that the document is 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

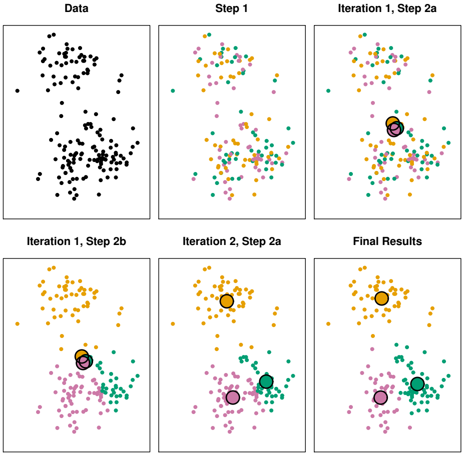
- Values of the word variables typically functions of within-document word counts

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

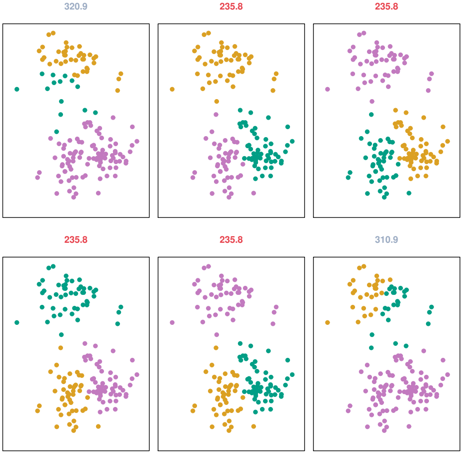
A *k*-Means Clustering Algorithm



Source: James et al. (2021)

1. Randomly assign initial clusters
2. Compute **centroid** of each cluster
3. Reassign clusters, mapping each observation to the nearest centroid
4. Recalculate centroids
5. Repeat Steps 3 and 4 until process converges, and no observations are reassigned at the next iteration

k-Means Clustering Identifies a Local (Not a Global) Optimum

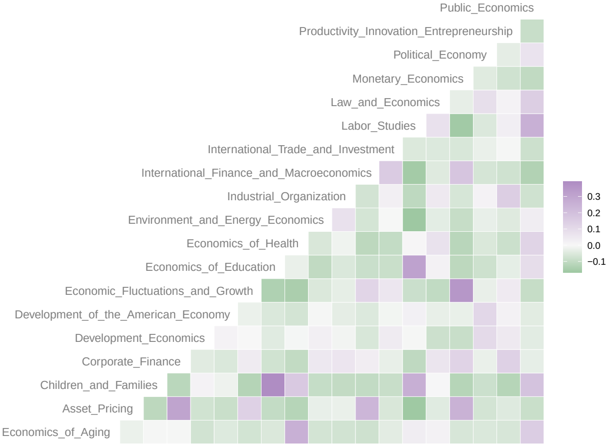


Source: James et al. (2021)

k-Means Clustering: Practical Considerations

- How much preprocessing, and how context-specific should it be?
- How many clusters? Should they be small or large?
  - k-means vs. hierarchical clustering
- How many random starts are needed to avoid local maxima? How robust are the clusters?
  - Check for (unexpected) clusters that are very small, exclude outliers or adapt preprocessing
  - Clustering can be very sensitive to small changes in sample construction (so beware)
- How to validate the clusters?
  - Can you hand code a sample of observations to assess the meaningfulness of the partition?
  - Do the groupings of words suggest that the clusters are meaningful groupings?

Example: *k*-Means Clustering 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)

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

Good news:

- With 50 initializations, chosen "best" clustering appears to be consistent across seeds

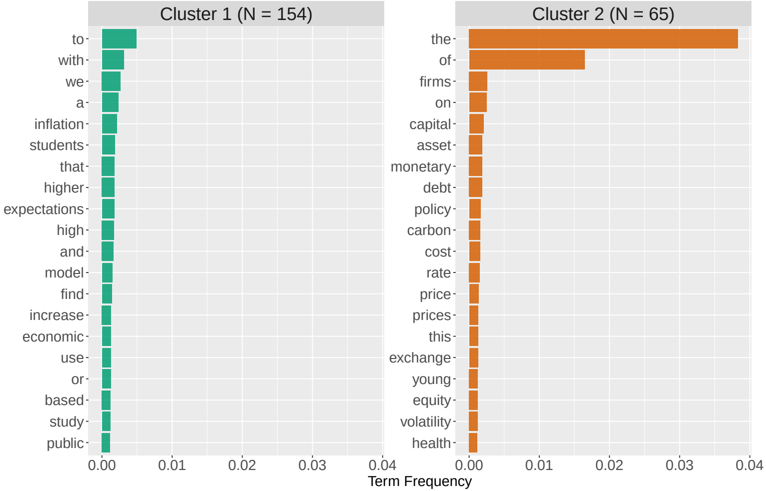
  $\Rightarrow$ Suggests we have identified a global maximum

- Cluster sizes are reasonable, no $N = 1$ clusters

Not-so-good news:

- Clusters do not line up with the split between NBER programs (C1: 77/77, C2: 39/26)
- Words most associated with each cluster suggest that we have fit the noise

Example: Words Most Associated with Each Cluster



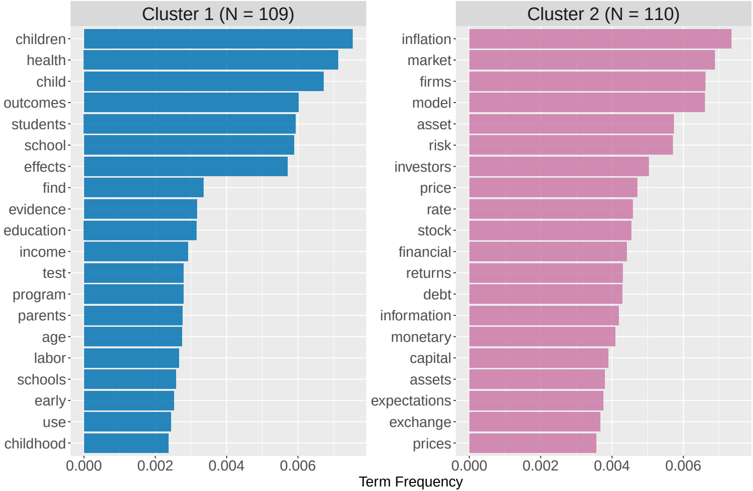
After Removing Stop Words: Magic Happens

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

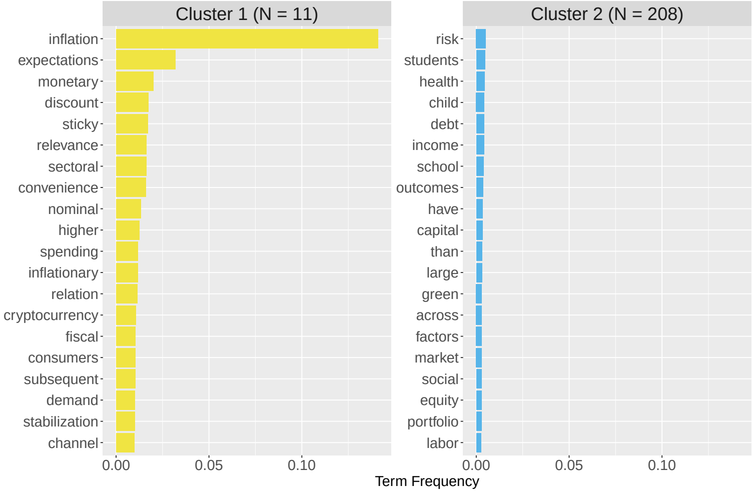
The four Asset Pricing papers included in Cluster 1:

- "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



Lab #11

Objective: implement *k*-means clustering in a sample of NBER working papers from two programs (not Asset Pricing + Children/Families and not Development + Children/Families)

The template for lab #11 implements *k*-means clustering with four clusters using papers from the (not mutually exclusive) Development Economics and Children and Families programs

- Clustering works well in R, somewhat less well in Python (weirdly)
- Your task is to adapt the code to cluster papers from your two favorite programs
- Change the preprocessing, number of clusters, etc. to arrive at a stable, sensible clustering