Text as Data

Ciencia de Datos y Econometría Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Text as Data: The Big Picture

- **Text is a vast source of data for research, business,etc**
- It comes connected to interesting "author" variables
  - What you buy, what you watch, your reviews
  - Group membership, who you represent, who you email
  - Market behavior, macro trends, the weather

Text as Data: Motivation

THE JOURNAL OF FINANCE • VOL. LXII, NO. 3 • JUNE 2007

Giving Content to Investor Sentiment: The Role of Media in the Stock Market

PAUL C. TETLOCK*

ABSTRACT

I quantitatively measure the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column. I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These and similar results are consistent with theoretical models of noise and liquidity traders, and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets.

Text as Data

- A passage in '*As You Like It*' from Shakepeare:

  All the world's a stage,
  and all the men and women merely players:
  they have their exits and their entrances;
  and one man in his time plays many parts...

Text as Data

- A passage in '*As You Like It*' from Shakepeare:

  All the world's a stage,
  and all the men and women merely players:
  they have their exits and their entrances;
  and one man in his time plays many parts...

- What the econometrian sees tokens:

world	stage	men	women	play	exit	entrance	time
1	1	2	1	2	1	1	1

From Text to Tokens

- Tokenization = splitting text into units
- Most often: words
- Example:
  "The economy is growing fast" $\rightarrow$ [the, economy, is, growing, fast]
- Tokens are the building blocks for Bag-of-Words

Why Word Counts?

- Texts are turned into signals, words to numbers
- **Distributional Hypothesis** (Harris, 1954): words in similar contexts have similar meanings
- Documents with similar word distributions likely discuss similar topics
- Counts give a first approximation of meaning

Counts as Approximation of Meaning

- Simplify: ignore order and grammar, keep word frequencies
- Frequent words reflect themes and emphasis
- Analogy to economics: start parsimonious, add frictions later

Bag-of-Words (BoW) Idea

- Once text has been tokenized
- Ignore grammar, word order, syntax, and context
- Represent each document as a vector of counts:

$$d = (c_1, c_2, \ldots, c_V)$$

- What is lost: structure and nuance
- What is kept: themes and emphasis
- Philosophy: sacrifice structure, keep signal

From Document to Corpus

- **Document**: a single text (article, tweet, speech, passage)
- Represented as a vector of word counts after tokenization

$$d = (c_1, c_2, \ldots, c_V)$$

- **Corpus**: a collection of documents
- Examples:
  - All Shakespeare plays
  - A set of news articles
  - Student essays in this class

Document-Term Matrix (DTM)

- Apply Bag-of-Words to every document in a corpus
- Stack results into a matrix:

$$X = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1V} \\ c_{21} & c_{22} & \dots & c_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ c_{D1} & c_{D2} & \dots & c_{DV} \end{bmatrix}$$

- Rows = documents (D)
- Columns = terms (V words in the vocabulary)
- Entries c_{ij} = count of word j in document i

Document-Term Matrix (DTM)

- A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection (corpus) of documents.
- Example 3 documents, A, B, y C:

1. El sol es una estrella.
2. Un buen viajante no tiene planes.
3. Juan tiene una mascota nueva

Text as Data: Cleaning and Tokenization

- Not all words are useful in the DTM
- **Cleaning:** remove elements that do not add meaning to content or structure
- There is no single recipe, depends on purpose and source

Cleaning: Common Steps

- Convert to lowercase, drop numbers and punctuation
  - But context matters: don't drop `:-)` from tweets!
- Remove **stopwords** (if, and, but, the, they, ...)
  - Caution: one person's stopword may be another's key term
- Remove rare words (very low frequency across docs)

Stemming vs Lemmatization

- Goal: reduce words to a base form
- **Stemming:** crude rules, chop endings
- **Lemmatization:** uses vocabulary and grammar
- Example: *organize, organizes, organizing* $\rightarrow$ *organize*

Stemming vs Lemmatization: Examples

- am, are, is $\rightarrow$ be
- car, cars, car's, cars' $\rightarrow$ car
- the boy's cars are different colors
  - Stemming: boy car be differ color
  - Lemmatization: boy car be different color

Stemming in Practice

- **Stemming** = rule-based process that chops word endings
- **English:**
  - Porter Stemmer (classic, simple rules)
  - Snowball Stemmer (successor, more flexible)
  - Lancaster Stemmer (more aggressive, often too short)
- **Spanish:**
  - Snowball Stemmer for Spanish (handles gender and plural forms)
  - Example: *niños, niñas, niño, niña* $\rightarrow$ *niñ*
- Pros: fast, language-specific implementations
- Cons: crude, may create non-words (e.g., *relational* $\rightarrow$ *relat*)

Lemmatization in Practice

- **Lemmatization** = reduce to dictionary form (lemma) using vocabulary + grammar
- Uses part-of-speech tagging to disambiguate
- **English:**
  - WordNet Lemmatizer (NLTK)
  - spaCy's lemmatizer (state-of-the-art, context-aware)
- **Spanish:**
  - spaCy Spanish models (es_core_news_sm, md, lg)
  - Freeling (NLP library for Spanish/Catalan)
- Pros: returns valid words, handles meaning and context
- Cons: slower, requires linguistic resources (POS tagging, dictionaries)

Stemming vs Lemmatization: Comparison

	Stemming	**Lemmatization**
Method	Rule-based (chop endings)	Dictionary + grammar
Output	Often non-words (*relat*)	Valid words (*relation*)
Speed	Fast, lightweight	Slower, needs resources
Accuracy	Crude, may conflate terms	Context-aware, more precise
English tools	Porter, Snowball, Lancaster	WordNet, spaCy
Spanish tools	Snowball (Spanish rules)	spaCy (es_core_news), Freeling
When should I use it?	Large-scale, noisy corpora	When interpretability matters

Trade-off: speed vs. accuracy

Bag-of-Words: Strengths and Limits

- **Strengths**
  - Simple, fast, often a strong baseline
  - Captures themes through word choice and frequency
- **Limitations**
  - Ignores context (*bank* = river or money)
  - Ignores order (“man bites dog” = “dog bites man”)
  - Very high-dimensional and sparse

N-grams: Adding Local Order

- An **n-gram** is a sequence of n consecutive words
- Examples:
  - Unigrams: economy, growth
  - Bigrams: central bank, machine learning
  - Trigrams: New York City, stock market crash
- Capture short-range context and common phrases

N-grams: Trade-offs

- Vocabulary size grows quickly with n (V^n possible n-grams)
- Many n-grams are rare $\rightarrow$ sparsity increases
- Only capture local context, not long-distance dependencies
- Still a useful bridge between Bag-of-Words and advanced models

Putting It Together

- **Cleaning:** remove noise and normalize
- **Bag-of-Words:** simplest representation, counts
- **N-grams:** enrich with short sequences
- **Next:** reduce dimensionality, uncover structure (topics, embeddings)

Example

Econometrica, Vol. 78, No. 1 (January, 2010), 35–71

WHAT DRIVES MEDIA SLANT? EVIDENCE FROM U.S. DAILY NEWSPAPERS

BY MATTHEW GENTZKOW AND JESSE M. SHAPIRO[1]

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to consumer preferences, which account for roughly 20 percent of the variation in measured slant in our sample. By contrast, the identity of a newspaper's owner explains far less of the variation in slant.

KEYWORDS: Bias, text categorization, media ownership.

Example

Gentzkow and Shapiro: What drives media slant? Evidence from U.S. daily newspapers (*Econometrica*, 2010)

- Build an economic model for newspaper demand that incorporates political partisanship (Republican vs Democrat)
  - What would be independent profit-maximizing "slant"?
  - Compare this to slant estimated from newspaper text.
- Use data from Congress to isolate the phrases
- Compare phrase frequencies in the newspaper with phrase frequencies in the 2005 Congressional Record to identify whether the newspaper's language is more similar to that of a congressional Republican or a congressional Democrat

Republican	Democratic
death tax	estate tax
tax relief	tax break
personal account	private account
war on terror	war in Iraq

Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/