Text as Data[1]

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

[1]This section is based on [Gentzkow et al., 2019]

Outline

Impact of New Technologies on Social Sciences Research

- New technologies have made availability
  - increasing share of human interaction, communication, and culture.
- Rich complement to the structured data traditionally used in research.
- Explosion of empirical economics research using text as data.

Application

- **Finance**,
  - text from financial news, social media, and company filings is used to predict asset price movements and study the causal impact of new information.
- **Macroeconomics**
  - forecast variation in inflation and unemployment, and estimate the effects of policy uncertainty.
- **Media economics**
  - text from news and social media is used to study the drivers and effects of political slant.
- **Industrial organization** and marketing,
  - text from advertisements and product reviews is used to study the drivers of consumer decision making.
- **Political economy**
  - text from politicians speeches is used to study the dynamics of political agendas and debate.

Applications and Challenges of Text Data in Economics Research

- Applications in various fields:
  - finance, macroeconomics, media economics, industrial organization, marketing, and political economy.
- Challenges: **High-Dimensionality**
  - A 30-word Twitter message using only the 1,000 most common English words has as many dimensions as the number of atoms in the universe.
- Statistical methods that analyze text are related to high-dimensional data analysis methods used in other fields (e.g., machine learning and computational biology).
  - Some are applied directly (e.g. lasso and penalized regressions).
  - Others are adapted to the specific structure of text data (e.g. topic models and multinomial inverse regression).

Three-Step Analysis of Text Data

1. Represent the raw text D as a numerical array C.
2. Map C to predicted values $\hat{V}$. of unknown outcomes V.
3. Use $\hat{V}$. in subsequent descriptive or causal analysis.

Outline

Step 1: Representation of Raw Text as Numerical Array

- Preliminary restrictions is necessary to **reduce the dimensionality of the data** to a manageable level.
  - Techniques cannot manage extremely high dimensional raw data (e.g., 1000^{30}-dimensional raw Twitter data).
- The elements of C are usually counts of **tokens**:
  - words, phrases, or other pre-defined features of text.
- Focus on features such as especially diagnostic words or phrases.
- The mapping from D to C leverages prior information about the structure of language to **reduce dimensionality** before any statistical analysis.

Outline

Step 2: Application of High-Dimensional Statistical Methods

- The second step involves the application of high-dimensional statistical methods.
- In this step, data is mapped to a predicted outcome, $\hat{V}$.
- Classic examples:
  - Determining whether an email is spam.
  - Sentiment prediction.
  - Predicting flu outbreaks from Google searches.

Focus on Prediction over Causal Inference

- The ultimate goal in most text analysis settings is **prediction**, not causal inference.
- The interpretation of the mapping from V to $\hat{V}$ is not usually of interest, as long as it generates accurate predictions.
- Examples: Scott and Varian (2014, 2015) used data from Google searches to produce high-frequency estimates of macroeconomic variables.

Applications of Text Analysis

- Groseclose and Milyo (2005) compared the text of news outlets to speeches of congresspeople to estimate the outlets' political slant.
- A large literature in finance following Antweiler and Frank (2004) and Tetlock (2007) uses text from the Internet or the news to predict stock prices.

Outline

Step 3: Inferring Causal Relationships and Structural Parameters

- Goal for social science studies: use text to **infer causal relationships or parameters of structural economic models**.
  - Stephens-Davidowitz (2014): Used Google search data to estimate local areas' racial animus and studied its causal effect on votes for Obama in the 2008 election.
  - Gentzkow and Shapiro (2010): Estimated each news outlet's political slant using congressional and news text, then studied the supply and demand forces that determine slant in equilibrium.
  - Engelberg and Parsons (2011): Measured local news coverage of earnings announcements and used the relationship between coverage and trading by local investors to separate the causal effect of news from other sources of correlation between news and stock prices.

Representing Text as Data

- Human interprets words in the context of other words.
- Most text analysis in the social sciences and machine learning simplifies this complexity.
- Simplifications for statistical analysis:
  - Dividing the text into individual documents.
  - Reducing the number of language elements considered.
  - Limiting the encoding of dependence among elements within documents.
- The result is a mapping from raw text to a numerical array, where **each row indicates the presence or count of a particular language token in a document**.

What Is a Document?

- The first step in constructing **C** is to divide raw text into individual documents.
  - Depend on **the level at which the attributes of interest are defined**.
  - For spam detection: individual emails.
    For predicting daily stock price movements: divide news text by day.
- In some cases, the definition of a document is not clear.
  - [Gentzkow et al., 2016] To predict legislators' partisanship from speeches,we could aggregate speech into speaker-day, speaker-year, or all speech by a given speaker during her time in Congress.
- Finer partitions ease computation, but limit the dependence we can capture.
- Important to check the sensitivity of results.
  - Theoretical guidance for the right level of aggregation is often limited.

Feature Selection - Part 1

- To manage the number of features, we can:
  - Eliminate elements like punctuation, numbers, HTML tags, proper names, etc.
  - Remove common words (stop words) and very rare words.
  - Filter by "term frequency-inverse document frequency" (**tf-idf**):
- **Term frequency (tf)** : the count of occurrences of a word in a document c_{ij}.
- **Inverse document frequency (idf)** : the log of one over the share of documents containing the word $\log(n)/d_j$, where $d_j = \sum_i \mathbb{1}(c_{ij} > 0)$.
- **Tf-idf** :the product of $tf_{ij} \times idf_j$.
  - Common and rare words have low tf-idf scores.
  - Words with tf-idf scores above a certain rank or cutoff are kept.

Feature Selection - Part 2

- Another step is stemming: replacing words with their root form.
- e.g. "economic", "economics", "economically" replaced by "economic"
  - The Porter stemmer [Porter, 1980] is a standard stemming tool for English language text.

Careful with Feature Selection

- These cleaning steps reduce the dimensionality of the data, easing computation and improving model interpretability.
- However, different elements may carry meaning in different contexts.
  - Example: One researcher's stop words are another's subject of interest. Dropping numerals from political text means missing references to the first 100 days or September 11.

n-grams and Bag-of-Words Representation

- Limit dependence among language elements for tractable representation.
- Bag-of-words: simplest and most common document representation.
  - Ignore the order of words.
  - Represent documents as vectors with word frequency counts.
- Example: "Good night, good night! Parting is such sweet sorrow." $\rightarrow$ "good night good night part sweet sorrow."
  Bag-of-words representation:
  - $c_{ij} = 2$ for $j \in \{\text{good,night}\}$,
  - $c_{ij} = 1$ for $j \in \{\text{part,sweet,sorrow}\}$,
  - and $c_{ij} = 0$ for all other words.

Using n-grams for Richer Modeling

- Count unique phrases of length n instead of individual words.
  - Example: Count 2-grams (bigrams) in the snippet above.
- Characteristics:
  - Allow limited description of word dependence, leading to richer modeling.
  - Useful in analysis of partisan speech [Gentzkow and Shapiro, 2010]: "death tax and tax break are phrases with strong partisan overtones that are not evident to look at single word.
  - Dimension of phrase tracking increases exponentially with order n.
- Most text analyses consider up to 2-grams or 3-grams.
- Best practice: start with single words, then evaluate if 2-grams or 3-grams are worth the extra time.

Richer Representations

- Computational linguistics offer methods for capturing richer text features.
- Syntax-informed text tokens: syntactic n-grams [Goldberg and Orwant, 2013].
  - Group words together when their meaning depends on each other.
- Alternative approach: Consider an ordered sequence of transitions between words.

Sentence Representation and Word Embedding

- Break document into sentences.
- Represent a sentence using binary $p \times s$ matrix S.
  - Nonzero elements indicate row-word occurrence in column-position.
  - p is the vocabulary length.
- Massive increase in data dimensions.
- Analyze data using word embedding: map words to $\mathbb{R}^K$ for $K \ll p$.
  - Sentences become sequences of points in K-dimensional space.

Outline

Idea of Document-Attribute Mapping

- Map the document-token matrix C to predictions $\hat{V}$ of an attribute V.
- Observed data can be partitioned into C_{train} and C_{test}. C_{train} has observed V_{train} and C_{test} has unobserved V.
- Size of C_{train} is $n_{\text{train}} \times p$ and V_{train} is $n_{\text{train}} \times k$, where k is the number of attributes to predict.
- Attributes in V can be observable (e.g., frequency of flu cases, movie reviews rating, unemployment rate) or latent (e.g., topics in a congressional debate).

Document-Attribute Mapping Methods

Methods to connect counts c_i to attributes v_i can be roughly divided into four categories:

- **Dictionary-based methods**
- **Text regression methods**
- **Generative models**
- **Word embeddings**

	How it works	Notes
Dictionary	specify $v_i = f(c_i)$ for known function $f(\cdot)$	no statistical inference most common
Text Reg.	begin from $p(v_i \mid c_i)$	
Generative	begin from $p(c_i \mid v_i)$	unsupervised & supervised
Word Emb.	similar meaning: similar vector	high value in the future

Outline

Dictionary-based - Idea

Dictionary-based methods:

- Use the **relative frequency of key words** to
  - classify documents into categories, or
  - measure the extent to which documents belong to particular categories.
- Dictionary to **measure tone** (e.g., Eshbaugh-Soha 2010)
  - A list of words: Dichotomously classified as positive or negative, or contain more continuous measures of their content.
  - Each word $m \in \{1, \cdots, M\}$ will have associated score s_m.
  - Simplest: $s_m = -1$ for words with negative tone, and $s_m = 1$ for words with positive tone.
  - If $N_i = \Sigma_{m=1}^{M} W_{im}$ words are used in document i, then dictionary methods can **measure the tone for any document** t_i as

$$t_i = \Sigma_{m=1}^{M} \frac{s_m W_{im}}{N_i}.$$

  - t_i can also be used to **classify documents into tone categories** (e.g., $t_i > 0$: positive, $t_i < 0$: negative)

Dictionary-based - General

In general, dictionary-based methods:

- Simply specify $\hat{v}_i = f(c_i)$ for some known function $f(\cdot)$, without involving statistical inference.
- Common in social science literature using text:
  Researchers define $f(\cdot)$ based on a prespecified dictionary of terms capturing particular text categories.
- Examples:
  - Tetlock (2007)
    c_i is a bag-of-words representation;
    v_i is the latent sentiment of Wall Street Journal columns;
    $f(\cdot)$ is defined using a dictionary called the General Inquirer.
  - [Baker et al., 2016]
    c_i is the count of newspaper articles containing certain terms;
    v_i is the degree of policy uncertainty;
    $f(\cdot)$ is the raw count of the prespecified terms divided by the total number of articles, averaged across newspapers.

Dictionary-based - Advantages

Dictionary-based methods are **easy and cheap to apply**:

- They identify words that separate categories and measure how often those words occur in texts (see Kellstedt 2000; Laver & Garry 2000; Burden & Sanberg 2003; Young & Soroka 2011).
- Finding the separating words is relatively easy.
- Many widely used off-the-shelf dictionaries available providing key words for many categories. (e.g., Bradley & Lang 1999; Hart 2000; Pennebaker, Francis & Booth 2001; Turney & Littman 2003).
- If documents are already coded into categories, dictionaries can be produced using existing methods. (e.g., Monroe, Colaresi & Quinn 2008; Taddy 2010; Diermeier et al. 2011).

Dictionary-based - Limitations

Dictionary-based methods **require close alignment** between word scores and their usage in a specific context.

- Applying dictionaries from one domain to another can lead to serious errors.
- Scholars **rarely validate the measures from dictionaries**.
- Although establishing granular scales of sentiment (e.g., tone) is useful for applications, humans are unable to produce the same granular measures reliably (Krosnick 1999).
  - E.g., Loughran & McDonald (2011) criticize the use of off-the-shelf dictionaries to measure tone in corporate earning reports.
  - Cancer, which has negative connotations in general contexts, could have a positive connotation in health-care company.

Dictionary-based - Solution

To effectively use dictionary methods in their future work, advances in the validation of dictionary methods must be made:

- Simplify the classification problem by using binary categories (e.g., positive or negative tone) and validate against human gold standards.
- Treat dictionary measures like unsupervised methods, establishing validity based on multiple standards.

Outline

Text Regression Methods

Text regression methods:

- Directly estimate the conditional outcome distribution, usually via the conditional expectation $E[v_i|c_i]$ of attributes v_i.
- To predict v_i from c_i, we would naturally regress the observed values of the former (V_{train}) on the corresponding values of the latter (C_{train}).
- Any generic regression technique can be applied, depending on the nature of v_i.
- However, the high dimensionality of c_i, where p is often as large as or larger than n_{train}, makes OLS and other standard techniques infeasible.
- So, we need techniques such as **penalized linear** or **logistic regression**.

Penalized Linear - Intro & Importance

Penalized linear models:

- The most popular strategy for high-dimensional regression in contemporary statistics and ML, particularly with L_1 penalization.
- Recommended for most text regression applications:
  - Linear models are intuitive and interpretable.
  - Fast, high-quality software is available for big sparse input matrices.
  - For simple text-regression tasks, it is seldom possible to do much better (e.g., out-of-sample prediction) than penalized linear models when the input dimension is near to the sample size.

Penalized Linear - Details & Estimator

- Linear models in this context are those where v_i depends on c_i through a linear index $\eta_i = a + x_i'\beta$, where
  - x_i a known transformation of c_i
  - $E[v_i|x_i] = f(\eta_i)$ for some known link function $f(\cdot)$.
- For the transformation $c_i \to x_i$, the best choice is application-specific. Common examples:
  - Identity $x_i = c_i$;
  - Normalization by document length $x_i = c_i/m_i$ with $m_i = \Sigma_j c_{ij}$;
  - Positive indicator $x_{ij} = 1[c_{ij} > 0]$.
- $l(a, \beta)$ loss function: unregularized objective proportional to the negative log likelihood, $-\log p(v_i|x_i)$. For example:
  - Gaussian (linear) regression: $l(a, \beta) = \Sigma_i (v_i - \eta_i)^2$;
  - Binomial (logistic) regression: $l(a, \beta) = -\Sigma_i [\eta_i v_i - \log(1 + e^{\eta_i})]$ $\forall v_i \in \{0, 1\}$.

- A penalized estimator is then the solution to

$$\min_{a,\beta}\{l(a,\beta) + n\lambda\Sigma_{j=1}^{p}\kappa_j(|\beta_j|)\}, \text{where}$$

  - $\lambda > 0$ controls overall penalty magnitude, and
  - $\kappa_j(.)$ are increasing cost functions, penalizing deviations of β_j from 0.
  - To choose optimal λ: cross-validation, AIC, BIC.

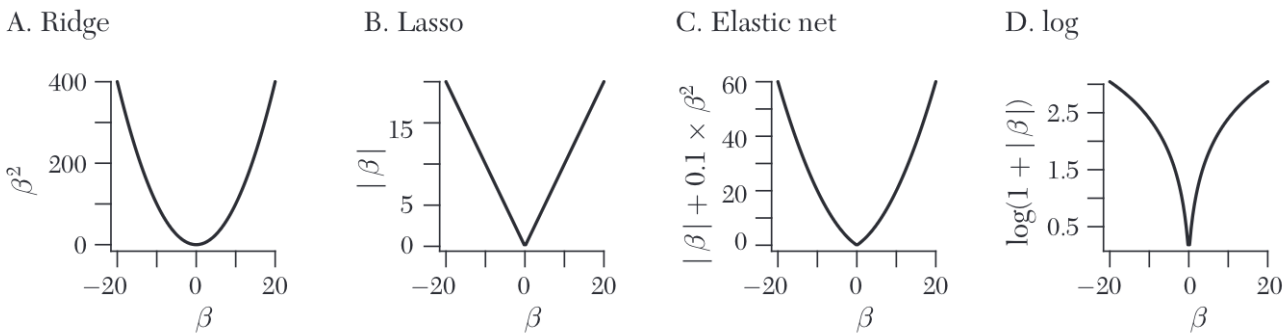


Figure 1

Note: From left to right, L_2 costs (ridge, Hoerl and Kennard 1970), L_1 (lasso, Tibshirani 1996), the "elastic net" mixture of L_1 and L_2 (Zou and Hastie 2005), and the log penalty (Candès, Wakin, and Boyd 2008).

Penalized Linear - Limit

- Dimension Reduction
  - If **predictors are highly correlated**, penalized linear model is suboptimal since it forces the coefficients on most regressors to be close to or exact 0 (shrinkage and variable selection).
  - The essence of dimension reduction:
    Use the simple average of the predictors as the sole predictor in a univariate regression.
  - 2 classic dimension reduction techniques are Principal Component Regression(PCR) and Partial Least Square Regression(PLS).

Penalized Linear - Alternatives

- Nonlinear Text Regression
  - Nonlinear regression method outperforms linear one if linear specification is **too restrictive**.
  - Common methods include **generalized linear models, support vector machines, regression trees, and deep learning**.
  - GLMs: include nonlinear functions of c_i (polynomial, interaction).
  - Regression Trees:
    - Sequentially sort data observations into bins based on values of the predictor variables;
    - Form predictions as the average value of the outcome variable within each partition.
- Bayesian Regression Methods: **Full Bayesian analysis** for high-dimensional regression. E.g., spike-and-slab.

Outline

Generative Model

Unlike text regression which begins from a model of $p(v_i|c_i)$, **generative model** begins from $p(c_i|v_i)$.

- In many cases, the underlying causal relationship runs **from outcomes to language** rather than the other way around.
  - E.g., Google searches about the flu do not cause flu cases to occur; rather, people with the flu are more likely to produce such searches.
  - E.g., Congresspeople's ideology is not determined by their use of partisan language; rather, people who are more conservative or liberal to begin with are more likely to use such language.
- From an economic point of view, the correct structural model of language in these cases **maps from v_i to c_i**.
- Modeling the underlying causal relationships can provide powerful guidance to **inference** and make the estimated model more **interpretable**.
- Generative models can be divided by whether the **attributes are observed or latent**.

Generative Model - Unsupervised

Unsupervised methods:

- We do not observe the true value of v_i for any documents.
- The function relating c_i and v_i is unknown, but we are willing to impose sufficient structure on it to allow us to infer v_i from c_i.
- In text analysis, the leading application is topic modeling (v_i are topics) and its variants (e.g., latent Dirichlet allocation, or LDA).

Generative Model - Supervised

Supervised methods:

- We observe training data V_{train} and we can fit our model, say $f_\theta(c_i; v_i)$ for a vector of parameters θ, to this training set.
- The fitted model $\hat{f}_\theta$ can then be inverted to predict v_i for documents in the test set and can also be used to interpret the structural relationship between attributes and text.
- Most common example: naive Bayes classifier, which treats counts for each token as independent with class-dependent means.

In some cases, v_i includes both observed and latent attributes for a **semi-supervised analysis**.

Supervised Learning > Dictionary

- Dictionary-based methods can assign texts to categories, but:
  - Requires pre-identified words to separate classes.
  - Inefficiencies arise when applied outside their original domain.
- In the **supervised learning methods**:
  - Human coders categorize documents by hand.
  - Algorithm learns to sort documents into categories using training set and words.

Advantages of supervised learning:

- **Domain specific**, avoid applying dictionaries outside intended use.
  - It requires scholar to develop coding rules and coherent definitions of concepts, leading to clarity in what are being measured and studied.
- **Easier to validate** with clear performance statistics.
  - There are clear statistics that summarize model performance

Supervised Learning - Steps

Three basic steps to do the supervised learning:

1. **Construct a Training Set:** Develop coding rules and define concepts.
2. **Apply Supervised Learning Method:** Learn the relationship between features and categories using the training set.
3. **Validate and Classify:** Assess model performance, validate the output, and classify remaining documents.

Supervised - Construct Training Set

Importance: No statistical model can repair a poorly constructed training set.

Training set construction:

- **Creating a Coding Scheme:**
  - Develop and iterate a coding scheme to address language ambiguities and nuanced concepts.
  - Use a concise codebook and revise it based on coder feedback and application to new documents.
- **Selecting Documents:**
  - Supervised learning methods use the **relationship between the features in the training set** to **classify the remaining documents in the test set**.
  - Ideally, training sets should be **representative of the corpus**.
  - Random sampling, either simple or stratified, is recommended.

Training Set Size:

Hopkins & King (2010): around 500 documents, with 100 to be sufficient.
The specific application determines the required number of documents.

Supervised - Apply Supervised Method I

After hand-classification, the documents are used to **train the supervised learning methods to learn about the test set**

- Classifying the individual documents into categories.
- Or measuring the proportion of documents in each category.

The common structure to do the classification:

- Suppose N_{train} documents ($i = 1, \cdots, N_{\text{train}}$) in training set, and each has been coded into one-of-K categories ($k = 1, \cdots K$).
- Each document i's category is represented by $Y_i \in \{C_1 \cdots C_K\}$, and the entire training set is represented as $Y_{\text{train}} = (Y_1, \cdots Y_{N_{\text{train}}})$.
- Each document i's features are contained in an M length vector W_i, which we collect in the $N_{\text{train}} \times M$ matrix W_{train}.

Supervised - Apply Supervised Method II

- Each supervised learning algorithm assumes that **there is some (unobserved) function that describes the relationship between the words and the labels**

$$Y_{\text{train}} = f(W_{\text{train}}).$$

- Each algorithm attempts to learn this relationshipestimating the function f with $\hat{f}$.
- $\hat{f}$ is then used to infer properties of the test set, $\hat{Y}$ testeither each document's category or the overall distribution of categoriesusing the test set's words W_{test}.

3 methods for inferring the relationship between words and categories:

- Individual Methods.
- Ensembles.
- Measuring Proportions.

Individual Methods: Naive Bayes Classifier I

Naive Bayes Classifier Overview: Based on Bayes' theorem, Naive Bayes is a simple yet powerful probabilistic classifier, often used in text classification. The model is based on the assumption that features (words) are independent given the class.

- **Objective:** Infer the probability that document i belongs to category k given its word profile W_i.
- **Formula:** $p(C_k|W_i) \propto p(C_k)p(W_i|C_k)$ where:
  - $p(C_k)$ is the prior probability of category k, estimated as $\hat{p}(C_k) = \frac{\text{Number of training documents in category } k}{N_{\text{train}}}$.
  - $p(W_i|C_k)$ assumes word independence within the category.

Individual Methods: Naive Bayes Classifier II

- **Estimation:** For a word m occurring j times in a document,

$$\hat{p}(W_{im} = j|C_k) = \frac{\text{Number of training documents in category } k \text{ with word } m \text{ used } j \text{ times}}{\text{Total documents in category } k}$$

- **Classification Rule:** The document is classified to the category maximizing the posterior probability,

$$\hat{f}(W_i) = \arg\max_k \left(\hat{p}(C_k) \prod_{m=1}^{M} \hat{p}(W_{im}|C_k) \right)$$

Ensemble Methods

Combining Classifiers:

- Individual classification methods can be accurate, but combining them can produce a superior classifier.
- Combining accurate and diverse classifiers improves overall accuracy (Jurafsky and Martin, 2009).

Benefits of Ensembles:

- Increase out-of-sample stability.
- Able to capture complex functional forms with simple classifiers (Dietterich, 2000; Hillard et al., 2008).

Ensemble Development Schemes:

- Super-learning: Assign weights to methods based on out-of-sample accuracy (van der Laan et al., 2007).
- Bagging: Repeatedly draw samples with replacement and classify out-of-sample cases.
- Boosting: Sequential training of classifiers, increasing weight on misclassified cases (Hastie et al., 2001).

Example: Stewart & Zhukov (2009) I

Research Context:

- Stewart and Zhukov (2009) compare foreign policy stances of civilian and military elites in Russian public statements.
- Corpus: 7920 Russian language public statements by political and military elites (1998-2008).
- Human coders developed a codebook to classify statements as restrained, activist, or neutral on Russian use of force.

Supervised Learning Approach:

- **Random Forest** model fitted using human-coded data to learn the relationship between words and classes.
- Remaining documents classified using the learned relationship.
- Ten-fold **cross-validation** performed on the training data to **compare machine and human classifications**.

Example: Stewart & Zhukov (2009) II

Random Forest v.s. human-coded classifications:

- **Accuracy**: Proportion of correctly classified documents (here, 0.65).
- **Precision**: Probability of correct classification given the machine's guess for a category (here, 0.65).
- **Recall**: Probability of machine correctly identifying a document in a category given human coding (here, 0.75).
- The differences between the precision and recall exemplify why the different statistics are useful.
  - The recall rate is higher than the precision here because the Random Forest algorithm guesses **too often** that a document is restrained.
  - It labels most of the human coder's restrained positions correctly.

Example: Stewart & Zhukov (2009) III

Figure: Human vs Supervised (Grimmer & Stewart (2013), 279)

		Training data		
		Restrained	*Activist*	*Neutral*
Machine	Restrained	111	31	28
	Activist	10	17	0
	Neutral	26	9	68

Outline

Word Embeddings

Word Embeddings:

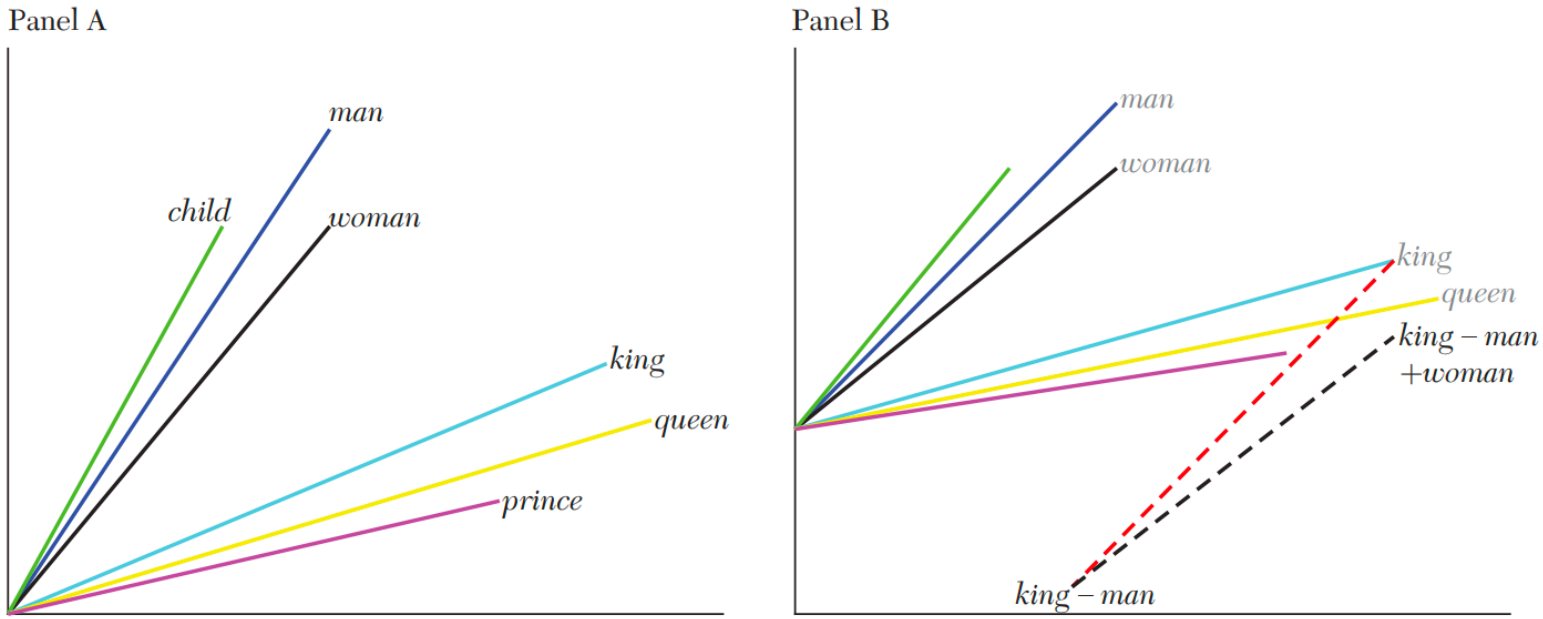
- Provide a richer representation of the underlying text than the token counts that underlie other methods.
- Allow words with **similar meanings** to have **similar vector representations**, which **lowers dimensional space**.
- Limited applications in economics to date, but their dramatic successes in deep learning and other machine learning domains suggest that they are likely to have high value in the future.
- Some popular techniques are **Word2Vec** (Mikolov et al. 2013) and Global Vector for Word Representation (**GloVe**, Pennington, Socher, and Manning 2014).

Word Embeddings - Example

Consider 6 words: {*king*, *queen*, *prince*, *man*, *woman*, *child*}.

- Combination *king* - *man* + *woman* is close to vector *queen*.

Figure: Word Embeddings (Gentzkow et al. (2019), 552.)



Outline

Practical Advice for Text Analysis Methods

- Dictionary-based methods
  - Suitable when strong, reliable prior information exists and data is weak.
  - Useful when limited or noisy training data is available.
- Text regression
  - Good choice for predicting single attributes with ample labeled training data.
- Generative models
  - Needed when multiple attributes are of interest and interdependencies must be controlled for.

Model Validation for Prediction

- Validate estimation approach performance.
- Check out-of-sample predictive performance on held-out data.
- Cross-validation for penalty selection.
- Reserve test data for the estimation of true average prediction error.

Model Validation for Descriptive or Causal Analysis

- Validate fitted model's accuracy for economic or descriptive quantities.
- Manual audits:
  - Informal: inspect documents and fitted $\hat{V}$.
  - Formal: human classification and quantitative evaluation of consistency.
  - Crucial for dictionary methods.
- Inspect estimated coefficients or model parameters directly.

References I


Baker, S. R., Bloom, N., and Davis, S. J. (2016).
Measuring economic policy uncertainty.
The quarterly journal of economics, 131(4):1593–1636.


Gentzkow, M., Kelly, B., and Taddy, M. (2019).
Text as data.
Journal of Economic Literature, 57(3):535–574.


Gentzkow, M. and Shapiro, J. M. (2010).
What drives media slant? evidence from u.s. daily newspapers.
Econometrica, 78(1):35–71.


Gentzkow, M., Shapiro, J. M., and Taddy, M. (2016).
Measuring group differences in high-dimensional choices: Method and application to congressional speech.
NBER Working Paper 22423, National Bureau of Economic Research.

References II


Goldberg, Y. and Orwant, J. (2013).
A dataset of syntactic-ngrams over time from a very large corpus of english books.
In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247.


Porter, M. F. (1980).
An algorithm for suffix stripping.
Program, 14(3):130–137.