DSME 6635: Artificial Intelligence for Business Research

Traditional NLP: Pre-processing and Word Representations

Renyu (Philip) Zhang

1

Agenda

- Natural Language Processing Framework
- Pre-processing
- Word Representation

2

2

Natural Language Processing

- Natural Language Processing (NLP): A subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

- Typical NLP:
  - Sentiment Classification
  - Machine Translation
  - Document Similarity
  - Topic Modelling
  - Etc.

- A classic NLP framework is a supervised learning framework where the inputs are texts, and the output is desired characteristics of these texts:
  - Sentiment Classifier: Text -> Sentiment Score
  - Review Classification: Text -> Review Problem
  - Machine Translation: Text -> Other language

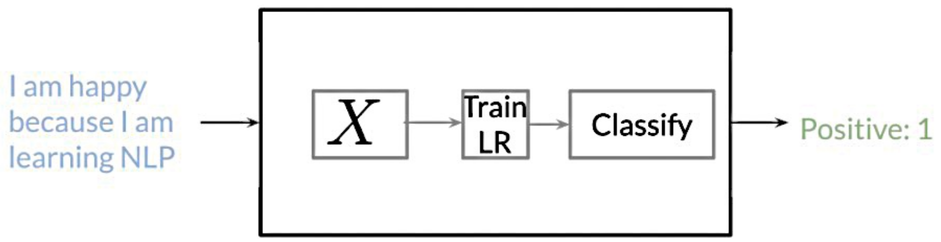
3

3

Classic NLP Framework

- Reference: https://www.coursera.org/specializations/natural-language-processing https://web.stanford.edu/~jurafsky/slp3/

- A classic NLP framework usually contains 2 parts:
  - Pre-processing: Text -> Numeric representations
  - Classification: Numeric representations -> outcome



4

4

NLP Roadmap

Attention is **all you need**
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …
Save Cite Cited by 36430 Related articles All 35 versions

Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… We introduce **BERT** and its detailed implementation in this … For finetuning, the **BERT** model is first initialized with the pre-trained … , we present **BERT** fine-tuning results on 11 **NLP** tasks. …
Save Cite Cited by 34042 Related articles All 41 versions

Probability & Vector Representations of NLP → Word2vec & Seq2seq → Transformer-based Models → Large Language Models (LLM)

Before 2010 | 2013-2017 | 2018-2022 | 2023-now

Efficient estimation of **word representations** in **vector space**
T Mikolov, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org
… **vector** X = **vector**("biggest")−**vector**("big") + **vector**("small"). Then, we search in the **vector space** for the **word** … question (we discard the input question **words** during this search). When the …
Save Cite Cited by 27227 Related articles All 48 versions

[PDF] **Glove**: **Global vectors** for word representation
J Pennington, R Socher… - Proceedings of the 2014 …, 2014 - aclanthology.org
… We use our insights to construct a new model for word representation which we call **GloVe**, for **Global Vectors**, because the **global** corpus statistics are captured directly by the model. …
Save Cite Cited by 26265 Related articles All 34 versions

[PDF] Language models are unsupervised multitask learners
A Radford, J Wu, R Child, D Luan, D Amodei… - OpenAI blog, 2019 - persagen.com
Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on taskspecific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset-matching or exceeding
Save Cite Cited by 2834 Related articles All 16 versions

Training language models to follow instructions with human feedback
L Ouyang, J Wu, X Jiang, D Almeida… - Advances in …, 2022 - proceedings.neurips.cc
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of …
Save Cite Cited by 4049 Related articles All 10 versions

5

5

NLP Roadmap

- New DL architecture allows us to not only improve the performance of traditional tasks but also enables us to achieve new tasks – machine language generation, chatbot, etc.
- The fundamental ways of thinking about words and pre-processing are the same across eras.
- Old architecture has one important advantage---explicit probabilistic model of linguistics (and possibly utilities). This transparency in modeling makes these kinds of model more useful in more economics-driven research with ML.
- A Typical NLP Task:

This is natural language → Pre-processing → Representation Learning → Downstream Task → Topic: NLP

6

6

Traditional NLP in Econ

JOURNAL ARTICLE

Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*

Stephen Hansen, Michael McMahon, Andrea Prat

The Quarterly Journal of Economics, Volume 133, Issue 2, May 2018, Pages 801–870, https://doi.org/10.1093/qje/qjx045
Published: 31 October 2017

PDF | Split View | Cite | Permissions | Share

Abstract

How does transparency, a key feature of central bank design, affect monetary policy makers' deliberations? Theory predicts a positive discipline effect and negative conformity effect. We empirically explore these effects using a natural experiment in the Federal Open Market Committee in 1993 and computational linguistics algorithms. We first find large changes in communication patterns after transparency. We then propose a difference-in-differences approach inspired by the career concerns literature, and find evidence for both effects. Finally, we construct an influence measure that suggests the discipline effect dominates.

JEL: D78 - Positive Analysis of Policy Formulation and Implementation, E52 - Monetary Policy, E58 - Central Banks and Their Policies
Issue Section: Article

ECONOMETRICA
JOURNAL OF THE ECONOMETRIC SOCIETY

Original Articles | Full Access

Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech

Matthew Gentzkow, Jesse M. Shapiro, Matt Taddy

First published: 25 July 2019 | https://doi.org/10.3982/ECTA16566 | Citations: 148

Get it @ NYU

SECTIONS | PDF | TOOLS | SHARE

Abstract

We study the problem of measuring group differences in choices when the dimensionality of the choice set is large. We show that standard approaches suffer from a severe finite-sample bias, and we propose an estimator that applies recent advances in machine learning to address this bias. We apply this method to measure trends in the partisanship of congressional speech from 1873 to 2016, defining partisanship to be the ease with which an observer could infer a congressperson's party from a single utterance. Our estimates imply that partisanship is far greater in recent years than in the past, and that it increased sharply in the early 1990s after remaining low and relatively constant over the preceding century.

ECONOMETRICA
JOURNAL OF THE ECONOMETRIC SOCIETY

Full Access

What Drives Media Slant? Evidence From U.S. Daily Newspapers

Matthew Gentzkow, Jesse M. Shapiro

First published: 08 February 2010 | https://doi.org/10.3982/ECTA7195 | Citations: 861

Get it @ NYU

PDF | TOOLS | SHARE

Abstract

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.

Text as Data

Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

Text as Data†

Matthew Gentzkow, Bryan Kelly, and Matt Taddy*

An ever-increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications. (JEL C38, C55, L82, Z13)

0. Pre-processing;

1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$;

2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$; and

3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis.

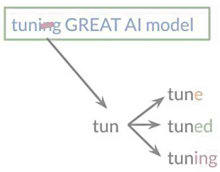
Agenda

- Natural Language Processing Framework
- Pre-processing
- Word Representation

9

Pre-processing

- References: https://nlp.stanford.edu/IR-book/pdf/02voc.pdf https://web.stanford.edu/~jurafsky/slp3/slides/2_TextProc_Mar_25_2021.pdf
- Text normalization: Transforming sentences into words.
- Text normalization includes 2 tasks: Word segmentation (i.e., tokenization) and word normalization:
  - **Elimination of non-words**: URL, HTML, handles, punctuations etc.
  - **Tokenization:** Parse strings into words.
  - **Stop-word removal**: Get rid of stop-words which are extremely common, such as "a, an, is, the, of…"
  - **Stemming**: Convert every word to its stem.
  - **Normalization**: Normalize accents and diacritics; change all letters into lower-cases.



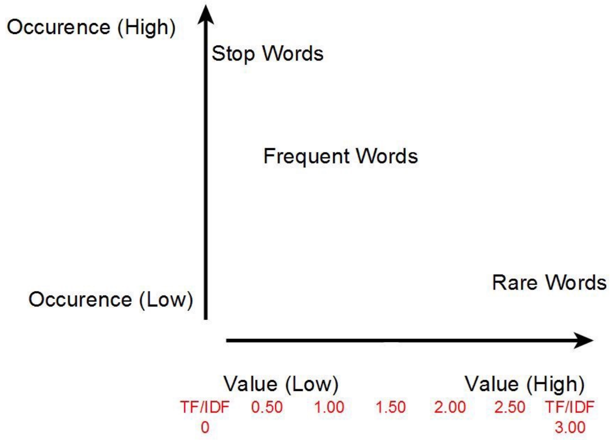
10

Term Frequency-Inverse Document Frequency

- Each word has different importance for a document/sentence.
- TF-IDF: A word appearing in **fewer documents** and appearing **more times** may be more important.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents



11

11

Agenda

- Natural Language Processing Framework
- Pre-processing
- Word Representation

12

12

Word Representation

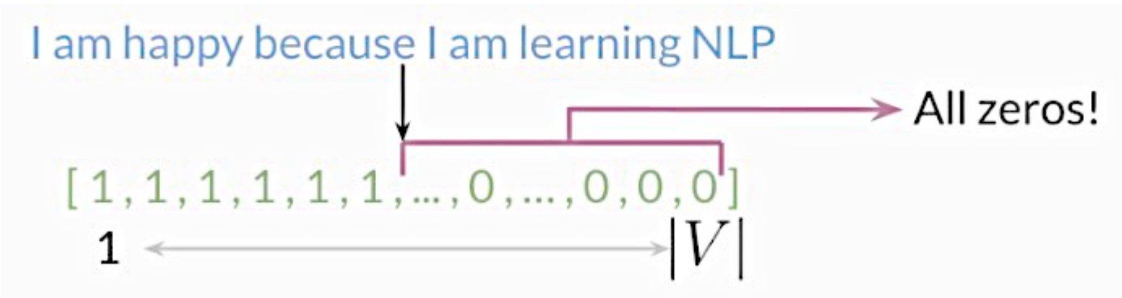
- With the vocabulary of words and word count, we can represent a sentence/document in different ways.
- **Frequentist** view: Represent words as vectors, which are low-dimensional projection of one-hot encoding of the words depending on its neighbors.
- **Bayesian** view: Represent words as probabilities; each word has a prior to be used and each sentence then has a conditional probability of words.

13

13

One-hot Encoding

- One-hot encoding: **Sparse representation**; think about the dummy variable in econometrics.
- You need k variables to represent a document if you have vocabulary length equal to k.



14

14

Low-Dimensional Dictionary

- Low dimensional dictionary representation: A vector whose length is the number of classes + 1.

Vocabulary	PosFreq (1)	NegFreq (0)
I	3	3
am	3	3
happy	2	0
because	1	0
learning	1	1
NLP	1	1
sad	0	2
not	0	1

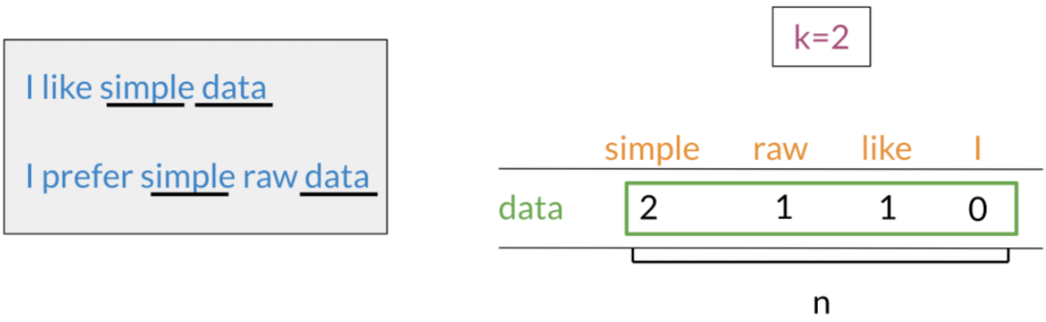
freqs: dictionary mapping from (word, class) to frequency

15

15

Low-Dimensional Neighbor Representation of Words

- You can use a words' neighbor words to represent a word.
- Obviously, this will take many unique words and the representation can be high-dimensional.



16

Low-Dimensional Document Representation of Words

- Reference: https://web.stanford.edu/~jurafsky/slp3/
- If you have multiple sets of documents and each one is different from others, you can use a word's occurrence in these document to represent a word's meaning.
- Basic idea: Similar words have similar vectors because they tend to occur in similar documents.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.5 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each word is represented as a row vector of length four.

17

Sentence Representation

- Add word vectors together.

Vocabulary	PosFreq (1)	NegFreq (0)
I	3	3
am	3	3
happy	2	0
because	1	0
learning	1	1
NLP	1	1
sad	0	2
not	0	1

freqs: dictionary mapping from (word, class) to frequency

- I am sad, because I am not learning NLP -> [x1, x2], x1 = ?, x2 = ?

18

Document Representation

- You can also use word occurrence to represent sentence and document.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

- This is called term-document matrix, allowing us to find similar documents.