

Want more?

`https://github.com/
matthewbrems/
nlp-fundamentals-python`

Introduction: Matt Brems (he/him)



Managing Partner & Principal Data Scientist, BetaVector

Distinguished Faculty, General Assembly

Marketing & Comms Director, Statistics Without Borders

Previously:

Growth + Computer Vision @ Roboflow

R&D Fairness/Bias in Data @ FINRA

Data Science Education @ General Assembly

Data Science @ Optimus Consulting

Enterprise Analytics @ Smucker's

M.S. Statistics @ The Ohio State University

Recommended Reads:

Data-Driven Thinking: "Factfulness"

Data Visualization: "Storytelling with Data"

Data Science: "Introduction to Statistical Learning with Applications in R" 3

Background Knowledge

- We assume no background in NLP.
- All code is written in Python, so experience is helpful. However, solutions are provided so a Python background is not required.
- Some experience with machine learning would make this workshop easier to follow, but is not necessary.

<https://github.com/matthewbrems/nlp-fundamentals-python>

Learning Objectives

- Clean text data with tokenization, base Python functions, and regular expressions.
- Learn lemmatizing and stemming.
- Transform data with `CountVectorizer` and `TFIDFVectorizer`.
- Fit machine learning models in `scikit-learn` and evaluate their performance.
- Build pipelines and `GridSearch` over NLP hyperparameters.

Run of Show

Introduction to Natural Language Processing (NLP)

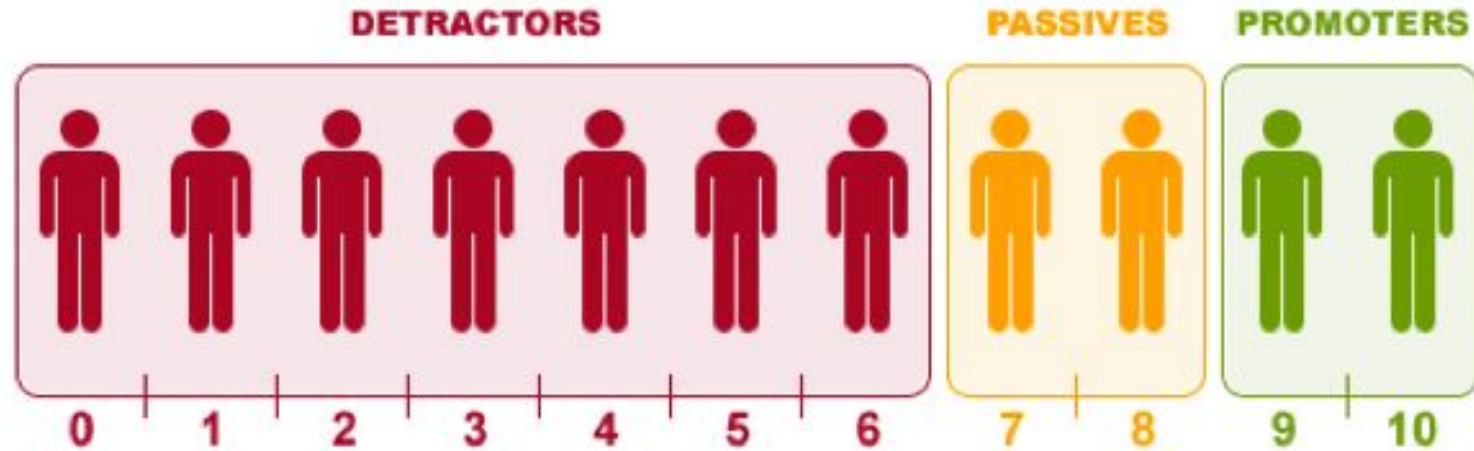
Cleaning Text Data

Converting Text Data to Model Features

Hyperparameters in NLP

Machine Learning with Pipelines in NLP

What is Natural Language Processing?



Net Promoter Score

=

% Promoters

-

% Detractors

What is Natural Language Processing?

```
model.doesnt_match(['angioplasty', 'appendectomy', 'cabg', 'bronchoscopy'])  
  
'appendectomy'
```


What is Natural Language Processing?

English - detected ▼



↔



German ▼

Can you please help me find the bathroom?

×

Können Sie mir bitte helfen, das Badezimmer zu finden?

Open in Google Translate

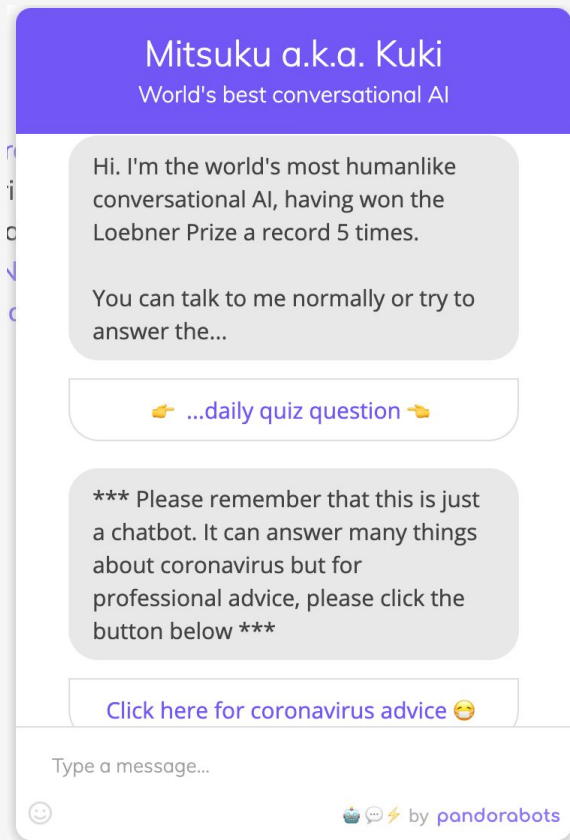
Feedback

translate.google.com ▼

Google Translate



What is Natural Language Processing?



Our goals with Natural Language Processing

1. **Our broad goal** with natural language processing is to get computers to understand language more like how humans understand language.
2. **Our more specific goal** with natural language processing in traditional machine learning is to convert our semi-structured text data into a dataframe of real numbers.
 - X is our input data.
 - Y is our output data.

WARNING: bias in NLP!

Analysis with NLP can only be as good as the data you provide it.

- If your data are biased, then your results will be biased.
- If your data are not biased... you're probably wrong.

WARNING: bias in NLP!

Word Embedding

- Paris is to France as Tokyo is to _____.

WARNING: bias in NLP!

Word Embedding

- Paris is to France as Tokyo is to Japan.
- Man is to king as woman is to _____.

WARNING: bias in NLP!

Word Embedding

- Paris is to France as Tokyo is to Japan.
- Man is to king as woman is to queen.
- Man is to computer programmer as woman is to _____.

WARNING: bias in NLP!

Word Embedding

- Paris is to France as Tokyo is to Japan.
- Man is to king as woman is to queen.
- Man is to computer programmer as woman is to homemaker.

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

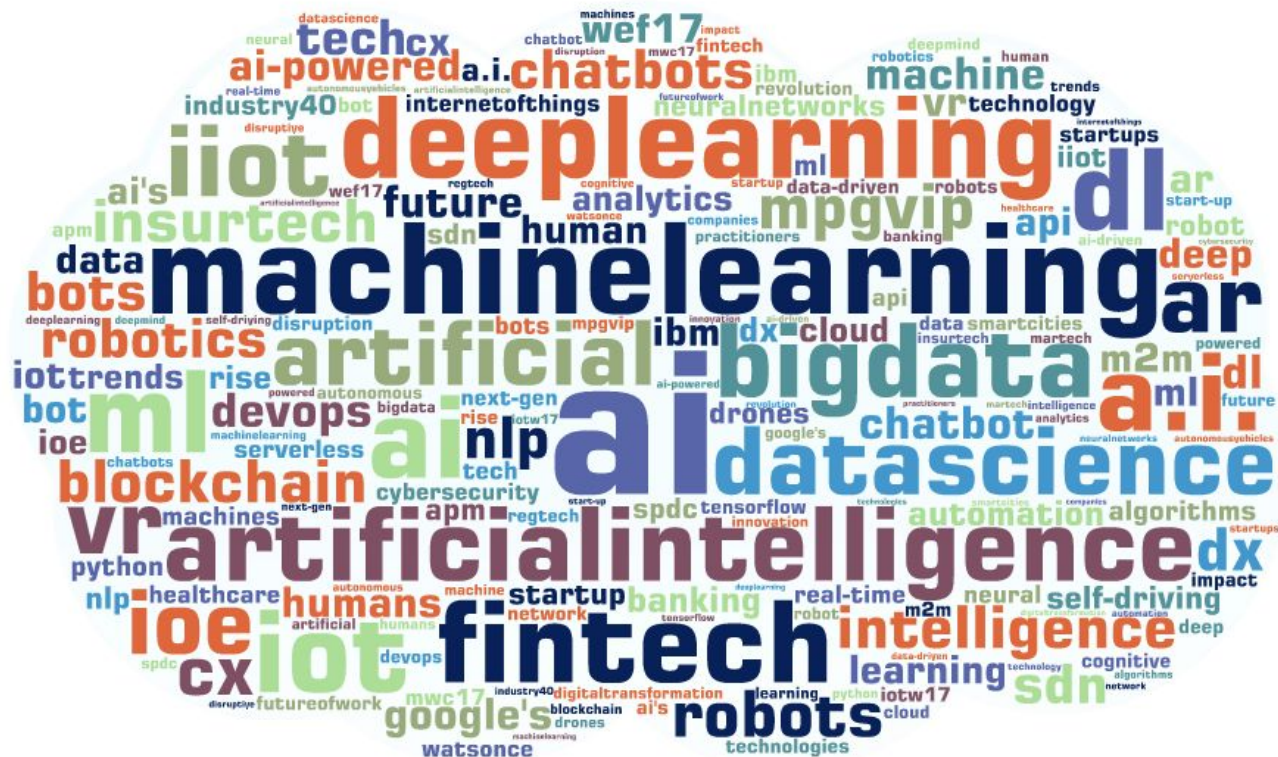
<https://arxiv.org/pdf/1607.06520.pdf>

WARNING: bias in NLP!

Ethical and Social Issues in Natural Language Processing @ Stanford

<https://web.stanford.edu/class/cs384/>

WARNING: word clouds are not data science!



Jeopardy Primer

- Round 1: Jeopardy!
 - Five clues in six categories.
 - Dollar amounts range from \$200 to \$1,000.
- Round 2: Double Jeopardy!
 - Five clues in six categories.
 - Dollar amounts range from \$400 to \$2,000.
- Round 3: Final Jeopardy!
 - One category, one question, dollar amount is a wager.



To the notebook!

`starter-code.ipynb`
`solution-code.ipynb`

Thank you!



LinkedIn: Matthew Brems

Twitter: @MatthewBrems

Github: MatthewBrems

Email: matt@betavector.com