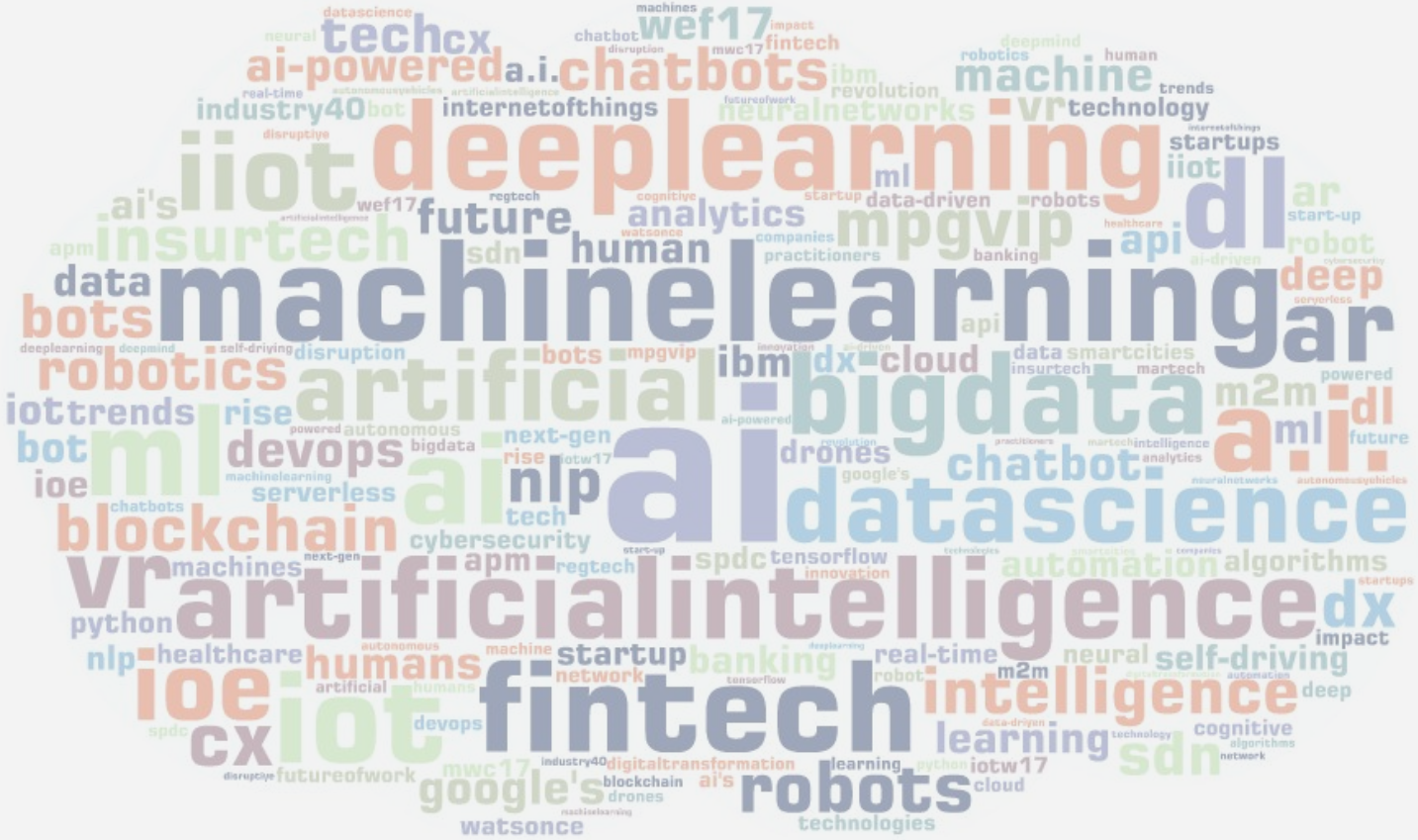


Natural Language Processing Fundamentals in Python

Matt Brems

Introduction: Matt Brems (he/him)

Growth Manager, Roboflow

Managing Partner, BetaVector

Marketing & Comms Director, Statistics Without Borders

Previously:

- Senior R&D Data Scientist, FINRA
- Lead Data Science Instructor, General Assembly
- Data Science Fellow, 0ptimus 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: "An Introduction to Statistical Learning with Applications in R"

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 regular expressions and tokenization.
- Learn lemmatizing and stemming, including how and when to use these techniques.
- 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

Module 1: Introduction to Natural Language Processing (NLP)

Module 2: Cleaning Text Data

Module 3: Converting Text Data to Model Features

Module 4: Hyperparameters in NLP

Module 5: Machine Learning with Pipelines in NLP

Run of Show

Module 1: Introduction to Natural Language Processing (NLP)

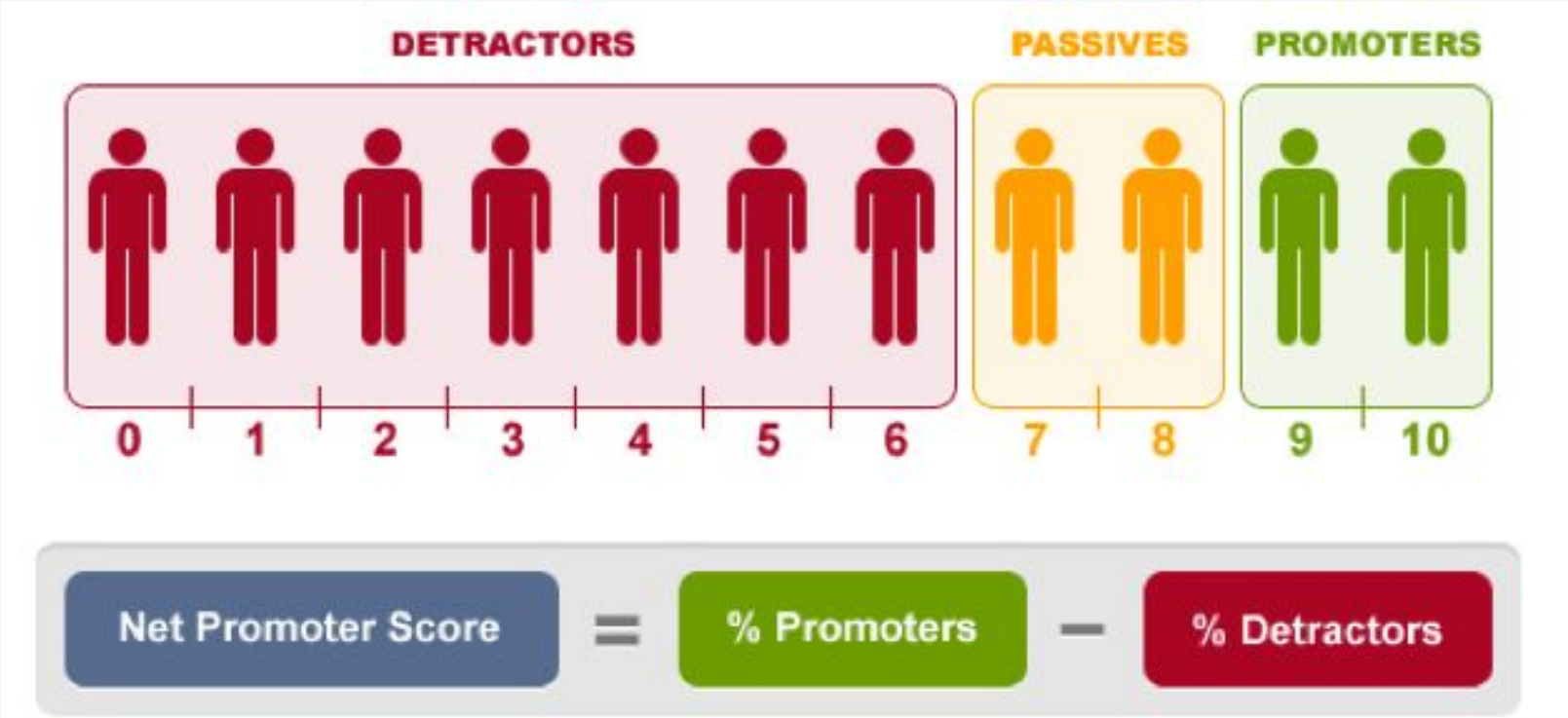
Module 2: Cleaning Text Data

Module 3: Converting Text Data to Model Features

Module 4: Hyperparameters in NLP

Module 5: Machine Learning with Pipelines in NLP

What is Natural Language Processing?

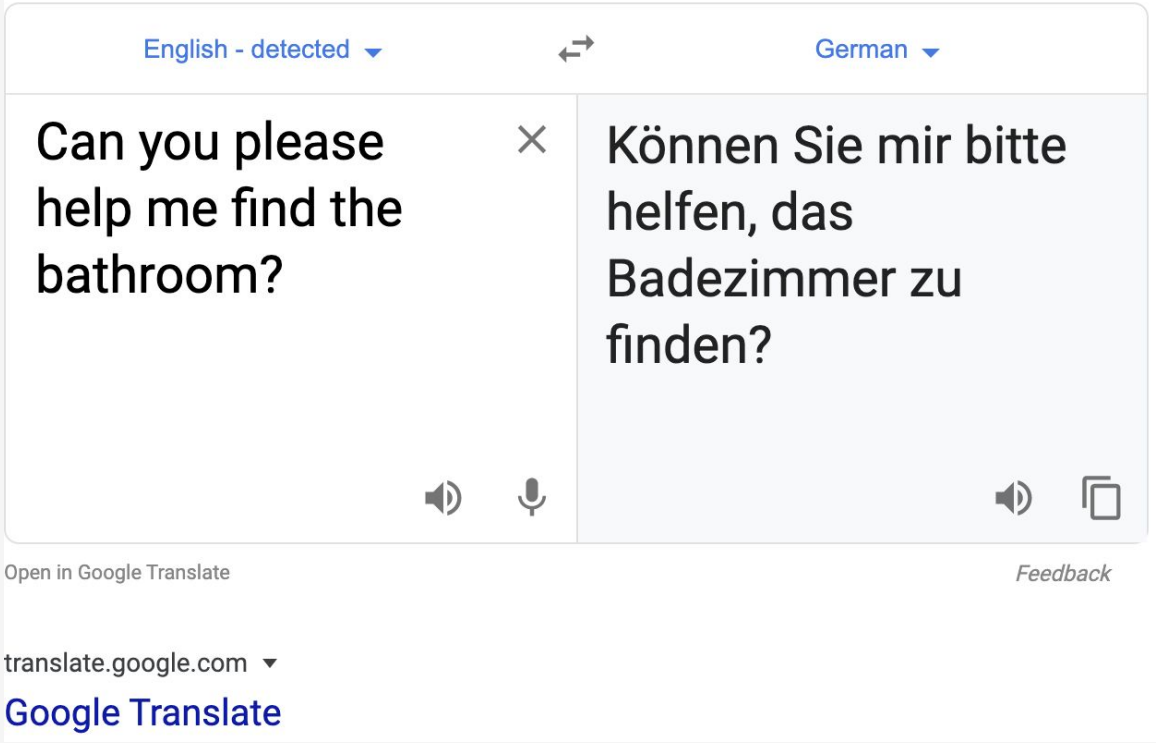


What is Natural Language Processing?

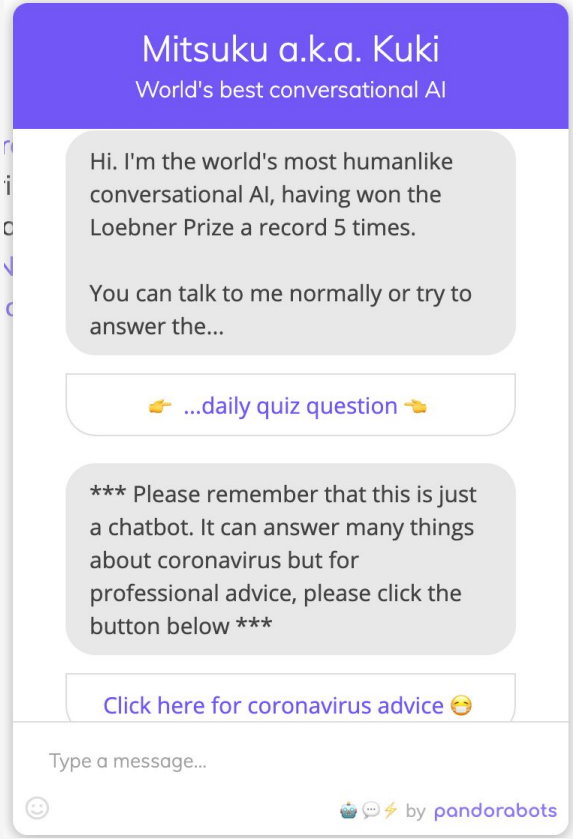
```
model.doesnt_match(['angioplasty', 'appendectomy', 'cabg', 'bronchoscopy'])
```

```
'appendectomy'
```

What is Natural Language Processing?



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

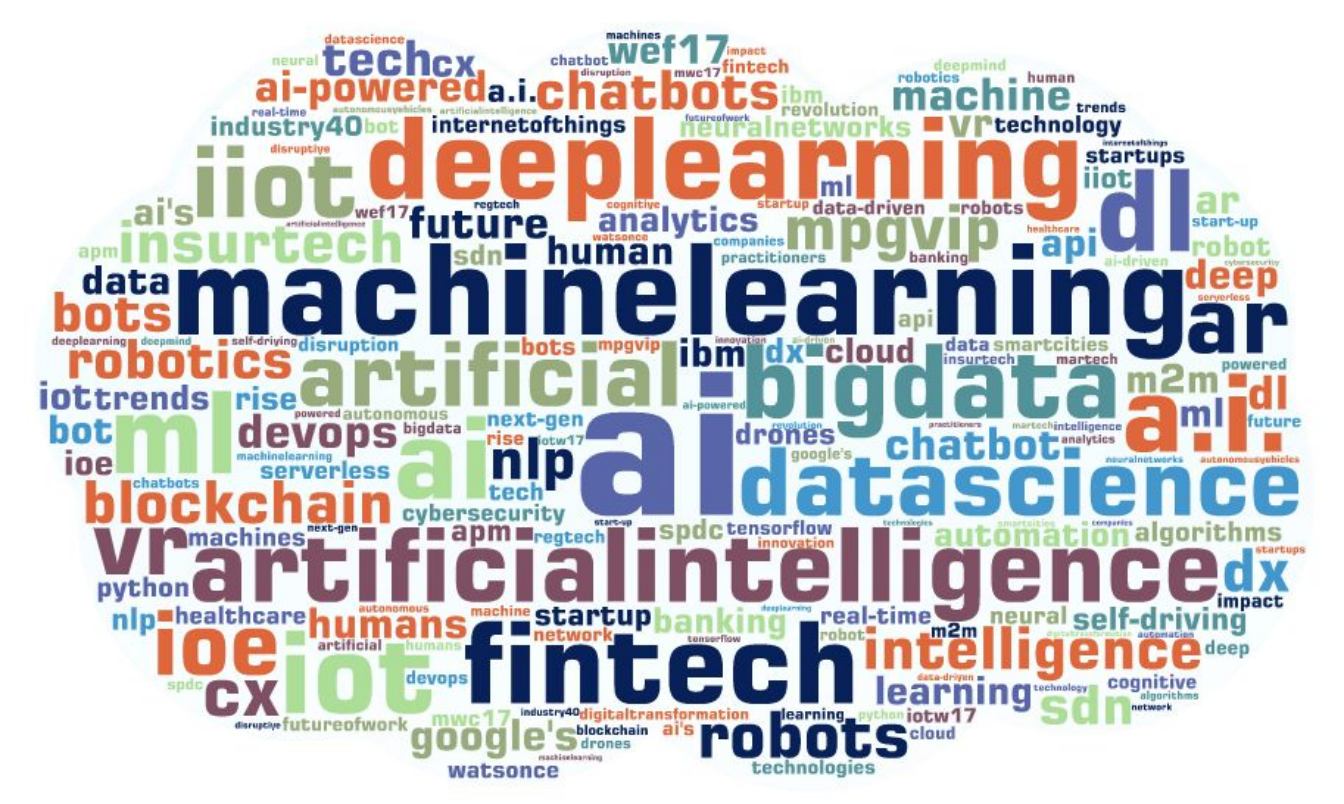
https://arxiv.org/pdf/1607.06520.pdf

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!



https://medium.com/hackernoon/what-real-fake-news-says-about-obamas-presidency-4bf42be71ff1

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!

THANK YOU

LinkedIn: Matthew Brems

Twitter: @matthewbrems

Github: matthewbrems

Email: matt@roboflow.com