General Assembly
Data Science Immersive Capstone:

Philosophical Factors
for Natural Language Processing

Thomas W. Ludlow, Jr.
February 12, 2019

Summary Topics

1. Philosophy and Technology
2. Project Approach
3. Source Data
4. Natural Language Processing Techniques
5. Classification Models
6. Results and Observations
7. Questions

Philosophy & AI

"Philosophy will be the key that unlocks artificial intelligence"
David Deutsch (The Guardian, 2012)

https://www.theguardian.com/science/2012/oct/03/philosophy-artificial-intelligence

Philosophy & AI

- Questions faced by Data Scientists and Artificial General Intelligence (AGI) Engineers are similar to those pondered by philosophers - *what does it mean to be self-aware?*

- Some argue that AGI is not possible without a body and/or soul

- Measurements of progress in AGI development are philosophical
  - *Turing Test* - Can we tell whether it's a machine?
  - *Mirror Test* - Can a machine infer facts about itself?

Philosophy as a Data Science Problem

How well can Machine Learning models recognize the ideas of historic philosophers?

Goal: Build a Natural Language Processing (NLP) model that identifies the philosophical "factors" for any text input using predictive similarity.

Project Approach

1. Obtain **philosophy texts** covering a wide historical range
2. Use **Topic Modeling** to identify distinct subjects in the texts
3. Apply **Word Vectors** to infer semantic meaning
4. Compare similarity to new texts using **Multi-class Classifier Modeling**

Source Data

- **Project Gutenberg**
  www.gutenberg.org
  Repository of 58,000+ free eBooks, in various file formats including *.txt*. Most selections for corpus from the Philosophy Bookshelf

- **ETEXT**
  www.textfiles.com/etext
  Collection of free *.txt* files for Fiction, Non-fiction, Reference, and By Author

Source Data

- Full corpus contains 30 full works, representing 20 authors
- Additional test corpus contains partial works from 17 of the corpus authors

- Selections include:

The Art of War Sun Tzu, 500 BC

The Republic Plato, 380 BC

The Rubaiyat
Omar Khayyam, 1120

The Prince
Niccolo Machiavelli, 1532

The Principles of Philosophy
Rene Descartes, 1644

Common Sense Thomas Paine, 1776

Also Sprach Zarathustra
Friedrich Nietzsche, 1891

The Analysis of Mind
Bertrand Russell, 1921

Source Data

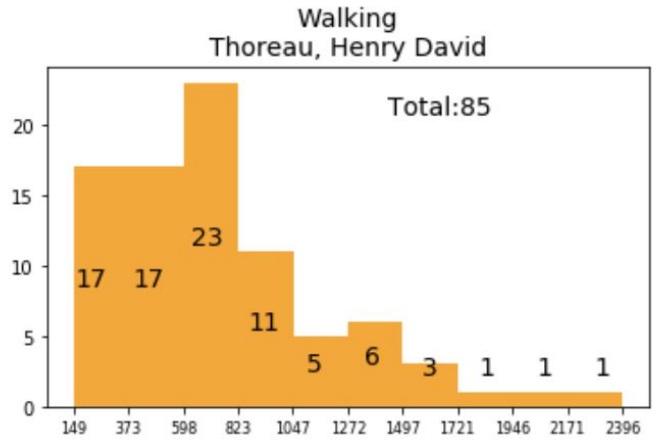
Text Loading

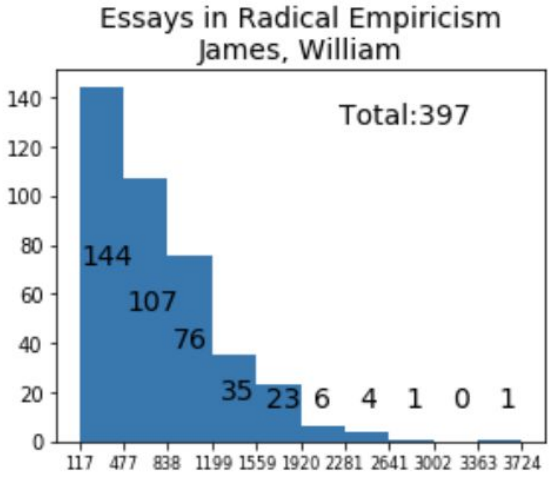
- Read text files line-by-line
- Remove text metadata
- Count the number and size of lines of text to determine paragraph boundaries
- Corpora
  - 16,060 paragraphs
  - 70,922 sentences
  - **1,887,382 words**

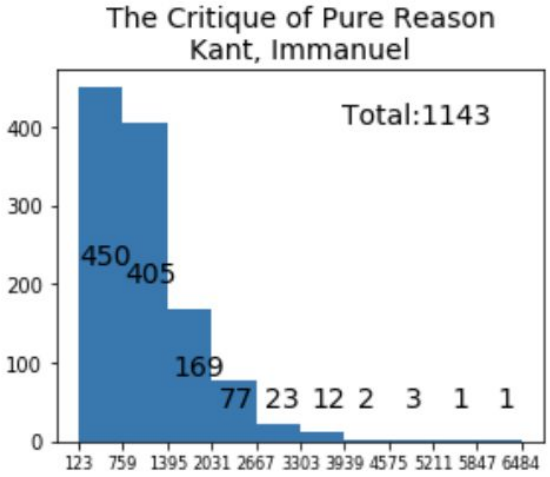
Source Data

Class Balancing

- Models work best when each class (author) has similar sized training data.
- Selected the ***minimum paragraph length*** based on balance.







Orange: ***Testing Set*** *Work Distribution*

Blue: ***Training Set*** *Work Distributions*

NLP Techniques

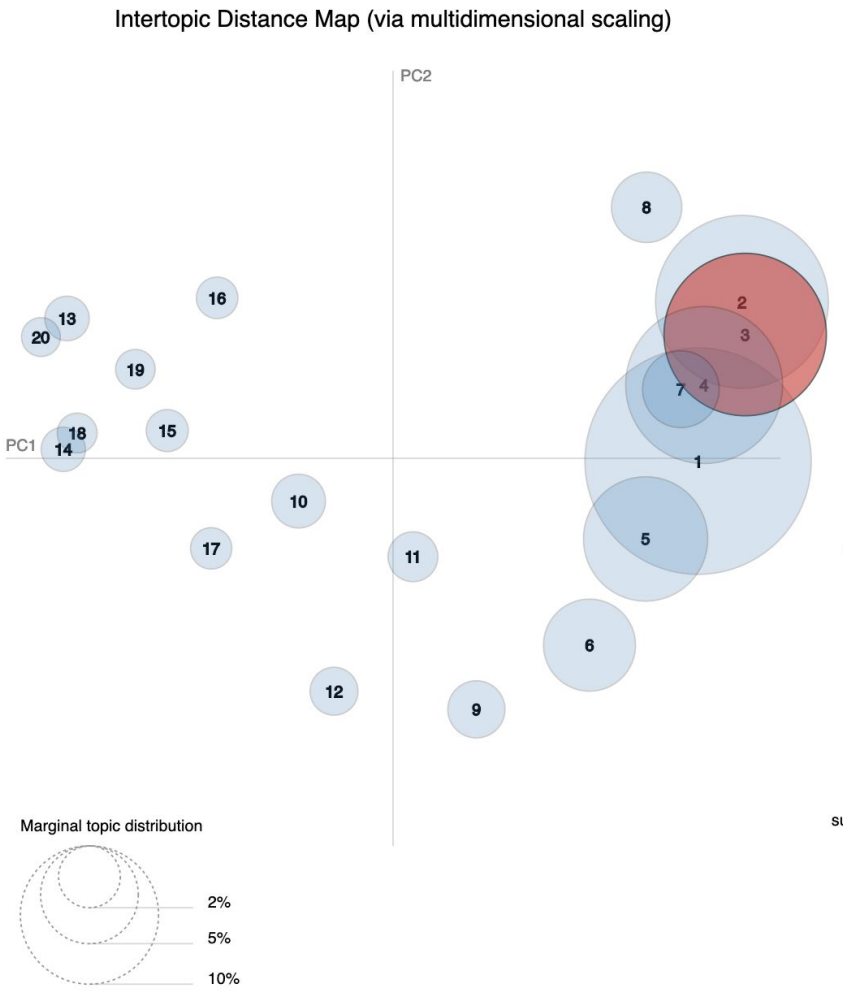
- **Python Tools and Libraries**
  - spaCy
  - Gensim
- **Natural Language Preprocessing**
  - Lemmatizing
  - Stop word removal
  - Roman numerals

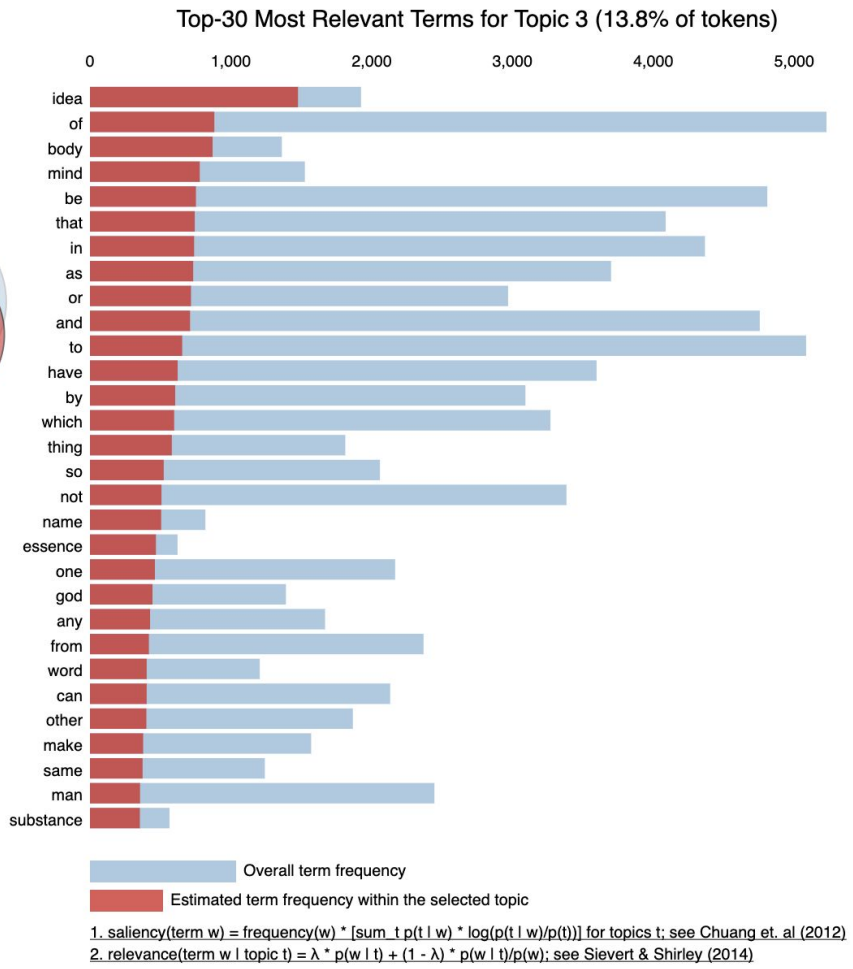
NLP Techniques: Topic Modeling

- **Latent Dirichlet Allocation (LDA)**
  - Organizes text into *n*-topics
  - Uses Bayesian Inference to determine topic probability

- **Document Vectors (Doc2Vec)**
  - Determines mathematical relationships between words in text
  - Enables semantic understanding of meaning of words, sentences, paragraphs

NLP Techniques: Topic Modeling





Multiclass Classification Models

Recurrent Neural Net

- Front-Feeding Sequential RNN (Keras)
- Versatility
- Weights to balance classes
- Tested on smaller corpus
- **Training Accuracy: 62.5%**
- **Testing Accuracy: 12.9%**

Results

Sample Query Output on Training Corpus "Bumper Stickers"

```
--------------------------------------------------------------------------------
Doc #: 28
Target Author: Nietzsche, Friedrich
Text:
 All ideations take place from particular perspectives

Philosophical Factors:

        Plato                                0.628
        Nietzsche, Friedrich                 0.065
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
Doc #: 34
Target Author: Rousseau, Jean-Jacques
Text:
 The people are sovereign in forming a political community to address problems in a commercial society

Philosophical Factors:

        Rousseau, Jean-Jacques               0.202
        Plato                                0.182
        Paine, Thomas                        0.151
        Sun Tzu                              0.119
```

Observations

Applications

- Literature
- Feature Engineering

Next Steps

- Classifier Models
- Support Vector Machine (SVM)
- Add bigrams, trigrams to corpus
- Longer modeling runs

Limitations

- GPU Processing speed
  - Preprocessing: 2 hours
  - LDA: 3 hours
  - Modeling Epochs
- Semantic interpretation
  - Tough to measure the depth of recognition / representation of document “meaning”

Questions?

What is mind? No matter.
What is matter? Never mind.

George Berkeley 1685-1753
(Subjective Idealist)