# Natural Language Processing

Data Science Immersive

FLATIRON SCHOOL

# Agenda:

- Introduction / History
- Basic Theory
- Tokens / Pre-processing
- Stems vs Lems
- Stop words
- Bag of Words
- TF-IDF

# Natural Language Processing

- What is natural language?

> ***“Any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation.”***
>
> - *Wikipedia*

# Brief History

- NLP generally started in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.

- Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing.

- In the 2010s, deep neural network-style machine learning methods became widespread in NLP, due in part to a flurry of results showing that such techniques can achieve state-of-the-art results in many natural language tasks.

# Natural Language Processing

- NLP is an area of machine learning focused on how to program computers for the processing and analyzation of large amounts of natural language data.

- With NLP, computers have the ability to understand, analyze, manipulate, and potentially generate human language.

  - **Final project idea?... Maybe you!**
  - https://github.com/AustinKrause/Mod_5_Text_Summarizer
  - https://github.com/jeussantiago/Your-Next-Book

# Some Data Science Use Cases

- Text classification
  - Is this email Spam or Ham?
- Topic Modelling
  - Unsupervised Clustering
- Sentiment analysis
  - Brand Monitoring
  - Product Analysis
- Voice Assistants
  - Siri
  - Alexa
  - etc...

# Quiz:

Arrange these terms in order from smallest content to largest:

Document, Bi-gram, Corpus, Sentence, Word, Paragraph, Corpora, N-gram

# Quiz Answers:

Arrange these terms in order from smallest content to largest:

Document, B-gram, Corpus, Sentence, Word, Paragraph, Corpora, N-gram

1) Word
2) Bi-gram
3) N-gram *(can be bi-gram)*
4) Sentence
5) Paragraph
6) Document
7) Corpus
8) Corpora

# Parsing and Understanding

What to compare?

- Tokens
  - Individual “words” (can be full sentences)
  - Bi-grams = 2 word combinations
  - N-grams = combinations of “N” number of words
- Documents
  - Collections of sentences/paragraphs
- Corpus
  - A collection of documents
- Corpora
  - More than one corpus

# Toolkits

- Natural Language Toolkit (NLTK)
- Textblob
- SpaCy
- GenSim

- Many more!

https://medium.com/activewizards-machine-learning-company/comparison-of-top-6-python-nlp-libraries-c4ce160237eb
https://elitedatascience.com/python-nlp-libraries

# Inputs for a normal machine learning model.

| sepal length | sepal width | petal length | petal width | class |
|---|---|---|---|---|
| 6.3 | 2.9 | 5.6 | 1.8 | Iris-virginica |
| 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| 5.6 | 2.9 | 3.6 | 1.3 | Iris-versicolor |
| 6.0 | 2.7 | 5.1 | 1.6 | Iris-versicolor |
| 7.2 | 3.6 | 6.1 | 2.5 | Iris-virginica |

# One-hot encoding for categorical data

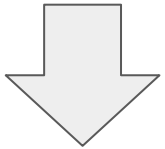
| sex | one-hot encoding | sex_female | sex_male |
|---|---|---|---|
| female | → | 1 | 0 |
| male | → | 0 | 1 |
| female | → | 1 | 0 |
| male | → | 0 | 1 |
| female | → | 1 | 0 |
| ... | ... | ... | ... |

# Dealing with text

| Doc_id | Document | Classification |
|---|---|---|
| 1 | “The impeachment trial in the Senate will …” | politics |
| 2 | “As of this afternoon, WeWork employees will no longer... | not-politics |
| 3 | “The Democratic Presidential debate will take place…” | politics |
| 4 | “Boeing is considering curbing production of the 737 Max...” | not-politics |

# Turning text data into a vector of numbers

| Doc_id | Document | Classification |
|---|---|---|
| 1 | “The impeachment trial in the Senate will …” | politics |
| 2 | “As of this afternoon, WeWork employees will no longer... | not-politics |



| Doc_id | The | impeachment | trial | ... | employees | will | Politics |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | | 1 | 1 | 1 |
| 2 | 0 | 0 | 0 | | 1 | 1 | 0 |

# Comparing sentences

Create a term frequency matrix for the following sentences:

1. “He can talk to the teacher this afternoon.”
2. “This afternoon, he can talk to a teacher.”

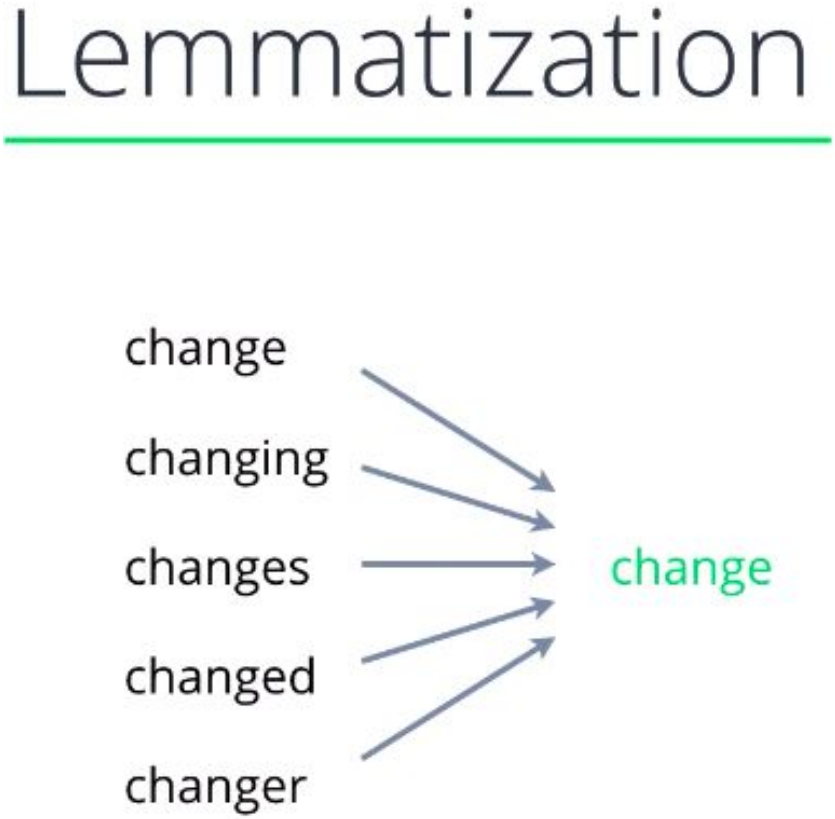
# Stemming

- Stemming is the process of reducing inflection in words to their root forms, such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.

- Stems are created by removing the suffixes or prefixes used with a word.

# Lemmatization

- Unlike Stemming, Lemmatization reduces the inflected words properly ensuring that the root word belongs to the language.
- In Lemmatization, the root word is called Lemma.
- A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.



# Stemming and Lemmatization Visualized

| | original_word | stemmed_word |
|---|---|---|
| 0 | trouble | troubl |
| 1 | troubled | troubl |
| 2 | troubles | troubl |
| 3 | troublemsome | troublemsom |

| | original_word | lemmatized_word |
|---|---|---|
| 0 | trouble | trouble |
| 1 | troubling | trouble |
| 2 | troubled | trouble |
| 3 | troubles | trouble |

# STOP words / “Punctuation?!”

- Stop Words:
  Words which are filtered out before or after processing of natural language data. Though "stop words" usually refers to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and not all tools even use such a list.

- Punctuation:
  Treat these characters similar to stop words. Can be removed in several different ways.

# Bag of Words

- Bag of Words: (BoW for short) is a way of extracting features from text for use in modeling.

- A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:
  - A vocabulary of known words.
  - A measure of the presence of known words.

- Called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether it knows if a word occurs in the document, not **where** in the document.

# Term Importance

Question: When comparing the similarity of the documents below, which are most similar?

- “Apple juice concentrate”
- “Orange juice concentrate”
- “Lemon powder mix”
- “Orange powder mix”
- “Cranberry powder mix”
- “Lemon juice concentrate”

# Which words are important?

- Term Frequency & Inverse Document Frequency (tf & idf)

Term Frequency

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Inverse Document Frequency

$$idf(w) = \log \frac{N}{df_t}$$

# TF-IDF

Applying Term Frequency-Inverse Document Frequency (tf-idf):

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

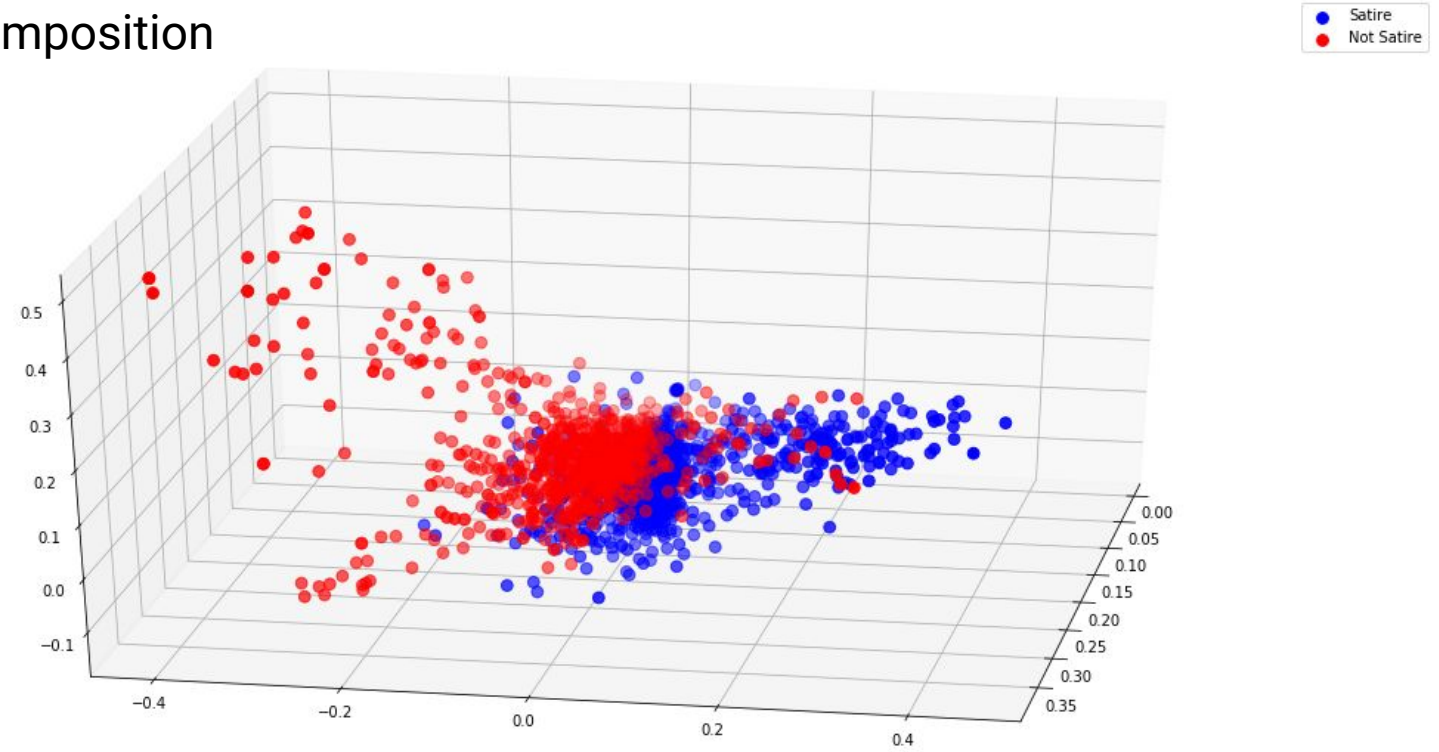
$tf_{i,j}$ = number of occurrences of $i$ in $j$

$df_i$ = number of documents containing $i$

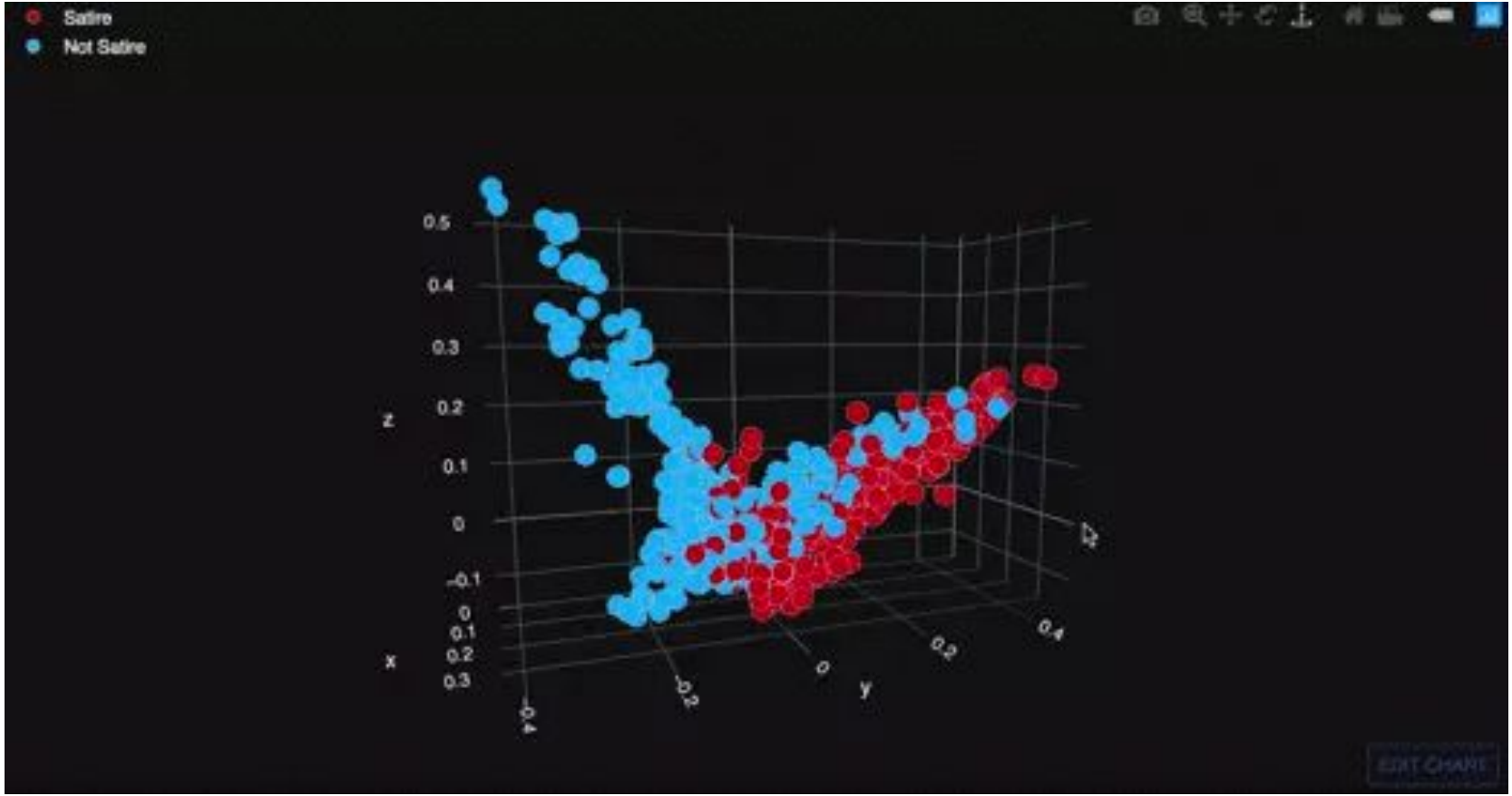
$N$ = total number of documents

# Visualizing Your Tf-idf Matrix

Singular Value Decomposition
(aka: SVD)



# Visualizing Your Tf-idf Matrix (Continued)



# Where to start...

We’ve some got options! (is there ever just one method?)

- Rule Based Approach
  - Regular expressions

- "Traditional" Machine Learning
  - Probabilistic modeling, likelihood maximization, and linear classifiers.
  - “You shall know a word by the company it keeps” - J. R. Firth 1957

- Deep Learning
  - Recurrent Neural Networks
    - LSTM (Long short-term memory)
    - https://skymind.ai/wiki/lstm

# Traditional vs State of the Art

Questions to ponder:

- Why use "traditional" machine learning (or rule-based) approaches for NLP?

- Why use deep learning over "traditional" machine learning?

# Closing Thoughts

Natural Language constantly changes…

**...so does Natural Language Processing!**

Some Resources:

https://www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html
https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e
https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b
https://machinelearningmastery.com/gentle-introduction-bag-words-model/
https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908
https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/
https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/