Hacker Hour

Natural Language Processing Basics

Archi Parekh

What is NLP?

- Branch of AI that focuses on having computers interpret language the same way humans do
- Mix of linguistics, statistics, and machine learning
- Some applications include translation, speech-to-text software, and chatbots

Text as Data

- Text is inherently qualitative data
- Quantitative analysis can yield unexpected insights

Preprocessing

- Super important step!
- Manipulate unstructured data into a consistent format
- Includes
  - Removing punctuation, converting to lowercase
  - Tokenization
  - Stemming
  - Removing common words

Word Frequency

- First way you can learn something about the text is calculating how many times each word appears
- Zipf’s Law: “rank-frequency distribution has an inverse relation”
    - The frequency of a word and its rank in the frequency distribution (like a list of most popular words) have an inverse relationship

Text Representations

Bag of Words

Represent text as a collection of all individual words

Pays no attention to context of a word in the text

N-gram

Represent text as a group of *N* consecutive words

Provides more context

Now what if we want to compare texts?

TF-IDF

- Term frequency inverse document frequency
- Used to see which words are most relevant to a given document in a corpus of documents
- **Term frequency:** # of times word *w* occurs in the document
- **Inverse document frequency:** Logarithm of the number of total documents divided by the number of documents containing word *w*
- **Tf-idf = tf * idf**

“Computer science is cool”

“The computer ate a mouse”

Idf = log (2/2) = 0

Tfidf = 0

Cosine Similarity

Use term frequencies to see how similar two documents are

Can be calculated by normalizing columns of a matrix and then multiplying it by its transpose

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \, ||\mathbf{B}||}$$

Cool applications

- Sentiment analysis
- Topic modeling
- Word embeddings
- And much more!

Topic Modeling

- Unsupervised machine learning method that groups documents into a specified number of topics
- Uses LDA (latent dirichlet allocation)
- Complex linear algebra
- Luckily we can use topic modeling library in R!

Sources

Lectures from Intro to Data Science (198:439) and Computational Social Science (920:360)

https://towardsdatascience.com/x%E1%B5%80x-covariance-correlation-and-cosine-matrices-d2230997fb7

https://www.tidytextmining.com