Natural Language Processing

Austin Buller & Jacob Kinser

What is NLP

- Helping computers understand and analyze text like humans, as well manipulate the incoming data and potentially generate human language.

Use Cases

- Virtual chatbots
- Spam detection
- Text summarization
- Text predictions/autocorrect

Process from Start to Finish

1. Read in raw text
2. Clean text and tokenize it
3. Fit to simple model
4. Final model selection

Breaking down text

- Removing punctuation and stop words
- Tokenization
  - Separating words into individual tokens
- Stemming vs lemmatization
  - Stemming
    - Uses the stem of the word
    - Less costly, but less accurate
  - Lemmatization
    - Takes the context that the word is used in
    - More costly, but more accurate

Vectorization

- Vectorization
  - A numeric representation of text
  - Encodes text as integers to create feature vectors
- Feature Vector
  - An n-dimensional vector of numerical features that represent some object
- Document term matrix
  - The table that holds the frequency data of the text after tokenization

Machine Learning

- Ensemble Method: technique that combines several week models into on strong model
  - Random Forest: Many week decision trees are made and combines into bigger and stronger trees
- Gradient Boosting: learning method based on focusing on mistakes, and improving on them