Basics of NLP

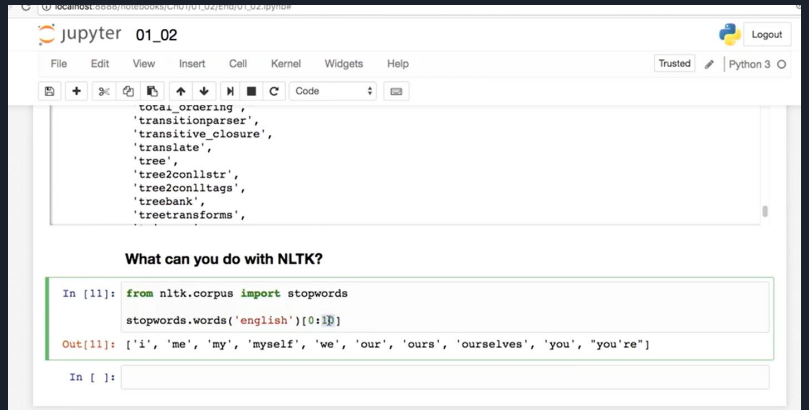
Kelly Jacobson, William Sengstock, Zach Witte

What is NLP?

- **Natural Language Processing** is the field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language.
- Real examples:
  - Spam filter
  - Auto-complete
  - Auto-correct
- Python does not see words. It only sees a string of characters, so it is our job to make sure Python understands what it is seeing.
  - Raw text needs to be converted to numbers
- Plenty of tools to help with this

NLTK

- Natural Language Toolkit is the most utilized package for handling NLP tasks in Python
- Once installed, can utilize multiple packages
  - (Ex) Command, “from nltk.corpus import stopwords”
    - Displays words that are used frequently, but don’t contribute much to a sentence’s meaning
- Includes libraries for tokenization, parsing, classification, stemming, tagging, and semantic reasoning
- LinkedIn videos goes into NLTK and how to use functions within it
- Jupyter Notebook
  - Works with NLTK, helps show visualization with the code being run



Machine Learning Pipeline

1. Have raw text - Model can't distinguish words, doesn't know what it's looking at
2. Tokenize - Tell the model what to look at, individual pieces of relevant information
3. Clean text - Remove stop words/Punctuation, stemming etc. Transforming the text into something easy to process, focus only on keywords
4. Vectorize - Convert to numeric form that a computer can understand
5. Machine learning algorithm - fit/train model to identify related words and understand context

Pre-processing/Tokenization

- Tokenization - break up sentences into words, words into characters
  - split into words, usually use whitespace as a delimiter
  - subword tokens use pieces of words to identify other words
    - Ex: “smart” is a subword in “smarter”
  - turn characters into tokens
  - the most frequently occurring words build your vocabulary
- NLTK has libraries for tokenization of sentences, words, characters, whitespace, and punctuation
- Normalization - convert all characters into a standard format that is easy to process
  - all lowercase, remove stop words, remove extra whitespace, remove special characters, etc.

Vectorizing Data

- Process of converting text to a form that Python and the machine can understand
- Encoding texts as integers to make feature vectors
  - Feature vectors: n-dimensional vector of numerical features that represent an object
  - In other words, taking a text message and converting it to a numerical vector that represents that text message
- Document Term Matrix
  - Gives numeric values of what the count is for each word in a sentence
  - Can be used to filter out spam
- Different Types
  - Count vectoring
  - N-gram vectoring
  - Inverse Document Frequency weighting

Machine Learning Algorithm

- Model will try and learn the relationship between words and labels
- Supervised Learning
  - Task of learning a function that makes predictions on unseen data
    - (Ex) Deciding whether an email is spam or not
- Unsupervised Learning
  - Deriving structure or patterns from the data when no data is given, self-discovery
    - (Ex) Grouping emails
- Random Forest Model: Makes multiple decision trees, combines predictions from all to produce final prediction
- Gradient Boosting: Iterative approach to combining weak learns into strong. Focus on mistakes