**CSC 417 Unit 3 Day 3 Outline**

1. Natural Language Processing
   1. Sentiment Analysis
      1. Determine ratio of positive to negative engagements (interactions) with a topic in a large body of text (e.g. are people reacting positively or negatively?)
         1. User reviews
         2. Tweets
      2. Process
         1. Preprocessing
            1. Tokenizing

Break down chunks of text into smaller pieces (sentences or words)

Also captures non-word strings – must decide how to handle (often removed from analysis)

* + - * 1. Remove “stop words”

Words that appear so frequently they throw off analysis (“and”, “the”, etc.)

Reduces volume of text to be analyzed

Can improve accuracy

* + - * 1. Normalize words

Condense all forms of a word into a single representation (e.g. count “perfect” and “perfectly” as the same word)

Stemming

Cut off word at stem

Will miss some relationships (e.g. see, saw)

Lemmatization

Relates all forms of a word to lemma (simplest form) (e.g. better -> good)

* + - * 1. Vectorize text

Transform tokens into vectors (numeric arrays that represent features)

Note text is processed as a numeric entity

* + - 1. Processing
         1. Create frequency distribution

Determine how often each word (stem or lemma) appears in the text

* + - * 1. Extracting concordance and collocation

Concordance = collection of word locations and context

Frequency

Position in text

Surrounding words

Collocation = sequence of words that frequently appear together

* + - * 1. Run sentiment analyzer

Categorizes text according to sentiment expressed (positive, negative, happy, sad, neutral, etc.)

Because text is vectorized, can use a wide variety of ML classifiers

* + 1. Python Tools
       1. ntlk
       2. spaCy