Natural Language Processing - Introduction

(https://github.com/learn-co-curriculum/dsc-nlp-section-intro) (https://github.com/learn-co-curriculum/dsc-nlp-section-intro/issues/new)

Introduction

This lesson summarizes the topics we'll be covering in this section and why they'll be important to you as a data scientist.

Foundations of Natural Language Processing (NLP)

In this section we will be covering Natural Language Processing (NLP), which refers to analytics tasks that deal with natural human language, in the form of text or speech.

Natural Language Tool Kit (NLTK)

We'll start by providing more context on the Natural Language Tool Kit (NLTK), one of the most popular NLP libraries used in Python. This library was developed by researchers at the University of Pennsylvania, and it has quickly become one of the most powerful and complete library of NLP tools available.

Regular Expressions

Data preprocessing is an essential part of NLP, and that's why being very familiar with **regular expressions** is extremely important. Regular expressions, or "Regex" is extremely useful for NLP. We can use regex to quickly pattern match and filter through text documents.

Feature Engineering for Text Data

Working with text data comes with a lot of ambiguity. Feature engineering for NLP is pretty specific, and in this section you'll learn some feature engineering techniques that are essential when working with text data. You'll learn how to remove stop words from your text, as well as how to create frequency distributions, representing histograms that give us an overview of the total number of times each word occurs in a given text corpus.

Additionally, you'll learn about stemming and lemmatization, which is the technique of removing suffixes from our words (and can enhance our text insight by creating frequency histograms *after* having performed stemming or lemmatization!). You'll also learn how to create bigrams, which creates an insight on how often two words occur together!

Context-Free Grammars and Part-of-Speech (POS) Tagging

In NLP, it is important to understand what context-free grammars and part-of-speech tagging are. Context-free grammars refer to bits of text that are grammatically correct, but feel like complete nonsense when considering the same bit of text on the semantic level. POS tagging refers to the act of helping a computer understand how to interpret a sentence. The context-free grammars (CFG) defines the rules of how sentences can exist. You'll see multiple examples on how to use both CFG and POS tagging, and why they are important!

Text Classification

We will finish off this section by explaining the general process to set text data up for classification problems.

Summary

In this section, you'll learn the foundations of NLP and different techniques to make a computer understand text!