16.5.1

Natural Language Processing

You and Jennifer feel you have a good understanding of PySpark and your ability to handle big data. Now it's time to focus on the next part of the challenge, which is handling text data. To start, you'll learn about natural language.

Natural language processing (NLP) is a growing field of study that combines linguistics and computer science for computers to understand written, spoken, and typed natural language. NLP is the process of converting normal language to a machine-readable format, which allows a computer to analyze text as if it were numerical data.

While NLP has a wide variety of use cases, and the field is rapidly growing, there are a few use cases that are particularly interesting:

- **Analyzing legal documents:** NLP can be used to analyze many types of legal documents. This can improve the outcome of a given case, as lawyers and staff can find critical information quickly.
- **U.S. Securities and Exchange Commission (SEC) filings:** NLP is used to analyze SEC filings for various businesses. Companies use NLP to analyze filings for real-time business intelligence.
- **Chatbots:** Chatbots are one of the most popular use cases. Chatbots can be used for selling products, customer support, and even medical help.

At this point, you might ask how this relates to big data. Due to the massive amounts of text data needed to drive insights, we'll have to learn how to manage that data. There are a number of important use cases to delve into:

- **Classifying text:** For many of the aforementioned use cases to work, a computer must know how to classify a given piece of text. Classification can mean a few different things in NLP. You can have classification of specific words, even specifying what the part of speech is. You can also classify what the text is as a whole.
- **Extracting information:** Many NLP tasks require the ability to retrieve specific pieces of information from a given document. Think of the case where we are extracting data from law documents. You might want to extract certain aspects of that document to present good cases.
- **Summarizing a document:** Summarization is a key aspect of NLP. It helps solve quite a few different problems. You can essentially create a model that summarizes a given document. This can be helpful to understand the high-level details of law documents, articles, and much more.

Retake

Context of Natural Language

Natural language can be complicated because the way it's written is not always how it is intended. Therefore, you might need the full context to understand the meaning.

Sarcasm is a great example. Say you had a bad experience at a restaurant. Your friend asks if you liked your meal and you reply "Oh, yeah, the food was amazing if you like dry, bland food." A friend familiar with your humor would understand your true intentions behind the quip. However, a straight reading, without detecting sarcasm, would give the impression you prefer dry, bland food.

Another challenge is interpreting the tone behind the text. For instance, snidely remarking "Great" and enthusiastically exclaiming "Great!" reveal two distinct tones but, in text, it is the same word.

These are just two examples of the complexity of dealing with natural language. In the next section, we'll show how a computer does its best to interpret language.