Intro to Natural Language Processing

Amol Mavuduru, AIS VP and Tech Lead

How this workshop is going to work...

- **First part**: a very brief slide presentation introducing the topic of NLP.
- **Second part:** a hands-on coding workshop where we will run code to train a sentiment classifier with over 90% accuracy.

Where to find the code for this workshop...

- You can find all of the code and materials for this workshop, including this presentation at this GitHub repo:

https://github.com/AmolMavuduru/IntroToNLP

What is natural language processing?

- A broad area of AI focused on processing and extracting meaning from human language data.
- Has a wide range of applications including:
  - Clinical medicine
  - Opinion mining
  - Marketing
  - Conversational interfaces

Extracting Features From Text

- Natural language processing frequently relies on techniques from **machine learning**.
- But... machine learning algorithms are designed to work with numbers, and raw text is not numerical data.
- In this workshop we will show you how to overcome to actually **transform text data into numerical data**.

Sentiment Analysis

- We will be focusing on this specific problem.
- Specifically applied to movie reviews.
- Given a string of text for a movie review, determine if the review is positive or negative.

The Dataset - IMDb Movie Reviews

- 50,000 IMDb movie reviews scraped from the web.
  - Even split between positive and negative
  - Used in the Stanford paper “Learning Word Vectors for Sentiment Analysis” (2011)

The Machine Learning Workflow

1. Preprocess data
2. Train models
3. Test models
4. Repeat steps 1-3 with different approaches to optimize model accuracy on the test data.