IBM course AI(mod2)

AI performs natural language processing

NLP sentence segmentation and tokens

NATURAL LANGUAGE PROCESSING

Computers are best at working with structured data, in which everything is neatly grouped and labeled. Unfortunately for machines, human language is anything but structured. You've been using language for most of your life. Your brain accomplishes this through some of the most complicated neural circuitry on Earth. But it is very difficult to create machines that can work with human language.

In NLP, machines segment sentences and extract meaning from "tokens" of human language

Human language is unstructured. Although it is loosely held together by rules of grammar, our language expresses information in many confusing ways. Unlike structured information, which can be arranged in tables or matrices with neatly labeled rows and columns, unstructured information is messy and difficult to understand. To see why, consider this famous joke by Groucho Marx.

> *One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know.*Adapted from Groucho Marx, 20th century comedian and movie star

To deal with the "messiness" of unstructured information, computers begin with one sentence at a time. This is called **sentence segmentation**. Computers then break the information into small chunks of information, called **tokens**, that can be individually classified. Once the tokens in text have been sorted into a structure based on what they mean, NLP can work with them.

The following activities show you how Groucho Marx's joke can be tokenized into useful categories called **entities** and **relationships**. You'll learn the meanings of these words as you continue.

> An **entity** is a noun representing a person, place, or thing. It's not an adjective, verb, or other article of speech.

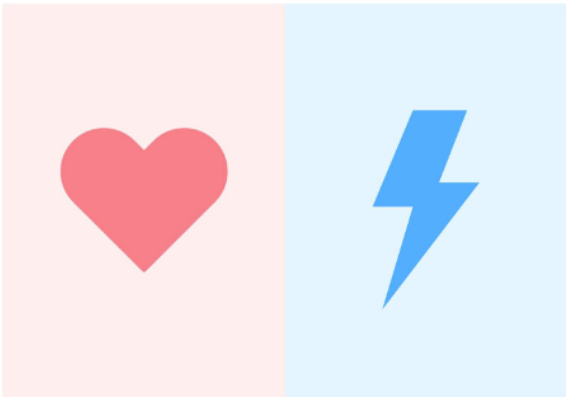
> A **relationship** is a group of two or more entities that have a strong connection to one another.

> A **concept** is something implied in a sentence but not actually stated. This is trickier because it involves matching ideas rather than the specific words present in the sentence.

Emotion detection and sentiment analysis are not the same thing

Emotion detection identifies distinct human emotion types.

For example, you can determine if the emotion being expressed is anger, happiness, or fear after reading a user's rating and comments in an online customer satisfaction survey.

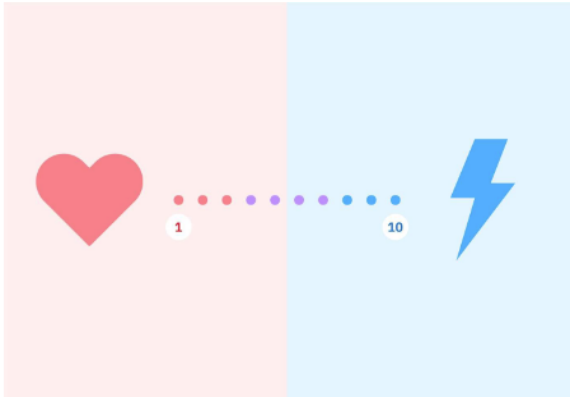


AI can be trained to classify emotions. Identifying the right emotional token can make a big difference when an AI system is reading a social media post or a customer service chat, in which different emotions significantly change the meaning of a sentence.

Sentiment analysis isn't a specific emotion —at least, not as computer scientists use the term. Instead, it's a measure of the strength of an emotion.

You can think of sentiment as a sliding scale between positive and negative, with neutral in the middle.

Sentiment analysis is a means of assessing if data is positive, negative, or neutral.



The classification problem

Human language makes classification challenging

Here's an old-fashioned riddle:

Why does your nose run and your feet smell?

Human language is full of terms that are vague or have double meanings. This is called a **classification problem**. In the riddle, **run** and **smell** each have two meanings.

- "A runny nose" means you have a cold and you need a tissues to wipe your nose.
- "A smelly foot" means that your foot has an unpleasant odor.

It might only take you a moment to understand the joke, but an AI system might have difficulty classifying its elements. Consider these examples:

- You can ship a box by train.
- When a building burns down, it burns up.
- You can fill in a form by filling it out.
- A wise guy is not the same as a wise person.

Classification can be more difficult for an AI system than identifying tokens because so much of classification depends on the context in which a sentence is contained. Compare **I went to the docks to ship my box** to **I went to the station to ship my box**. Both sentences indicate where a box's travel begins, but neither specifies how it will travel. An AI system must associate the word **ship** with either the word **station** or **docks**, and then relate that association with the right concept: either train or boat.

How does an AI system deal with this problem? After ingesting several thousand instances in which shipping from a dock results in boat travel, while shipping from a station leads to shipping by rail, the AI system identifies the frequency in which places and kinds of travel are linked. Gradually, the system gets better at classification and makes fewer mistakes. However, as with humans, an AI system's classification will never be 100% perfect. (That's why well-designed AI systems give not only a response, but also a confidence value.)