Artificial Intelligence Masterclass

Natural Language Processing

H.M. Samadhi Chathuranga Rathnayake

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Introduction to NLP

- Natural language processing (NLP) is a field that focuses on making natural human language usable by computer programs.
- NLTK, or Natural Language Toolkit, is a Python package that you can use for NLP.
- A lot of the data that you could be analyzing is unstructured data and contains human-readable text.

Tokenizing

- Tokenizing makes it simple to divide text into words or sentences.
- This will enable you to deal with shorter passages of text that, even when read separately from the rest of the text, are still largely cohesive and intelligible.
- It's the initial stage in structuring unstructured data so that it may be analyzed more easily.
- When you're analyzing text, you'll be tokenizing by word and tokenizing by sentence.
- Here's what both types of tokenization bring to the table:
  - Tokenizing by word
    - This allows you to identify words that come up particularly often
  - Tokenizing by sentence
    - When you tokenize by sentence, you can analyze how those words relate to one another and see more context.

Stop Words

- Stop words are words that you want to ignore, so you filter them out of your text when you’re processing it.
- Very common words like 'in', 'is', and 'an' are often used as stop words since they don’t add a lot of meaning to a text in and of themselves.

Stemming

- Stemming is a text processing task in which you reduce words to their root, which is the core part of a word.
- For example, the words “helping” and “helper” share the root “help.”
- Stemming allows you to zero in on the basic meaning of a word rather than all the details of how it’s being used.
- NLTK has more than one stemmer, but the most popular one is the Porter stemmer.
- There are two ways, the stemming can go wrong.
  - **Understemming** happens when two related words should be reduced to the same stem but aren’t. This is a false negative.
  - **Overstemming** happens when two unrelated words are reduced to the same stem even though they shouldn’t be. This is a false positive.

POS Tagging

- Part of speech is a grammatical term that deals with the roles words play when you use them together in sentences.
- Tagging parts of speech, or POS tagging, is the task of labeling the words in your text according to their part of speech.
- In English, there are eight parts of speech.

POS Tagging

Part of speech	Role	Examples
Noun	Is a person, place, or thing	mountain, bagel, Poland
Pronoun	Replaces a noun	you, she, we
Adjective	Gives information about what a noun is like	efficient, windy, colorful
Verb	Is an action or a state of being	learn, is, go
Adverb	Gives information about a verb, an adjective, or another adverb	efficiently, always, very
Preposition	Gives information about how a noun or pronoun is connected to another word	from, about, at
Conjunction	Connects two other words or phrases	so, because, and
Interjection	Is an exclamation	yay, ow, wow

Lemmatizing

- Like stemming, lemmatizing reduces words to their core meaning, but it will give you a complete English word that makes sense on its own instead of just a fragment of a word like 'discoveri'.

Chunking

- While tokenizing allows you to identify words and sentences, chunking allows you to identify phrases.
- Chunking makes use of POS tags to group words and apply chunk tags to those groups.
- Chunks don’t overlap, so one instance of a word can be in only one chunk at a time.

Chinking

- Chinking is used together with chunking, but while chunking is used to include a pattern, chinking is used to exclude a pattern.

Named Entity Recognition (NER)

- Named entities are noun phrases that refer to specific locations, people, organizations, and so on.
- With named entity recognition, you can find the named entities in your texts and also determine what kind of named entity they are.

Named Entity Recognition (NER)

- There are several NE types

NE type	Examples
ORGANIZATION	Georgia-Pacific Corp., WHO
PERSON	Eddy Bonte, President Obama
LOCATION	Murray River, Mount Everest
DATE	June, 2008-06-29
TIME	two fifty a m, 1:30 p.m.
MONEY	175 million Canadian dollars, GBP 10.40
PERCENT	twenty pct, 18.75 %
FACILITY	Washington Monument, Stonehenge
GPE	South East Asia, Midlothian