Universiti Malaysia PAHANG

Engineering • Technology • Creativity

CHAPTER 4
Natural Language Processing

ARTIFICIAL INTELLIGENCE (BSD2513)
DR. KU MUHAMMAD NA'IM KU KHALIF

Content

Chapter 4.1: Introduction to Natural Language Processing (NLP)

Chapter 4.2: Tokenizing Text Data

Chapter 4.3: Stemming and Lemmatization

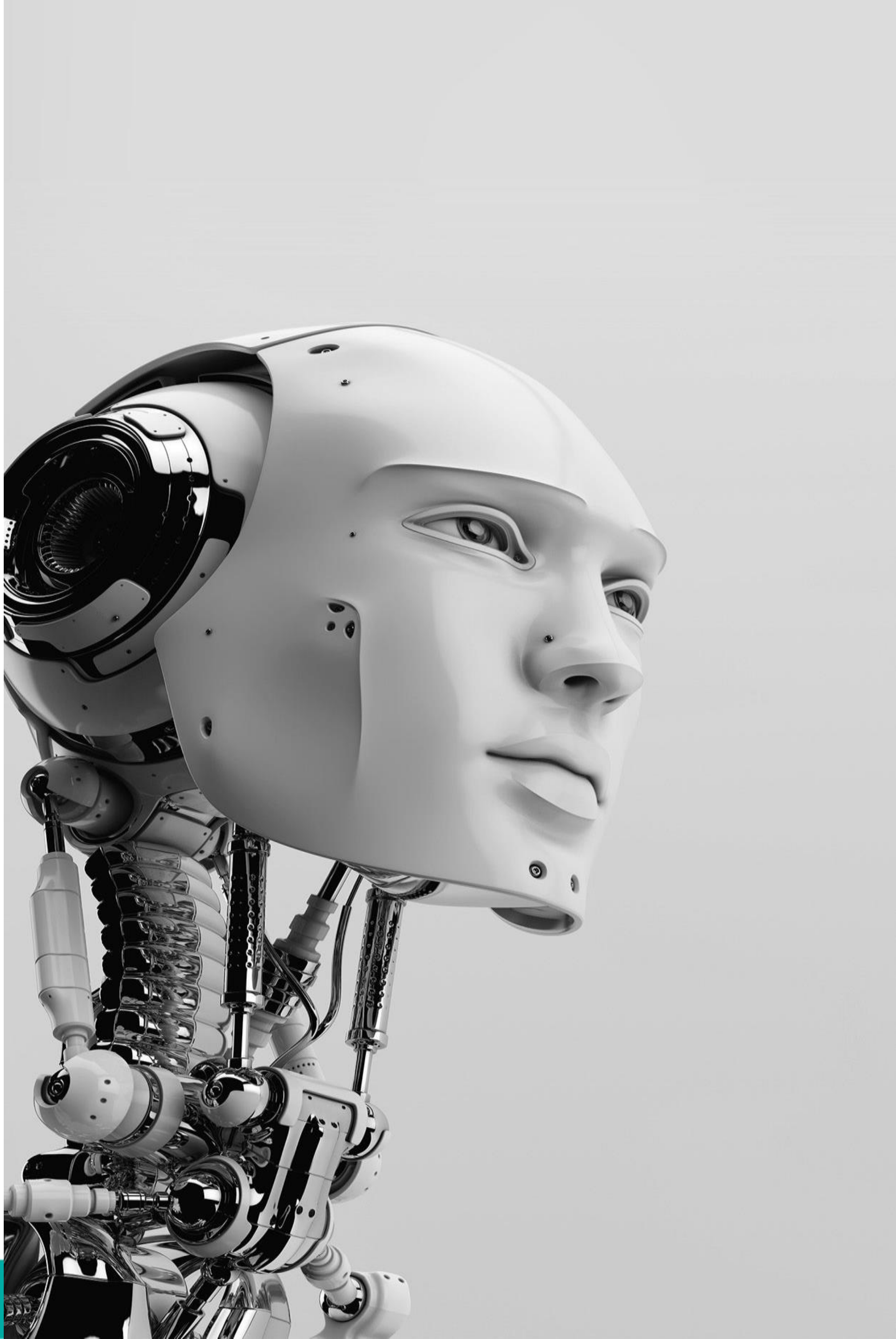
Chapter 4.4: Part of Speech (POS) Tagging and Chunking

Chapter 4.5: Building a Sentiment Analyzer

Chapter 4.1: Introduction to Natural Language Processing (NLP)

By the end of this topic, you should be able to:

- learn about the basic concepts of natural language processing (NLP) in artificial intelligence fields.
- understand the important and useful of NLP in real world applications.



An Overview of NLP

- Everything we express (verbally or in writing) carries huge amounts of information. The topic we choose, our tone, our selection of words, everything adds information that can be interpreted and value extracted.
- In theory, we can understand and predict human behaviour using that information.
- But there is a problem: one person may generate hundreds or thousands of words in a declaration, each sentence with its corresponding complexity. The situation is unmanageable if you want to scale and analyse several hundreds, thousands or millions of people or declarations in a given geography.

- Data generated from conversations, declarations. Even tweets are examples of unstructured data.
- **Unstructured data** doesn’t fit neatly into relational databases' traditional row and column structure and represent the vast majority of data available in the actual world. It is messy and hard to manipulate. Nevertheless, thanks to the advances in disciplines like machine learning, a big revolution is going on.
- Nowadays, it is no longer about trying to interpret a text or speech based on its keywords (the old-fashioned mechanical way), but about understanding the meaning behind those words (the cognitive way). This way, detecting figures of speech like irony or even performing sentiment analysis is possible.
- Having computers that can understand human language is one of the breakthroughs that will make computers even more helpful. **NLP** provides the foundation to begin to understand how this might be possible.

Natural Language

- Natural language refers to the way we, as human beings, communicate with each other.
- Namely, speech and text.
- We are surrounded by text.
- Think about how much text you see each day:
  a. Signs
  b. Menus
  c. Email
  d. SMS
  e. Web Pages

- Now think about speech. As a species, we may speak to each other more than we write. It may even be easier to learn to speak than to write. Voice and text are how we communicate with each other.
- Given the importance of this type of data, we must have methods to understand and reason about natural language, just like we do for other types of data.

"Can anyone, *anyone*, tell me how a semicolon is used other than in emoticons?!"

Challenges of Natural Language

- Working with natural language data is not solved. It has been studied for half a century and is hard.

It is hard from the standpoint of the child, who must spend many years acquiring a language … it is hard for the adult language learner, it is hard for the scientist who attempts to model the relevant phenomena, and it is hard for the engineer who attempts to build systems that deal with natural language input or output. These tasks are so hard that Turing could rightly make fluent conversation in natural language the centerpiece of his test for intelligence.

— Page 248, Mathematical Linguistics, 2010.

- Natural language is primarily hard because it is messy. There are few rules. And yet we can easily understand each other most of the time.

Human language is highly ambiguous … It is also ever changing and evolving. People are great at producing language and understanding language, and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings. At the same time, while we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language.

— Page 1, Neural Network Methods in Natural Language Processing, 2017.

From Linguistics to Natural Language Processing

Linguistics

- Linguistics is the scientific study of language, including grammar, semantics, and phonetics.
- Classical linguistics involved devising and evaluating rules of language. Significant progress was made on formal methods for syntax and semantics, but for the most part, the interesting problems in natural language understanding resist clean mathematical formalisms.
- Broadly, a linguist studies language, but perhaps more colloquially, a self-defining linguist may be more focused on being out in the field.
- Mathematics is the tool of science. Mathematicians working on natural language may refer to their study as mathematical linguistics, focusing exclusively on using discrete mathematical formalisms and theory for natural language (e.g. formal languages and automata theory).

Computational Linguistics

- Computational linguistics is the modern study of linguistics using computer science tools. Yesterday's linguistics maybe today's computational linguist, as using computational tools and thinking has overtaken most fields of study.

Computational linguistics is the study of computer systems for understanding and generating natural language. … One natural function for computational linguistics would be the testing of grammars proposed by theoretical linguists.

— Pages 4-5, Computational Linguistics: An Introduction, 1986.

- Large data and fast computers mean that writing and running software can discover new and different things from large text datasets.
- In the 1990s, statistical methods and machine learning began to and eventually replaced the classical top-down rule-based approaches to language, primarily because of their better results, speed, and robustness. These approaches to studying natural language now dominate the field; it may define the field.
- Computational linguistics also became known by the name of natural language process or NLP.

Statistical Natural Language Processing

- In NLP, to reflect the more engineer-based or empirical statistical methods approach.
- The statistical dominance of the field also often leads to NLP being described as Statistical Natural Language Processing, perhaps to distance it from the classical computational linguistics methods.

I view computational linguistics as having both a scientific and an engineering side. The engineering side of computational linguistics, often called natural language processing (NLP), is largely concerned with building computational tools that do useful things with language, e.g., machine translation, summarization, question-answering, etc. Like any engineering discipline, natural language processing draws on a variety of different scientific disciplines.

— How the statistical revolution changes (computational) linguistics, 2009.

- Linguistics is a large topic of study, and, although the statistical approach to NLP has shown great success in some areas, there is still room and great benefit from the classical top-down methods.

Roughly speaking, statistical NLP associates probabilities with the alternatives encountered in the course of analyzing an utterance or a text and accepts the most probable outcome as the correct one. … Not surprisingly, words that name phenomena that are closely related in the world, or our perception of it, frequently occur close to one another so that crisp facts about the world are reflected in somewhat fuzzier facts about texts. There is much room for debate in this view.

— Page xix, The Oxford Handbook of Computational Linguistics, 2005.

Natural Language Processing

- As data science practitioners interested in working with text data, we are concerned with the tools and methods from the field of Natural Language Processing.
- We have seen the path from linguistics to NLP in the previous section. Now, let's look at how modern researchers and practitioners define what NLP is all about.
- In perhaps one of the more widely textbooks written by top researchers in the field, they refer to the subject as "linguistic science," permitting discussion of classical linguistics and modern statistical methods.

The aim of a linguistic science is to be able to characterize and explain the multitude of linguistic observations circling around us, in conversations, writing, and other media. Part of that has to do with the cognitive size of how humans acquire, produce and understand language, part of it has to do with understanding the relationship between linguistic utterances and the world, and part of it has to do with understand the linguistic structures by which language communicates.

— Page 3, Foundations of Statistical Natural Language Processing, 1999.

- They go on to focus on inference through the use of statistical methods in natural language processing.

Statistical NLP aims to do statistical inference for the field of natural language. Statistical inference in general consists of taking some data (generated in accordance with some unknown probability distribution) and then making some inference about this distribution.

— Page 191, Foundations of Statistical Natural Language Processing, 1999.

- In their text on applied natural language processing, the authors and contributors to the popular NLTK Python library for NLP describe the field broadly as using computers to work with natural language data.

We will take Natural Language Processing — or NLP for short –in a wide sense to cover any kind of computer manipulation of natural language. At one extreme, it could be as simple as counting word frequencies to compare different writing styles. At the other extreme, NLP involves “understanding” complete human utterances, at least to the extent of being able to give useful responses to them.

— Page ix, Natural Language Processing with Python, 2009.

- Statistical NLP has turned another corner and is now firmly focused on using deep learning neural networks to perform inference on specific tasks and develop robust end-to-end systems.
- In one of the first textbooks dedicated to this emerging topic, Yoav Goldberg concisely defines NLP as automatic methods that take natural language as input or produce natural language as output.

Natural language processing (NLP) is a collective term referring to automatic computational processing of human languages. This includes both algorithms that take human-produced text as input, and algorithms that produce natural looking text as outputs.

— Page xvii, Neural Network Methods in Natural Language Processing, 2017.

Reference: https://machinelearningmastery.com/natural-language-processing/

Definitions:

Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages.

(Diego Lopez Yse, 2019)

"No, hashtags are not a part of speech."

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

(Jason Browniee, 2019)

NLP Applications

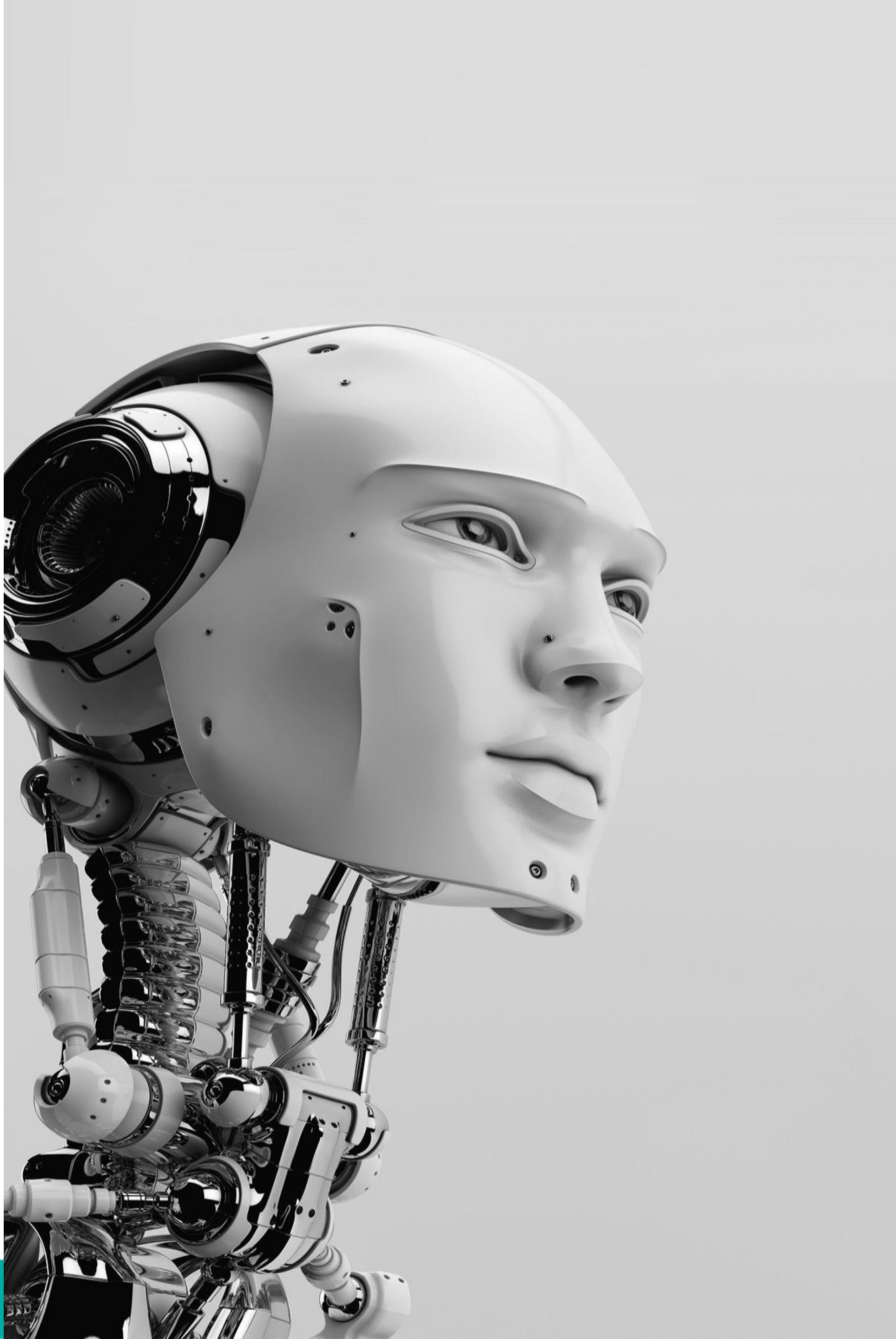
1. Language Translator
2. Social Media Monitoring
3. Chatbots
4. Survey Analysis
5. Targeted Advertising
6. Hiring and Recruitment
7. Voice Assistants
8. Grammar Checkers
9. Email Filtering

Reference: https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/

Chapter 4.2:
Tokenizing Text Data

By the end of this topic, you should be able to:

- use a python package called Natural Language Toolkit (NLTK) to build the NLP applications.
- apply the tokenization to deal with text and break it down into smaller pieces for analysis.



Installation of Packages

You can install it by running the following command:

```
$ pip3 install nltk
```

You can find more information about NLTK at http://www.nltk.org.
In order to access all the datasets provided by NLTK, we need to download it. Open a Python shell by typing the following:

```
$ python3
```

We are now inside the Python shell. Type the following to download the data:

```
>>> import nltk
>>> nltk.download()
```

We will also use a package called `gensim`. It is a robust semantic modeling library that's useful for many applications. It can be installed by running the following command:

```
$ pip3 install genism
```

You might need another package, called `pattern`, for `gensim` to function properly. You can install it by running the following command:

```
$ pip3 install pattern
```

You can find more information about `gensim` at https://radimrehurek.com/gensim.

Tokenizing Text Data

- When we deal with text, we need to break it down into smaller pieces for analysis. To do this, tokenization can be applied. Tokenization is the process of dividing text into a set of pieces, such as words or sentences. These pieces are called tokens.
- Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units is called token.
- The tokens could be words, numbers or punctuation marks. In tokenization, smaller units are created by locating word boundaries. These are the ending point of a word and the beginning of the next word. These tokens are considered as a first step for stemming and lemmatization.

Natural Language Processing

['Natural', 'Language', 'Processing']

Why is Tokenization required in NLP?

- Before processing a natural language, we need to identify the words that constitute a string of characters. That’s why tokenization is the most basic step to proceed with NLP (text data). This is important because the meaning of the text could easily be interpreted by analyzing the words present in the text.
- Let’s take an example. Consider the below string:

“This is a cat.”

- What do you think will happen after we perform tokenization on this string? We get [‘This’, ‘is’, ‘a’, cat’].
- There are numerous uses of doing this. We can use this tokenized form to:

  i. Count the number of words in the text.

  ii. Count the frequency of the word, that is, the number of times a particular word is present.

Methods to Perform Tokenization in Python

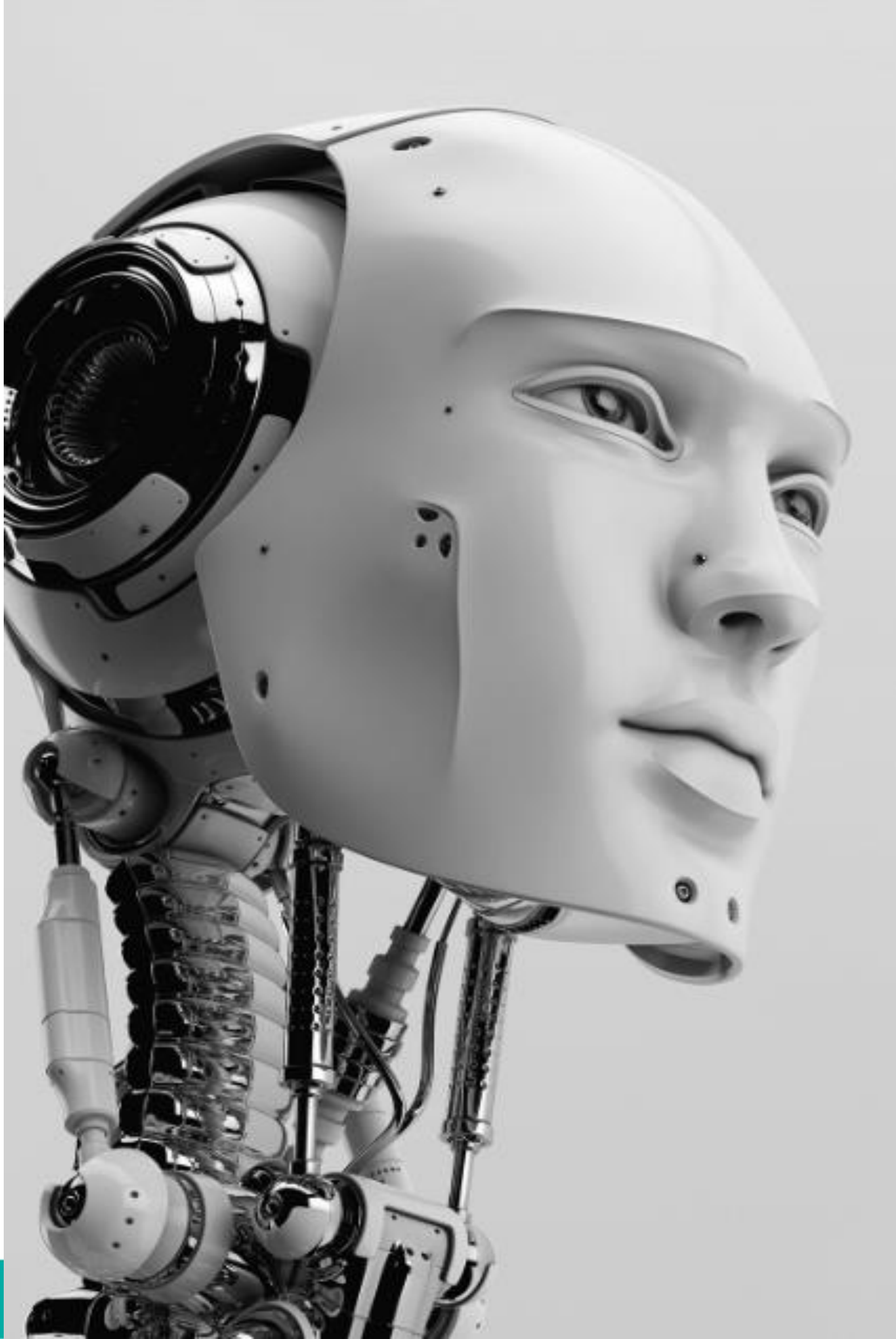
1. Tokenization using Python’s split() function
2. Tokenization using Regular Expressions (RegEx)
3. Tokenization using NLTK
4. Tokenization using the spaCy library
5. Tokenization using Keras
6. Tokenization using Gensim

Commonly, Tokenization breaks the given string for word and sentence.

Chapter 4.3: Stemming and Lemmatization

By the end of this topic, you should be able to:

- apply the stemming and lemmatization techniques to convert words to their base forms.



Converting Words to Their Base Forms Using Stemming

- Working with text means working with many variations. We must deal with different forms of the same word and enable the computer to understand that these words have the same base form. For example, the word sing can appear in many forms, such as singer, singing, song, sung, and so on. This set of words share similar meanings. This process is known as **stemming**. Stemming is a way of producing morphological variants of a root/ base word. Humans can quickly identify these base forms and derive context.
- When analysing text, it's helpful to extract these base forms. Doing so enables the extraction of valuable statistics derived from the input text. Stemming is one way to achieve this. The goal of a stemmer is to reduce words from their different forms into a common base form. The heuristic process cuts off the ends of words to extract their base forms.

What is Stemming?

- When Stemming is applied to the words in the corpus, the word gives the base for that particular word.
- It is like a tree with branches, you are removing the branches till their stem. Eg: fix, fixing, fixed gives fix when stemming is applied.
- There are different types through which Stemming can be performed. Some of the popular ones which are being used are:
  1. Porter Stemmer
  2. Lancaster Stemmer
  3. Snowball Stemmer

Converting Words to Their Base Forms Using Lemmatization

- **Lemmatization** is another method of reducing words to their base forms. In the previous section, we saw that some of the base forms that were obtained from those stemmers didn't make sense.
- Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.
- Lemmatization is like stemming, but it brings context to the words. So, it links words with similar meanings to one word. For example, all three stemmers said that the base form of calves is calv, which is not a real word. Lemmatization takes a more structured approach to solve this problem.

  a. Rocks: rock

  b. Corpora: corpus

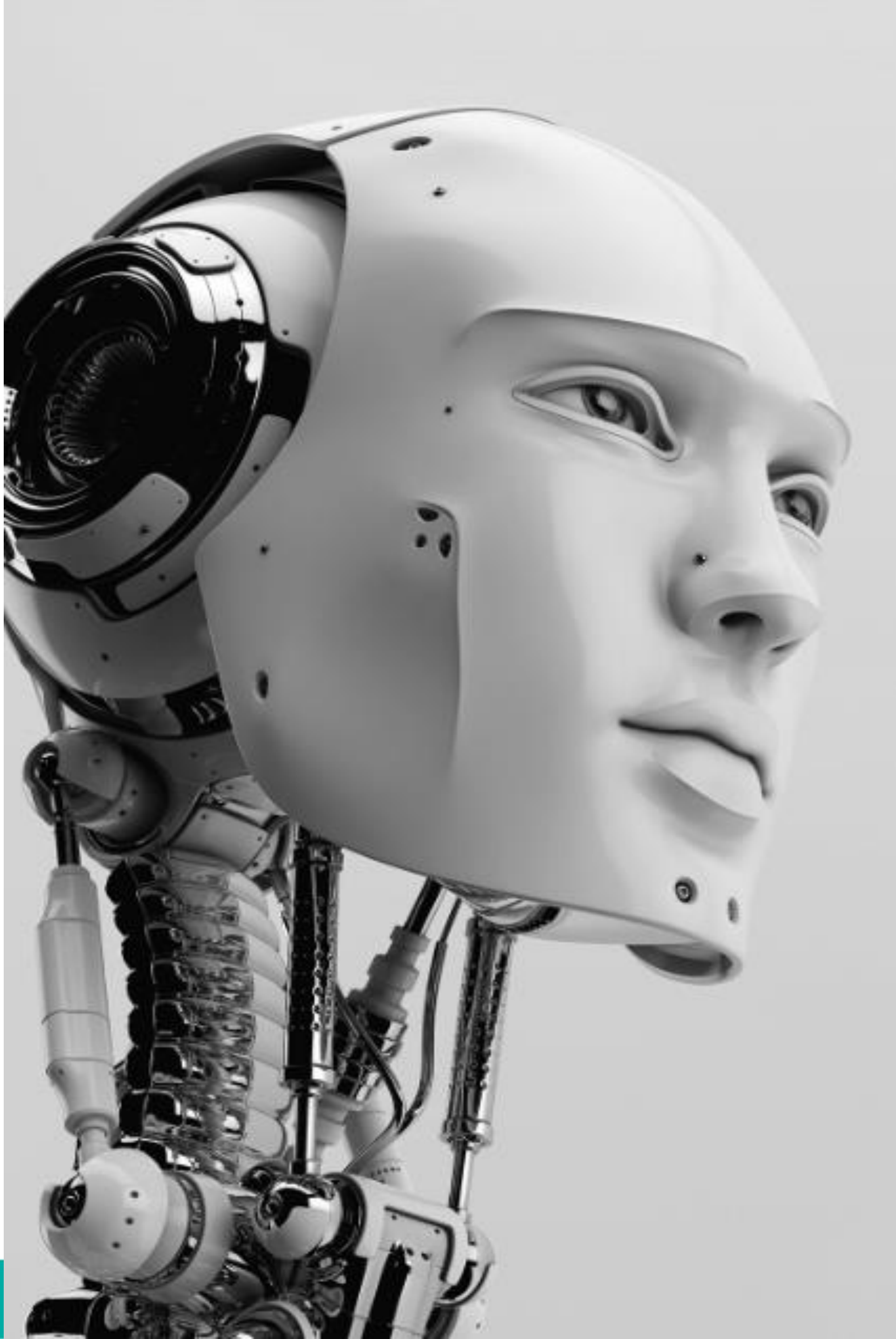
  c. Worse: bad

- There is a slight difference between lemmatization and stemming, them is lemmatization cuts the word to gets its lemma word meaning it gets a much more meaningful form than what stemming does. The output we get after lemmatization is called ‘lemma’.
- There are many methods through which lemma can get obtained and lemmatization can be performed. Some of them are WordNet Lemmatization, TextBlob, Spacy, Tree Tagger, Pattern, Genism, and Stanford CoreNLP lemmatization. Lemmatization can be applied from the mentioned libraries.

Chapter 4.4: Part of Speech (POS) Tagging and Chunking

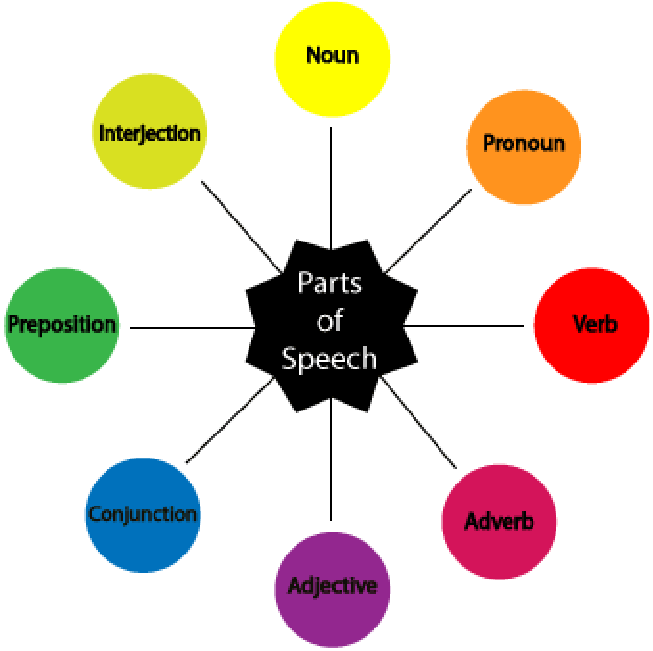
By the end of this topic, you should be able to:

- Learn Part of Speech (POS) tagging and chunking in NLP.
- learn how to dividing text data into chunks in NLP problems.



What is Part of Speech (POS)?

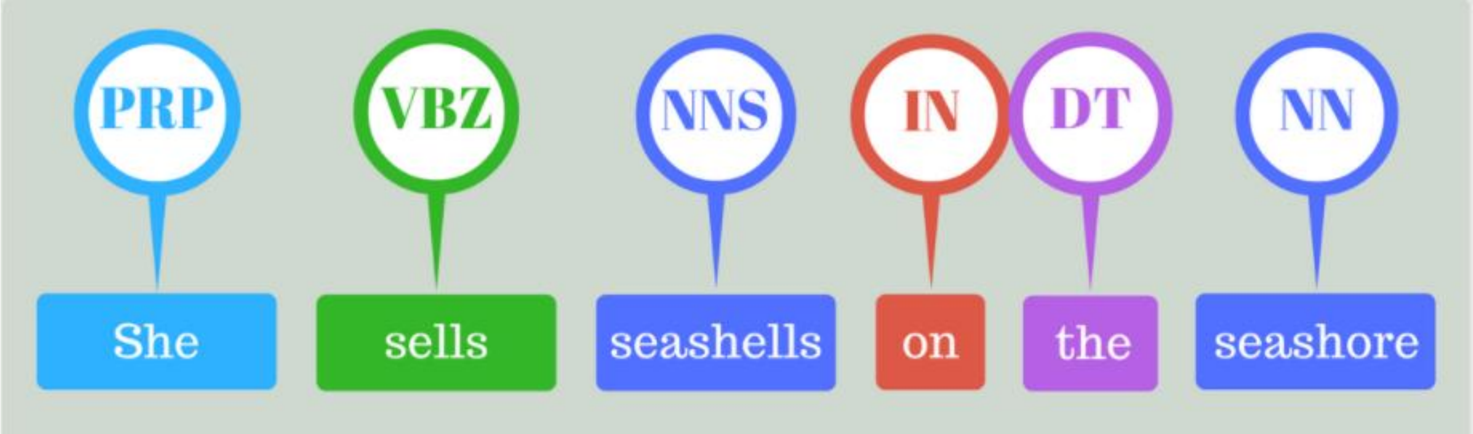
- The part of speech explains how a word is used in a sentence. There are eight main parts of speech - nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections.



- Noun (N)- Daniel, London, table, dog, teacher, pen, city, happiness, hope
- Verb (V)- go, speak, run, eat, play, live, walk, have, like, are, is
- Adjective(ADJ)- big, happy, green, young, fun, crazy, three
- Adverb(ADV)- slowly, quietly, very, always, never, too, well, tomorrow
- Preposition (P)- at, on, in, from, with, near, between, about, under
- Conjunction (CON)- and, or, but, because, so, yet, unless, since, if
- Pronoun(PRO)- I, you, we, they, he, she, it, me, us, them, him, her, this
- Interjection (INT)- Ouch! Wow! Great! Help! Oh! Hey! Hi!

- Most POS are divided into sub-classes. POS Tagging means labelling words with their appropriate Part-Of-Speech.

How does POS Tagging works?



- POS tagging is a supervised learning solution that uses features like the previous word, next word, first letter capitalised etc. NLTK has a function to get pos tags, and it works after tokenization process.
- Tagging tells you whether words are nouns, verbs, adjectives, etc., but it doesn't give you any clue about the structure of the sentence or phrases in the sentence, so group together (connected items or words) so that they can be stored or processed as single concepts. For example, “Prime Minister of Malaysia”

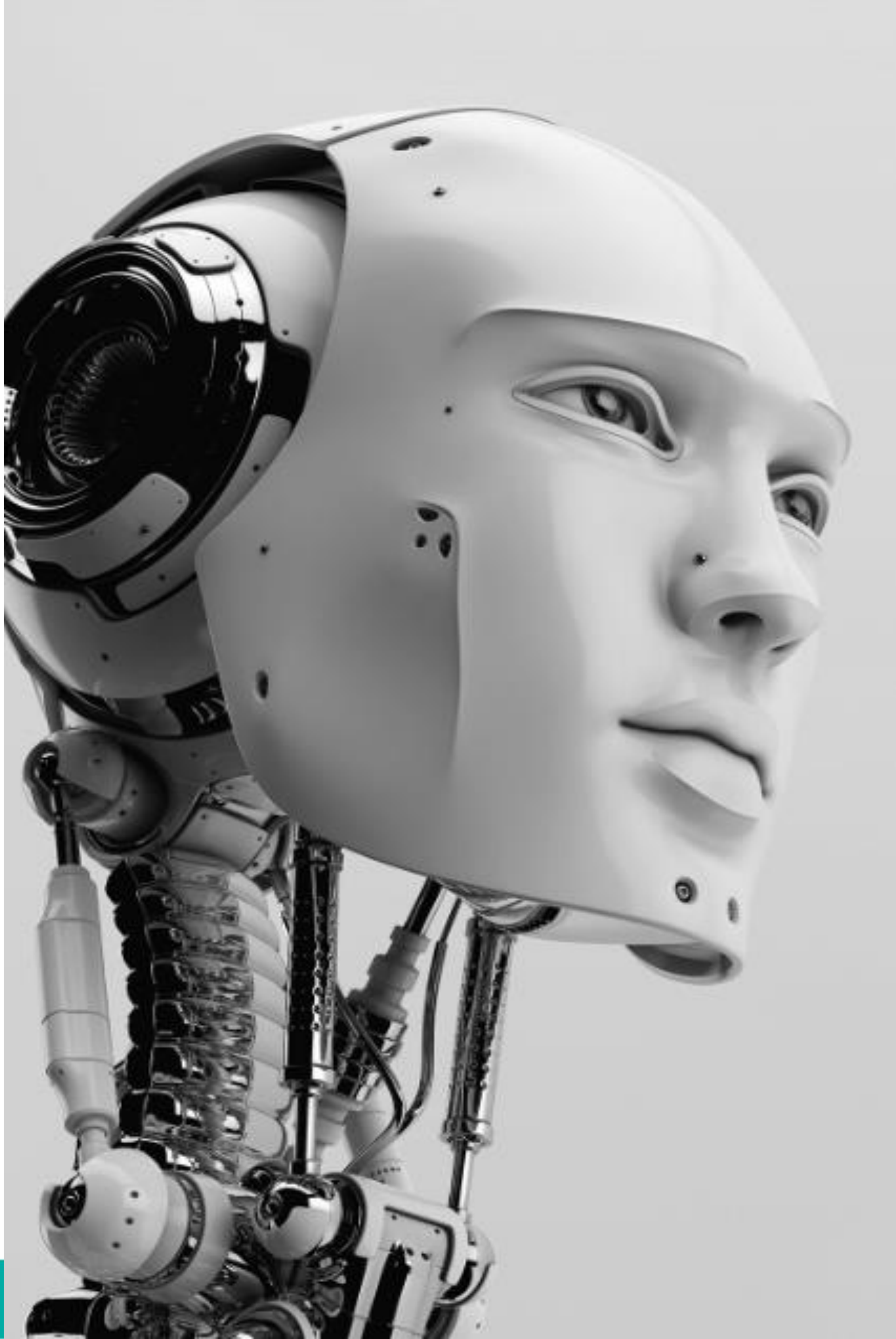
Dividing Text Data into Chunks

- Text data usually needs to be divided into pieces for further analysis. This process is known as **chunking**. This is used frequently in text analysis.
- The conditions that are used to divide the text into chunks can vary based on the problem at hand.
- This is not the same as tokenization, where text is divided into pieces. During chunking, we do not adhere to any constraints, except that the output chunks need to be meaningful.
- When we deal with large text documents, it becomes essential to **divide the text into chunks to extract meaningful information**.

Chapter 4.5: Building a Sentiment Analyzer

By the end of this topic, you should be able to:

- learn how to build a sentiment analyzer in NLP problem.



Building a Sentiment Analyzer

- Sentiment analysis is the process of determining the sentiment of a piece of text.
- For example, it can be used to determine whether a movie review is positive or negative. This is one of the most popular applications of natural language processing. We can add more categories as well, depending on the problem at hand.
- This technique can be used to get a sense of how people feel about a product, brand, or topic. It is frequently used to analyze marketing campaigns, opinion polls, social media presence, product reviews on e-commerce sites, and so on.

Let's see how to determine the sentiment of a movie review.

- We will use a Naive Bayes classifier to build this sentiment analyser. First, extract all the unique words from the text. The NLTK classifier needs this data to be arranged in the form of a dictionary to ingest it.
- Once the text data is divided into training and testing datasets, the Naive Bayes classifier will be trained to classify the reviews into positive and negative.
- Afterward, the top most informative words to indicate positive and negative reviews can be calculated and displayed.
- This information is interesting because it shows what words are being used to denote various reactions.

Links to Read

https://machinelearningmastery.com/natural-language-processing/

https://www.ibm.com/cloud/learn/natural-language-processing

https://www.sas.com/en_nz/insights/analytics/what-is-natural-language-processing-nlp.html

https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/