

# NLP Introduction

講者：Isaac

# Outline

---

- ▶ Why NLP
- ▶ What is Natural Language Processing
- ▶ NLP Applications
- ▶ NLTK Introduction



---

Why NLP?



# Why NLP?

---

- ▶ goal of Natural Language Processing(NLP) is to fill the gap how the humans communicate(natural language) and what the computer understands(machine language)



---

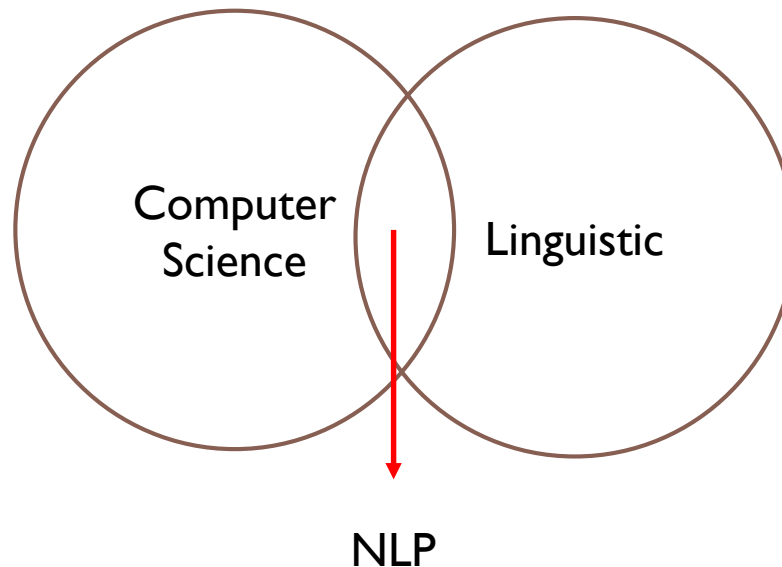
# What is Natural Language Processing



# What's NLP

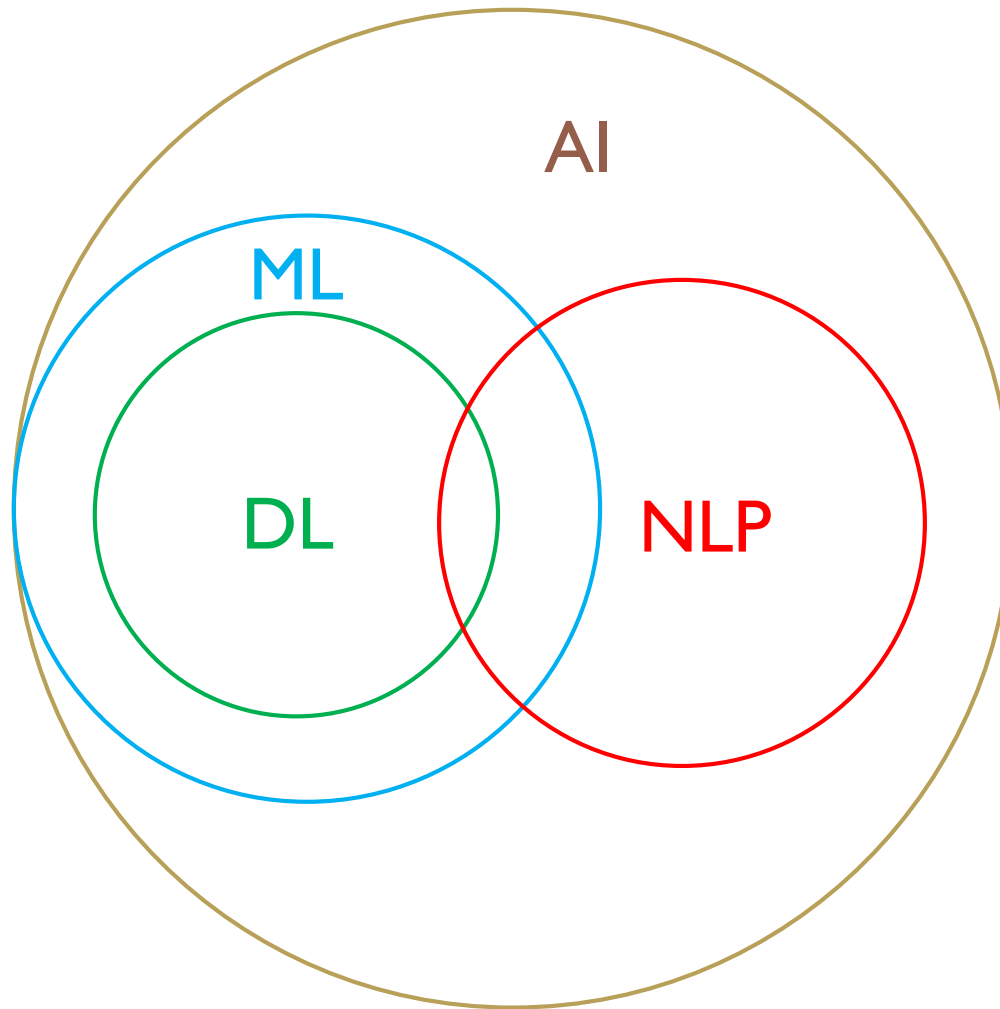
---

- ▶ Natural Language Processing(NLP) is the technology used to aid computers to understand the human's natural language
- ▶ hard to teach computers to understand how we communicate



# Relationship between AI, ML, DL, and NLP

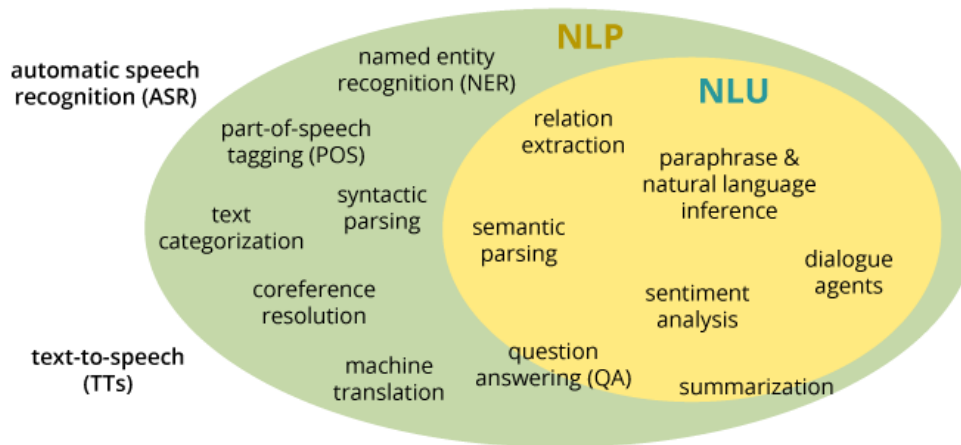
---



# NLU V.S. NLP V.S. ASR

---

## Terminology: NLU vs. NLP vs. ASR



<https://www.kdnuggets.com/2019/07/nlp-vs-nlu-understanding-language-processing.html>



# What's Corpus

---

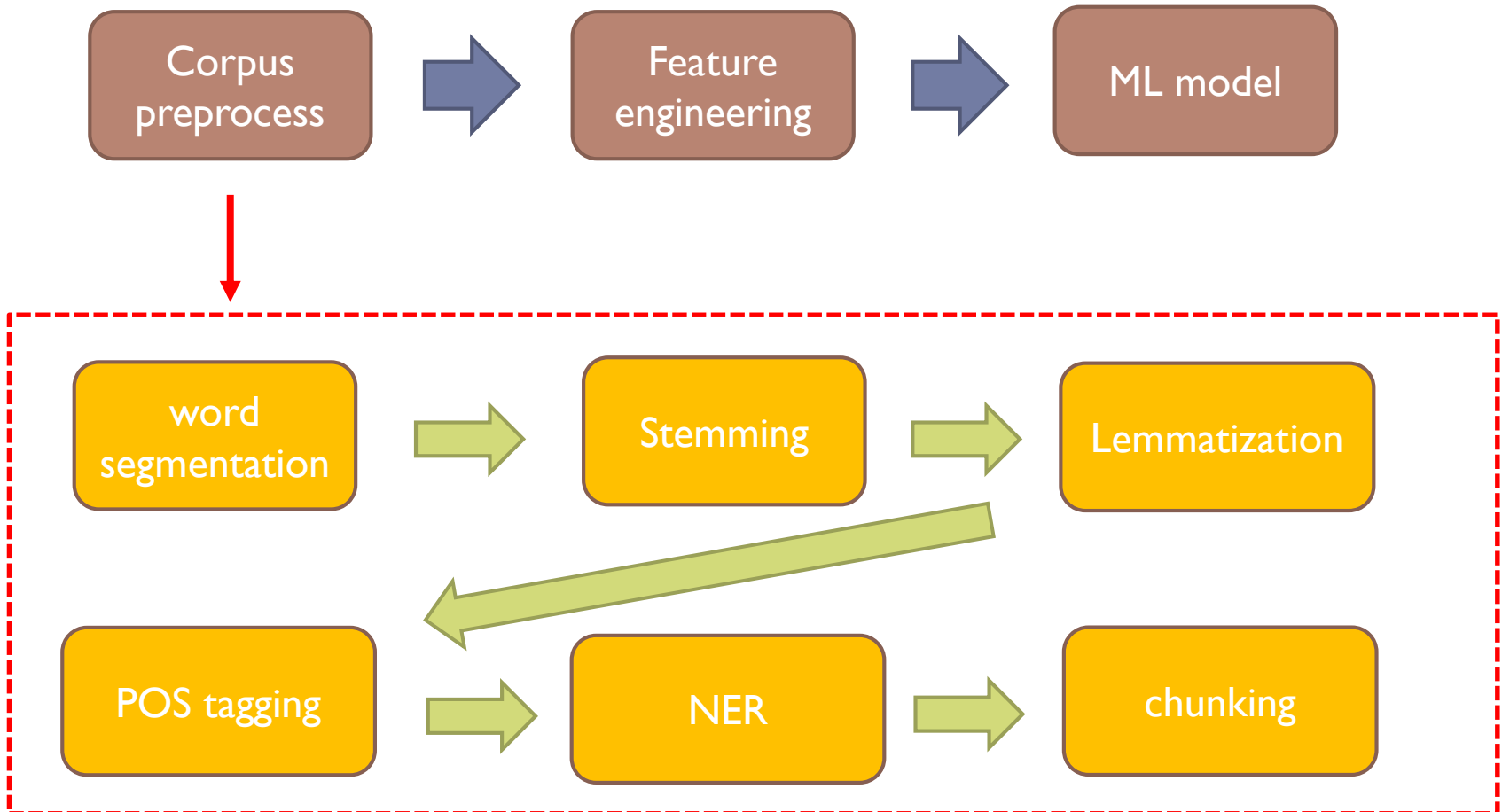
- ▶ corpus refers to a collection of texts
  - ▶ may be single language or multiple languages
  - ▶ common corpus including Google Books Ngram, Brown, American National



**WIKIPEDIA**  
The Free Encyclopedia

# Flows in NLP

---



# Techniques used in NLP?

---

- ▶ **Syntax**

- ▶ If arrangement of words in a sentence makes grammatical sense

- ▶ **Semantics**

- ▶ the meaning that is conveyed by a text

# syntax techniques

---

- ▶ **Syntax techniques**
  - ▶ Word segmentation
  - ▶ Stemming
  - ▶ Lemmatization
  - ▶ Part-of-speech tagging
  - ▶ Parsing
  - ▶ Sentence breaking

# syntax - word segmentation

---

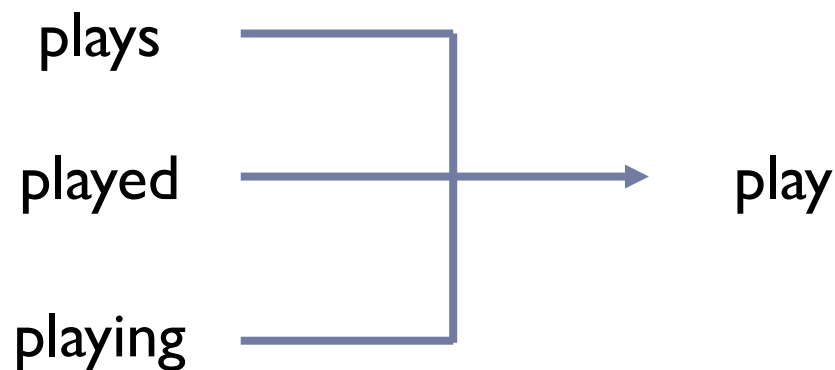
- ▶ task of splitting text into words
  - ▶ Like Chinese/Japanese characters
- ▶ Famous open source
  - ▶ Jieba (<https://github.com/fxsjy/jieba>)

我|電腦|當機|了

# syntax - Stemming

---

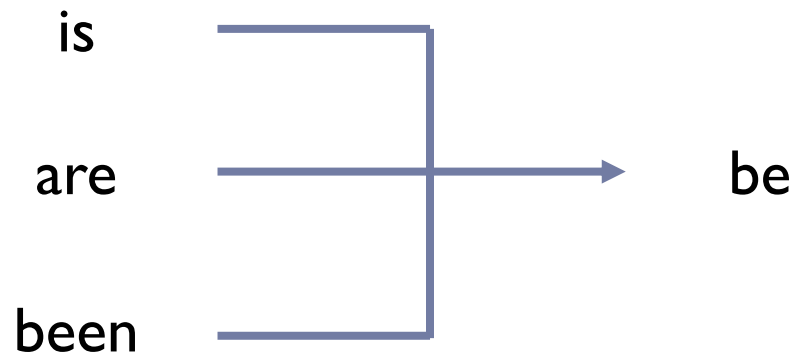
- ▶ cutting the inflected words to their root form
- ▶ famous method including Porter, Snowball, and Lancaster



# syntax -Lemmatization

---

- ▶ reducing the inflected forms of a word into a single form for easy analysis



# syntax - Part-of-speech tagging

---

- ▶ Identifying the part of speech for every word



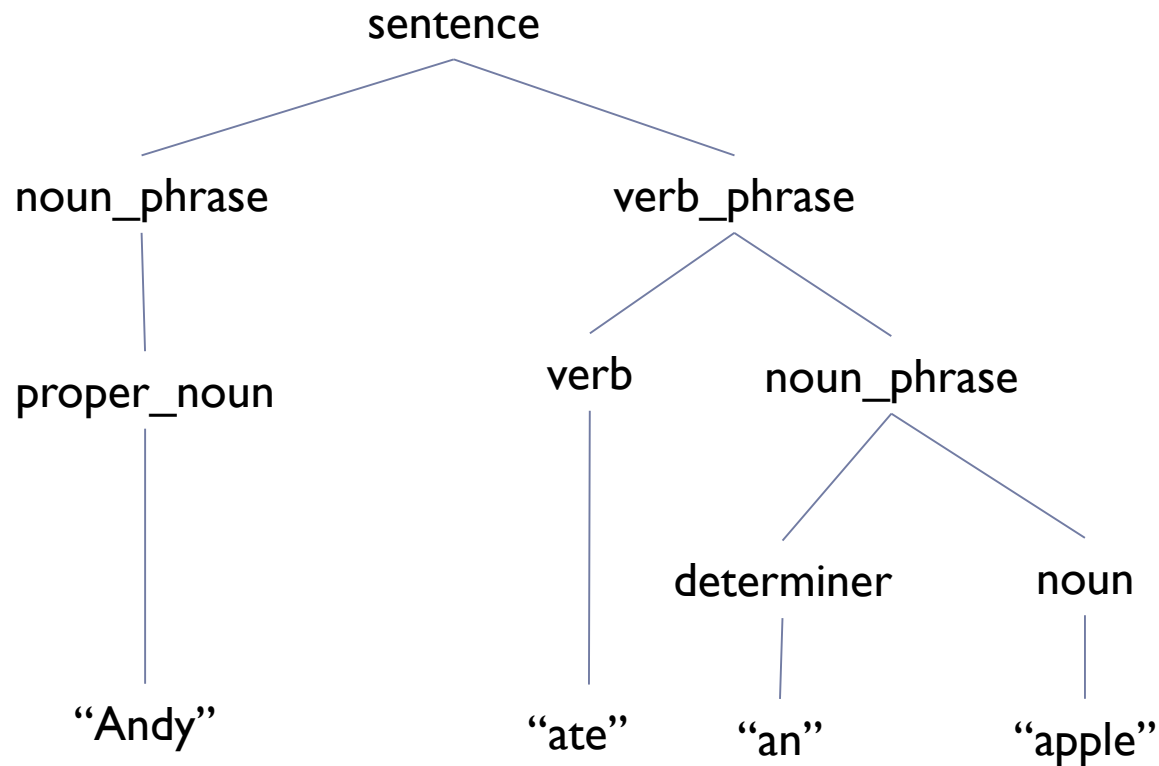
She sells seashells on the seashore



# syntax - Parsing

---

## ► Grammatical analysis of a sentence



# syntax - Parsing

---



**I saw a girl with a telescope**

# syntax - Sentence breaking

---

- ▶ Placing sentence boundaries on a continuous text
- ▶ Also called sentence tokenize, sentence boundary disambiguation, sentence segmentation

Hello world. This blog post is about sentence segmentation. It is not always easy to determine the end of a sentence. One difficulty of segmentation is periods that do not mark the end of a sentence. An ex. is abbreviations.



- Hello world.
- This blog post is about sentence segmentation.
- It is not always easy to determine the end of a sentence.
- One difficulty of segmentation is periods that do not mark the end of a sentence.
- An ex. is abbreviations.

# Semantics techniques

---

- ▶ **Syntax techniques**
  - ▶ Named entity recognition (NER)
  - ▶ Word sense disambiguation
  - ▶ Natural language generation

# Semantics - Named Entity Recognition (NER)

- Finding references to entities in text and labeling them with their location and type

In fact, the **Chinese** NORP market has the **three** CARDINAL most influential names of the retail and tech space – **Alibaba** GPE, **Baidu** ORG, and **Tencent** PERSON (collectively touted as **BAT** ORG), and is betting big in the global **AI** GPE in retail industry space. The **three** CARDINAL giants which are claimed to have a cut-throat competition with the **U.S.** GPE (in terms of resources and capital) are positioning themselves to become the ‘future **AI** PERSON platforms’. The trio is also expanding in other **Asian** NORP countries and investing heavily in the **U.S.** GPE based **AI** GPE startups to leverage the power of **AI** GPE. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing **one** CARDINAL, with an anticipated **CAGR** PERSON of **45%** PERCENT over **2018 - 2024** DATE.

To further elaborate on the geographical trends, **North America** LOC has procured **more than 50%** PERCENT of the global share in **2017** DATE and has been leading the regional landscape of **AI** GPE in the retail market. The **U.S.** GPE has a significant credit in the regional trends with **over 65%** PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as **Google** ORG, **IBM** ORG, and **Microsoft** ORG.

# Semantics - Word sense disambiguation

---

- ▶ Assign the most appropriate meaning to a word within a given context

“The **bank** will not be accepting cash on Saturdays.”

“The river overflowed the **bank**.”

# Semantics - Natural language generation

---



---

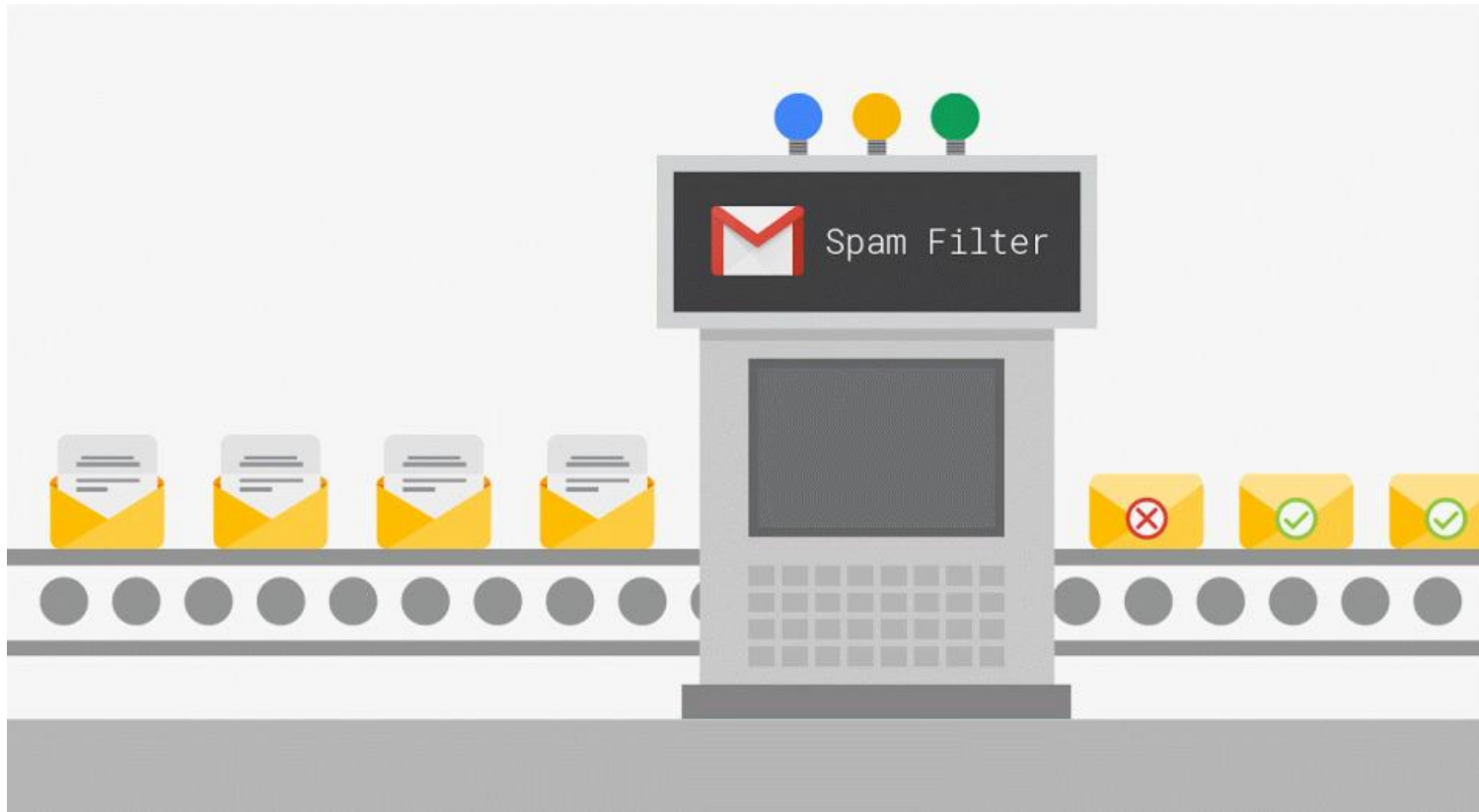
# NLP Applications





# Spam detection

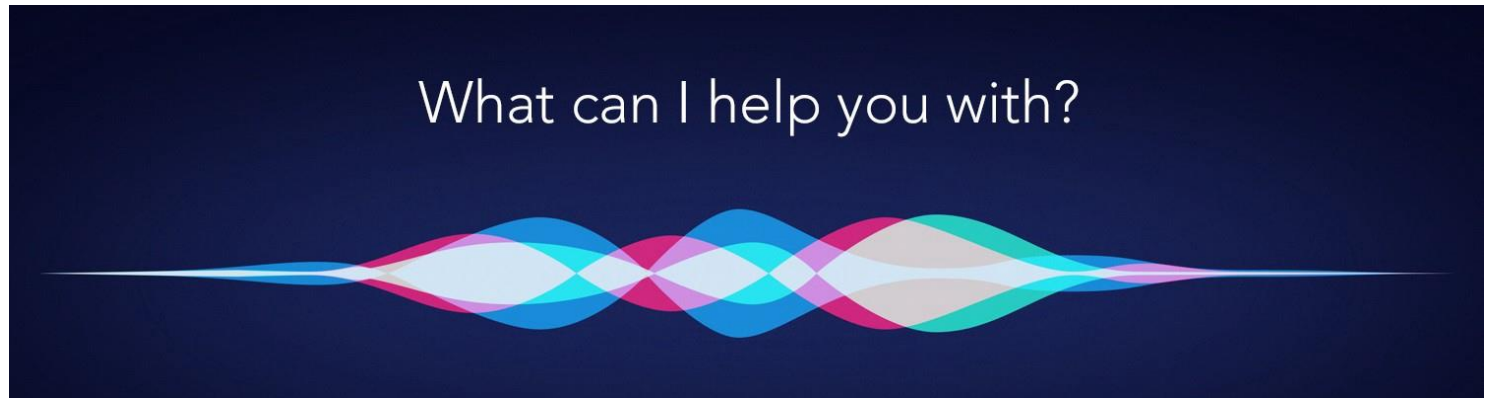
---



Source: <https://i.gifer.com/Ou1t.gif>

# Siri

---



Source: <https://www.analyticsindiamag.com/behind-hello-siri-how-apples-ai-powered-personal-assistant-uses-dnn/>

# Spell checking

---



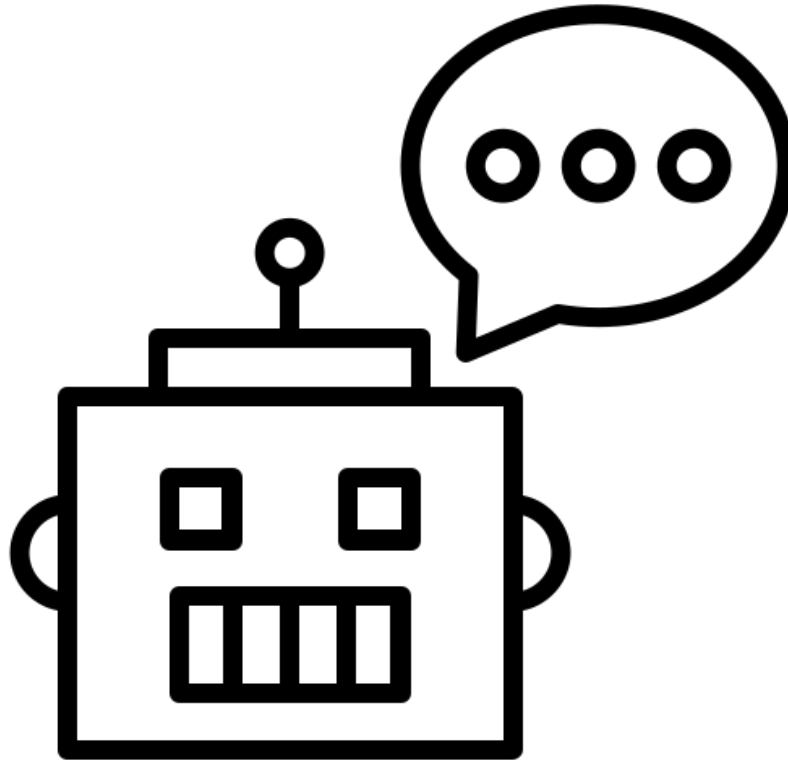
# Automatic Summarization

---



# Chatbots

---



---

# NLTK Introduction



# What's NLTK?

---

- ▶ NLTK (Natural Language Toolkit) is a open source for work with human language data
  - ▶ provide many corpora and lexical resources
  - ▶ <https://www.nltk.org/>



# Four kinds of corpus in NLTK

---

- ▶ **isolate corpus**
  - ▶ collection of natural language text
- ▶ **categorized corpus**
  - ▶ collection of different kinds of categorized text
- ▶ **overlapping corpus**
  - ▶ collection of different kinds of categorized text but some of category are overlapped
- ▶ **temporal corpus**
  - ▶ collection of text in a period of time



# Common open source in NLP

---

