



# Chapter 1

Introduction to Natural Language Processing (NLP)

# Outline

- What is NLP?
- What is Unstructured Text?
- What is Structured Text?
- Translate Unstructured to a Structured Format
- NLU vs NLG
- NLP Goals
- Level of understanding in NLP
- Example of NLP Tools & Services
- Examples of NLP Technology

# What is NLP?

- Natural Language Processing (NLP) is a subfield in
  - Artificial Intelligence (Machine Learning (ML) as its part),
  - Linguistics,
  - Cognitive Science and Computer Sciencethat enables machines to analyze and generate natural language data.
- NLP starts with something called unstructured text.

# What is Unstructured Text?

- What does "unstructured" mean in a data context?
  - Text is commonly referred to as unstructured data.
  - There is definitely structure behind text.
  - There really is structure behind text, there is proper spelling, punctuation, proper sentence construction, and proper thought development.
  - BUT that doesn't allow the text to be considered structured in the eyes of the computer.
  - Text did not fit into a standard database management system (DBMS).

# What is Structured Text?

- Structured data is data that nicely fits inside a standard database management system.
  - The computer expects data to be in records (a key and other attributes).
- One of the interesting questions becomes:
  - How can unstructured data be translated into a structured format?

# Translate Unstructured to a Structured Format

- UNSTRUCTURED -

ADD EGGS AND MILK  
TO MY SHOPPING LISTS

NLP

NLU



NLG

- STRUCTURED -

<SHOPPIN LIST>

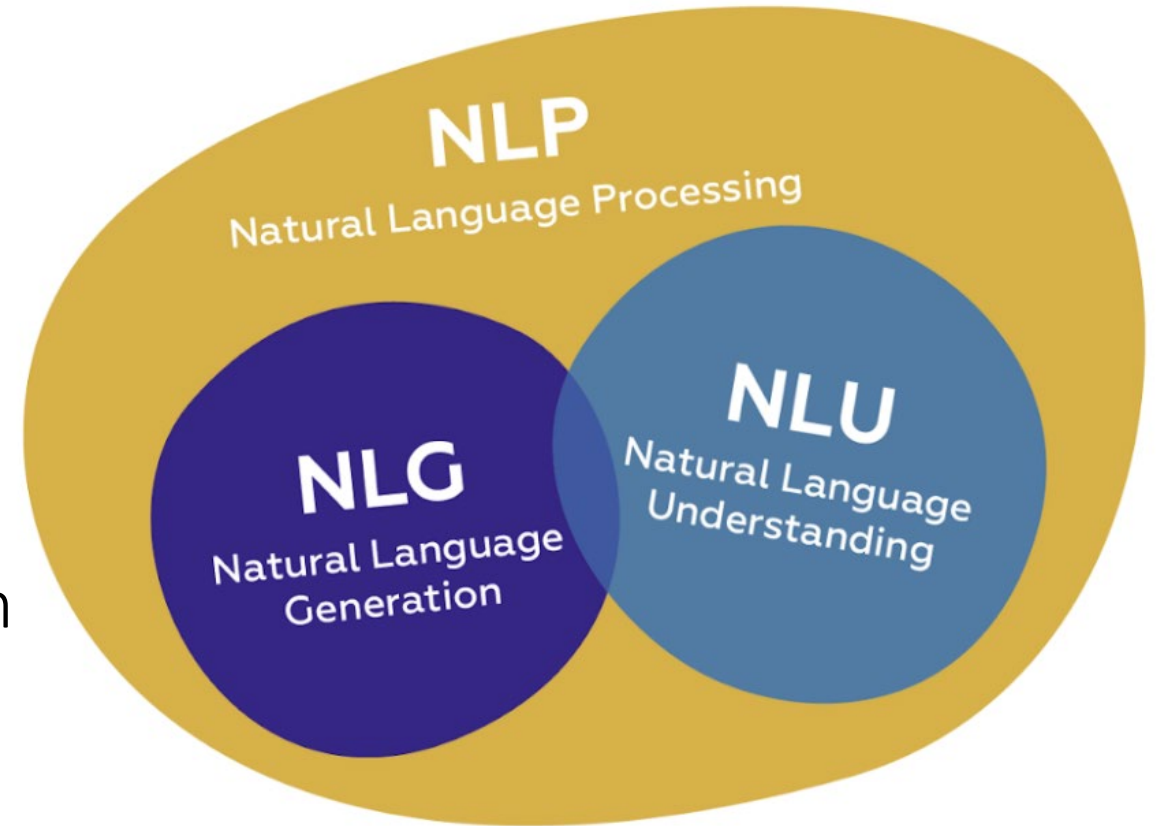
<ITEM>EGGS</>

<ITEM>MILK</>

</>

# NLU vs NLG

- Both are main branches of the NLP.
- NLU involves transforming human language into a machine-readable format.
- NLG involves the processing and conversation of the information from the computer language to the understandable human language.

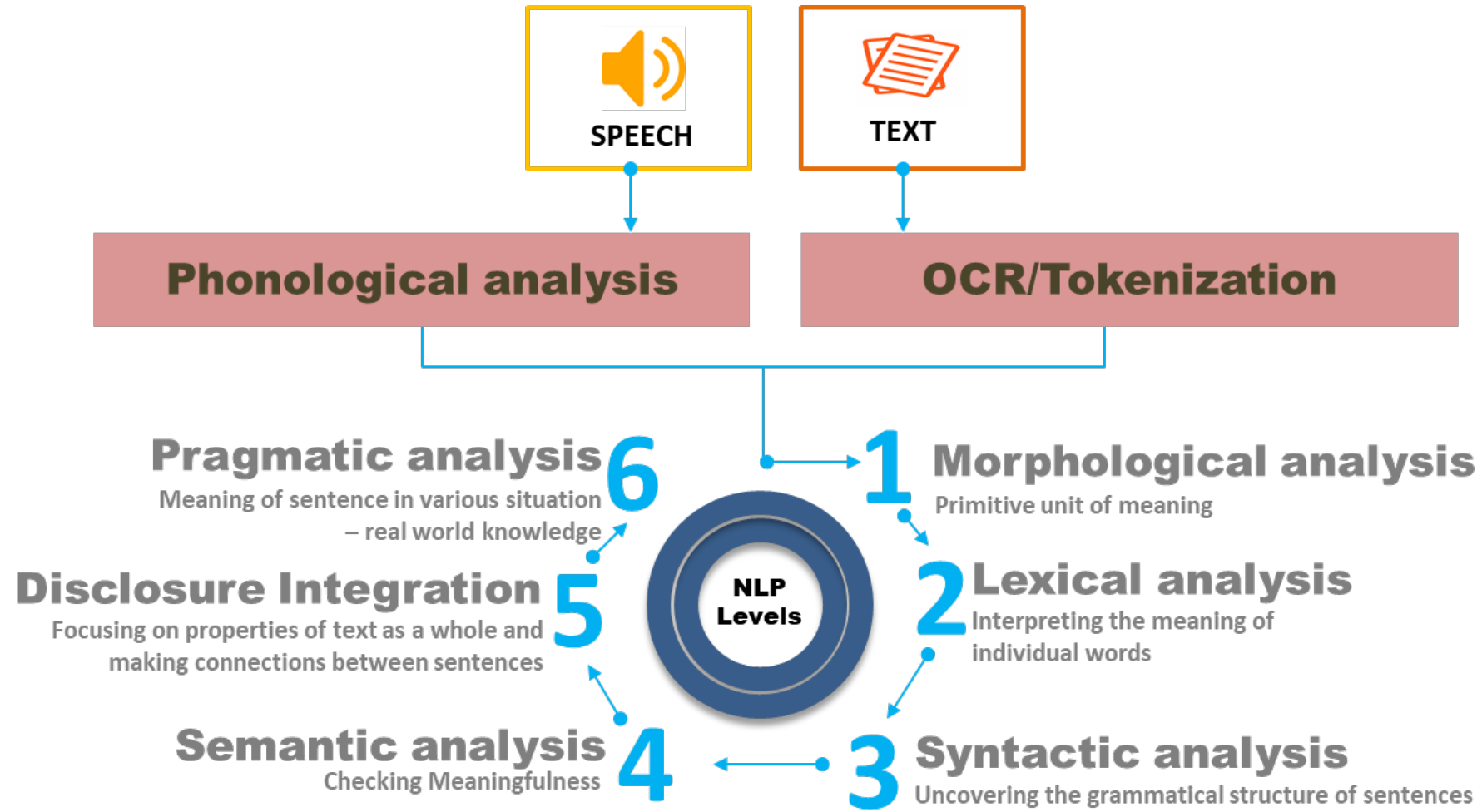


# NLP Goals

- The main goal of natural language processing (NLP) is to design and build computer systems that are able to
  - Process and analyze natural languages like Thai or English,
  - Understand the contents of data inputs (e.g., speech), and
  - Generate their outputs in a natural language.

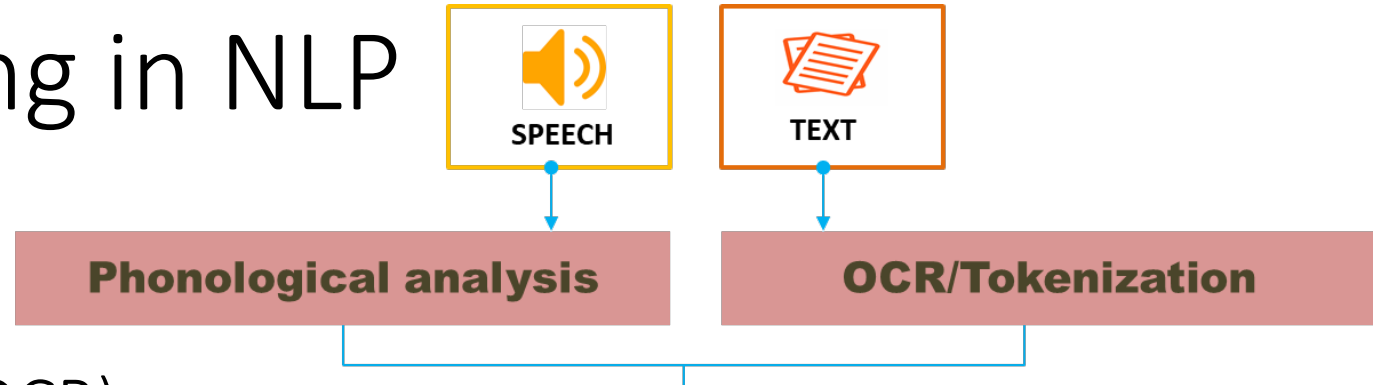


# Level of understanding in NLP



# Level of understanding in NLP

- Phonological Analysis:
  - Interpreting speech sounds.
- Optical Character Recognition(OCR):
  - OCR is the mechanical conversion of images of typed, handwritten or printed text into machine-encoded text. Also, from a scanned document, a photo of a document.
- Tokenization:
  - It is the first step in any NLP.
  - A tokenizer breaks unstructured data and natural language text into chunks of information.
    - breaks text paragraph into sentences/words.



# Level of understanding in NLP



- Morphological Analysis:
  - It studies and understanding the structure of words.
  - It identifies how a word is produced through the use of morphemes that is the smallest units of meanings.
  - It can broken down words into three morphemes (prefix, stem, and suffix). , e.g., the word: “unhappiness ”.
    - Prefix: un-
    - Stem: happy
    - Suffix: -ness

# Level of understanding in NLP

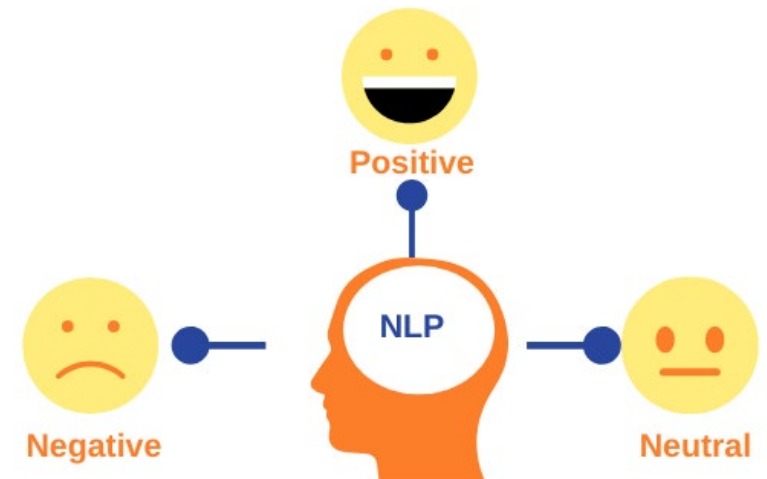


- Lexical Analysis:
  - Analyzing the structure of words
  - Dividing the whole text into paragraphs, sentences, and words.
  - Two techniques are used as follows:
    - Stemming:
      - This technique is to reduce words to their dictionary root.
      - Stemming identifies the common root form of a word by removing or replacing word suffixes (e.g. “flooding” is stemmed as “flood”)
    - Lemmatization:
      - This technique is to reduce and consider the meaning of the word in the evaluation.
      - Lemmatization identifies the inflected forms of a word and returns its base form (e.g. “better” is lemmatized as “good”).

# Level of understanding in NLP



- Syntactic Analysis (Parsing):
  - It is the process of analyzing the natural language with the rules of formal grammar to find out the dictionary meaning of any sentence.
  - Understanding in the sentence patterns of language (Subject, Verb, Object, Preposition)
- Semantic Analysis:
  - Understanding the context of any text and understanding the emotions.
  - It is used in tools such as machine translations, chatbots, search engines and text analytics.



# Level of understanding in NLP



- Discourse Analysis:
  - Focuses on the properties of the text as a whole that convey meaning by making connections between component sentences.
  - It focus on any aspect of linguistic behaviors, e.g.,
    - Study of particular patterns of pronunciation,
    - Sentence structure,
    - Semantic representation,
    - Ambiguity resolution
  - Example: **John** go to school, **he** loves NLP course.

# Level of understanding in NLP



- Pragmatic Analysis:
  - It analyze what the given text basically means.
  - It explains how extra meaning is read in text.
  - This requires much world knowledge (i.e., the understanding of intentions, plans, and goals).
  - For examples:
    - “Close the window?”
    - “Do you have a watch?”

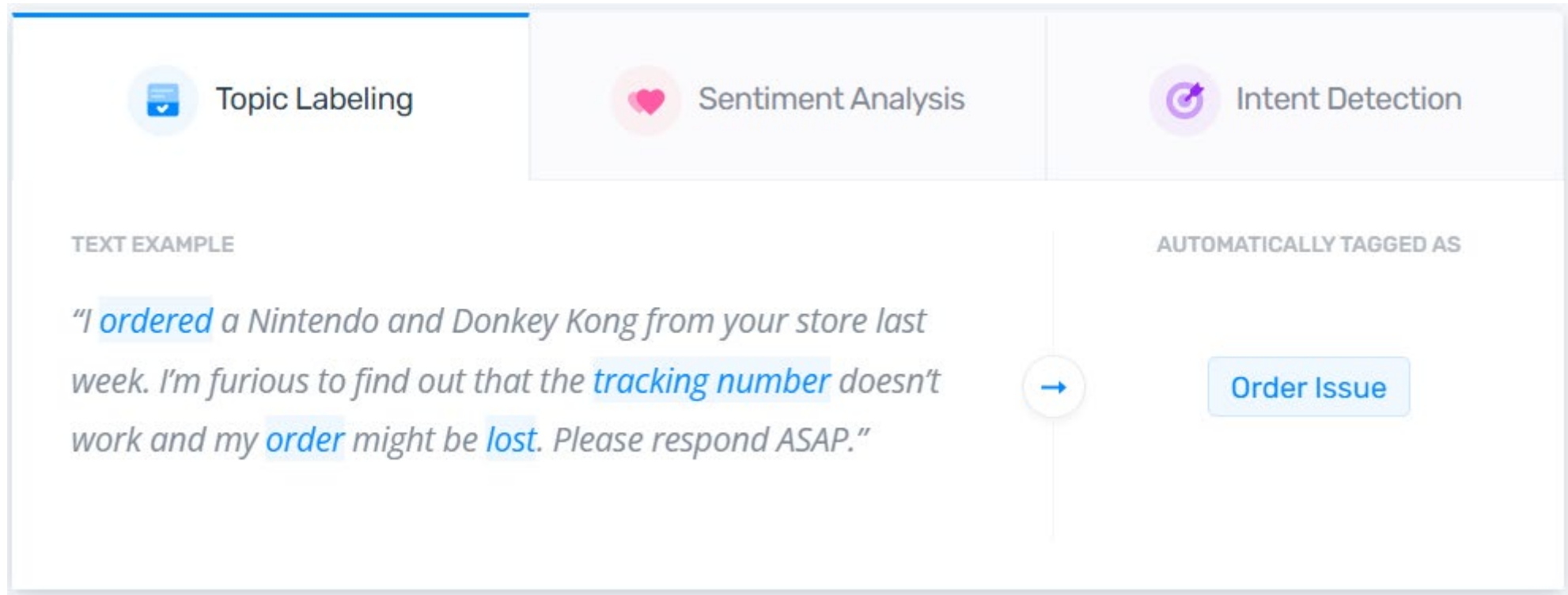
# Example of NLP Tools & Services

1. [MonkeyLearn](#) | NLP made simple
2. [Aylien](#) | Leveraging news content with NLP
3. [IBM Watson](#) | A pioneer AI platform for businesses
4. [Google Cloud NLP API](#) | Google technology applied to NLP
5. [Amazon Comprehend](#) | An AWS service to get insights from text
6. [NLTK](#) | The most popular Python library
7. [Stanford Core NLP](#) | Stanford's fast and robust toolkit
8. [TextBlob](#) | An intuitive interface for NLTK
9. [SpaCy](#) | Super-fast library for advanced NLP tasks
10. [GenSim](#) | State-of-the-art topic modeling

<https://monkeylearn.com/blog/natural-language-processing-tools/>



# Example of NLP Tools: MonkeyLearn



The screenshot displays the MonkeyLearn web interface for text classification. At the top, there are three tabs: 'Topic Labeling' (selected), 'Sentiment Analysis', and 'Intent Detection'. Below the tabs, the 'TEXT EXAMPLE' section contains a customer complaint: "I **ordered** a Nintendo and Donkey Kong from your store last week. I'm furious to find out that the **tracking number** doesn't work and my **order** might be **lost**. Please respond ASAP." The words 'ordered', 'tracking number', 'order', and 'lost' are highlighted in blue. An arrow points from this text to the 'AUTOMATICALLY TAGGED AS' section, which shows a single tag: 'Order Issue'.

Topic Labeling Sentiment Analysis Intent Detection

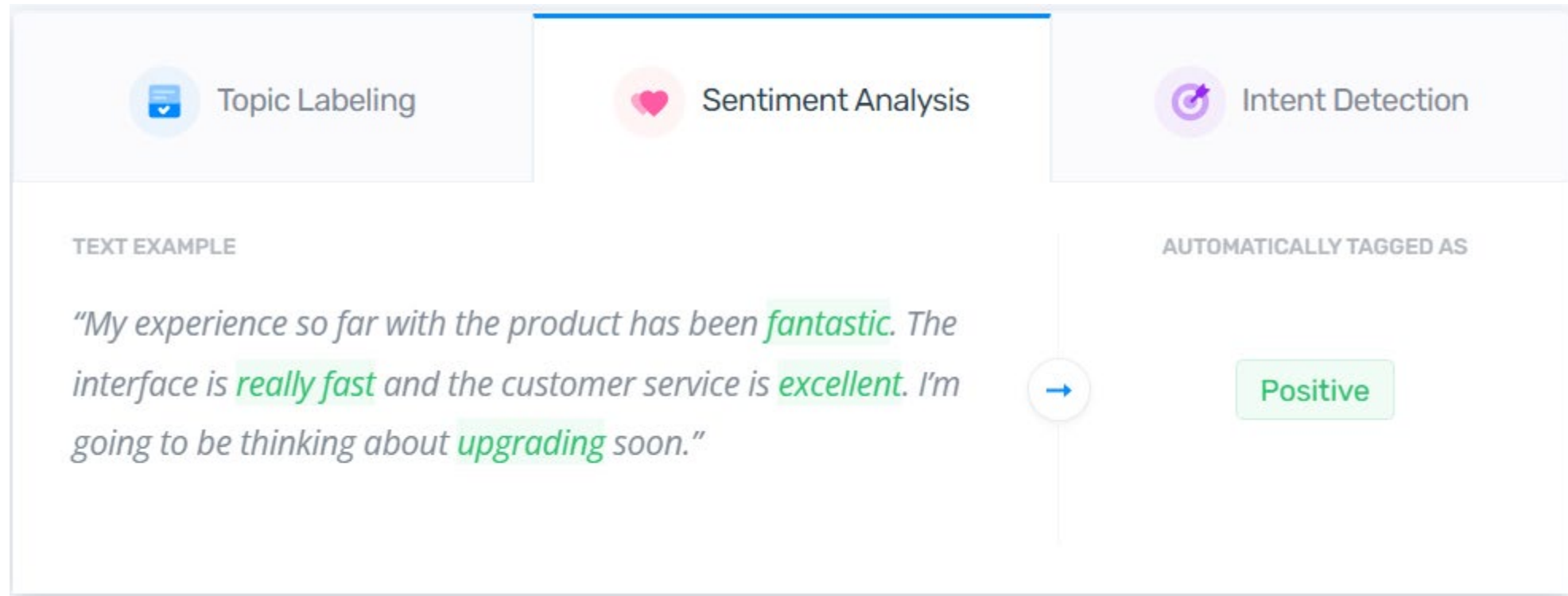
TEXT EXAMPLE

*"I **ordered** a Nintendo and Donkey Kong from your store last week. I'm furious to find out that the **tracking number** doesn't work and my **order** might be **lost**. Please respond ASAP."*

AUTOMATICALLY TAGGED AS

Order Issue

# Example of NLP Tools: MonkeyLearn



# Example of NLP Tools: MonkeyLearn

The screenshot displays the MonkeyLearn web interface. At the top, there are three tabs: 'Topic Labeling' (with a document icon), 'Sentiment Analysis' (with a heart icon), and 'Intent Detection' (with a target icon). The 'Intent Detection' tab is currently selected. Below the tabs, on the left, is a 'TEXT EXAMPLE' section containing the text: "Hi Jack, thanks for the email, your platform looks promising. Can we schedule a call tomorrow to see a demo? Please let me know when you are available. Thanks, Koko." The words 'looks promising', 'schedule a call', and 'see a demo?' are highlighted in purple. On the right, under the heading 'AUTOMATICALLY TAGGED AS', there is a single purple button labeled 'Interested in Demo'. A blue arrow points from the text example to this button.

# Example of NLP Tools: MonkeyLearn

The screenshot displays the MonkeyLearn interface for Feature Extraction. At the top, there are three tabs: 'Feature Extraction' (selected), 'Keyword Extraction', and 'Entity Extraction'. Below the tabs, a 'TEXT EXAMPLE' is provided: "The specs of the laptop are: Refurbished Dell Black 14" E6420 with Intel Core i5 Processor, 6GB Memory, 320GB Hard Drive and Windows 10 Home". To the right, under 'AUTOMATICALLY TAGGED AS', the system has identified several entities: 'Dell' as a 'BRAND', '14"' as 'SCREEN SIZE', 'Intel Core i5' as 'CPU', '6GB' as 'RAM', and '320GB' as 'HARD DRIVE'. Each entity is shown in a colored box with its category label in a smaller box next to it.

Feature Extraction

Keyword Extraction

Entity Extraction

TEXT EXAMPLE

"The specs of the laptop are: Refurbished Dell Black 14" E6420 with Intel Core i5 Processor, 6GB Memory, 320GB Hard Drive and Windows 10 Home"

AUTOMATICALLY TAGGED AS

Dell BRAND

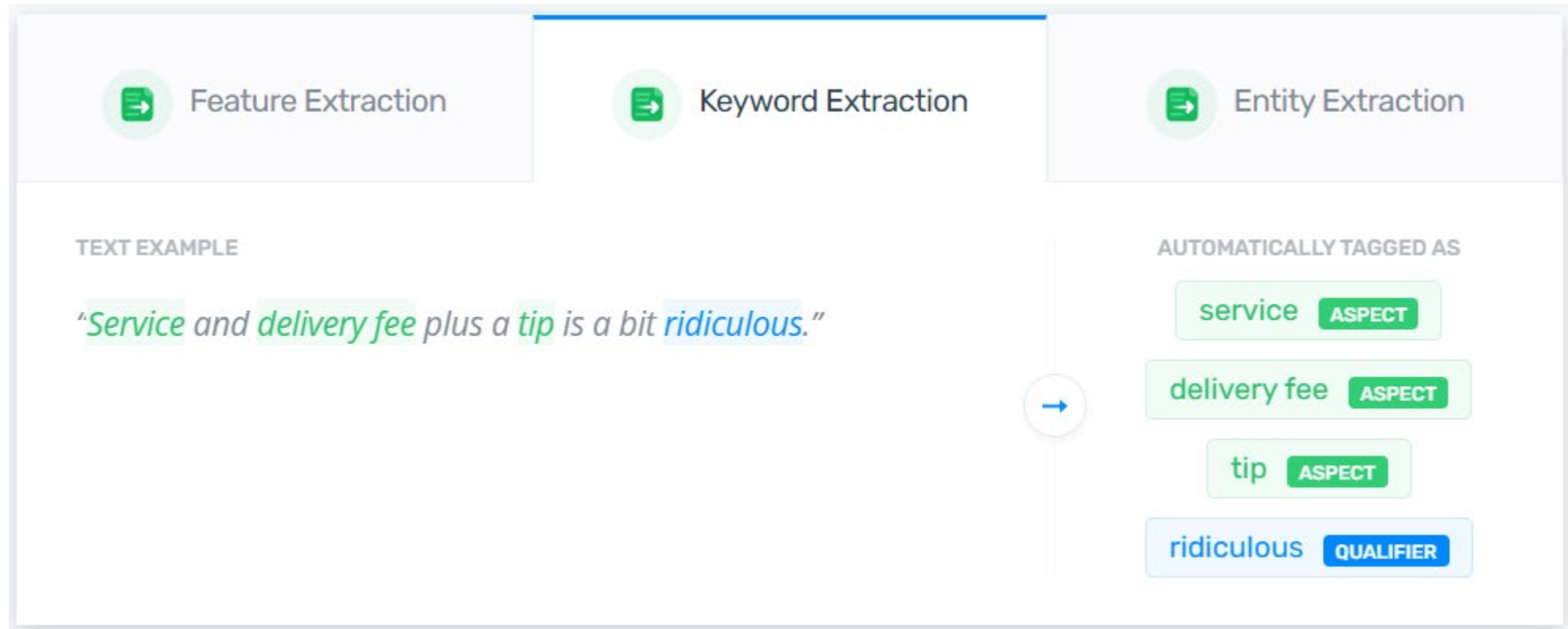
14" SCREEN SIZE

Intel Core i5 CPU

6GB RAM

320GB HARD DRIVE

# Example of NLP Tools: MonkeyLearn



# Example of NLP Tools: MonkeyLearn

The screenshot displays the MonkeyLearn interface for text analysis. At the top, there are three tabs: 'Feature Extraction', 'Keyword Extraction', and 'Entity Extraction'. The 'Entity Extraction' tab is selected. Below the tabs, a text example is provided: "Ron Gilbert from LucasArts had two inspirations that led the adventures of Guybrush Threepwood. One was a novel written by Tim Powers, and the other was a ride at Disneyland." The text is automatically tagged with entities. On the right, a list of entities is shown, each with a name and a category label in a colored box: Ron Gilbert (PEOPLE), LucasArts (COMPANIES), Guybrush Threepwood (PEOPLE), Tim Powers (PEOPLE), and Disneyland (PLACES). A blue arrow points from the text example to the entity list.

Feature Extraction   Keyword Extraction   Entity Extraction

TEXT EXAMPLE

*"Ron Gilbert from LucasArts had two inspirations that led the adventures of Guybrush Threepwood. One was a novel written by Tim Powers, and the other was a ride at Disneyland."*

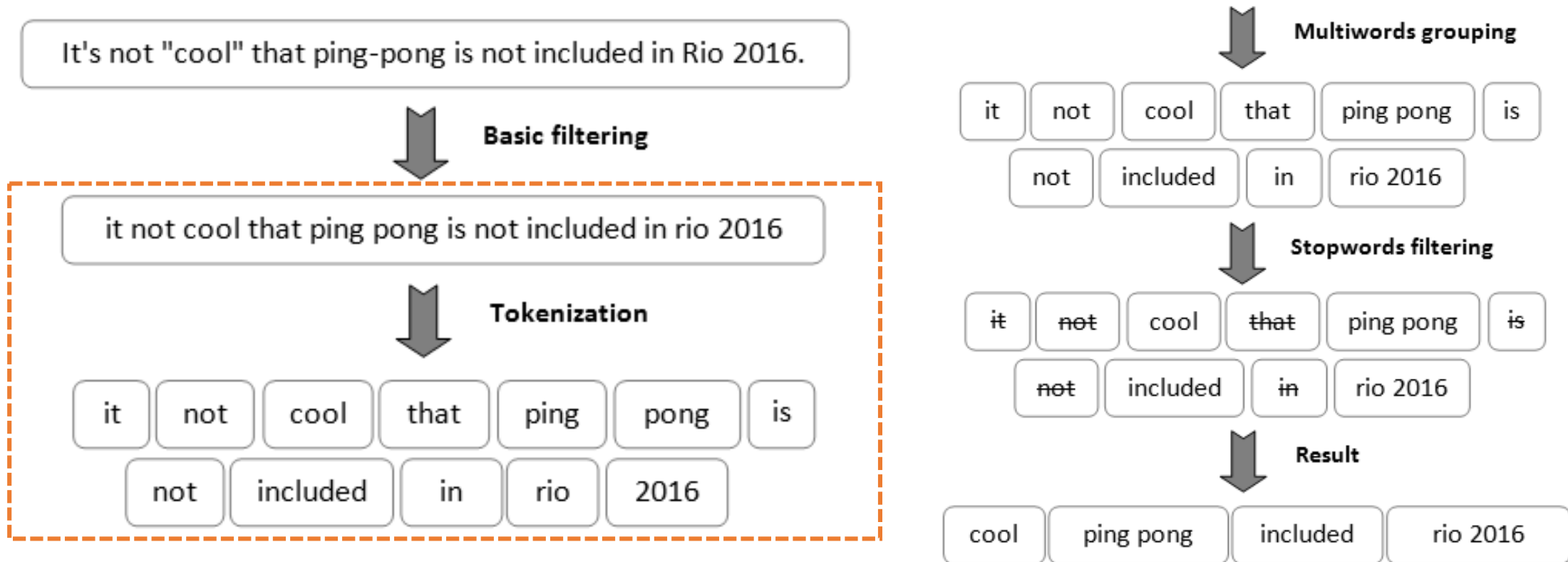
AUTOMATICALLY TAGGED AS

- Ron Gilbert PEOPLE
- LucasArts COMPANIES
- Guybrush Threepwood PEOPLE
- Tim Powers PEOPLE
- Disneyland PLACES

# NLP Libraries: NLTK

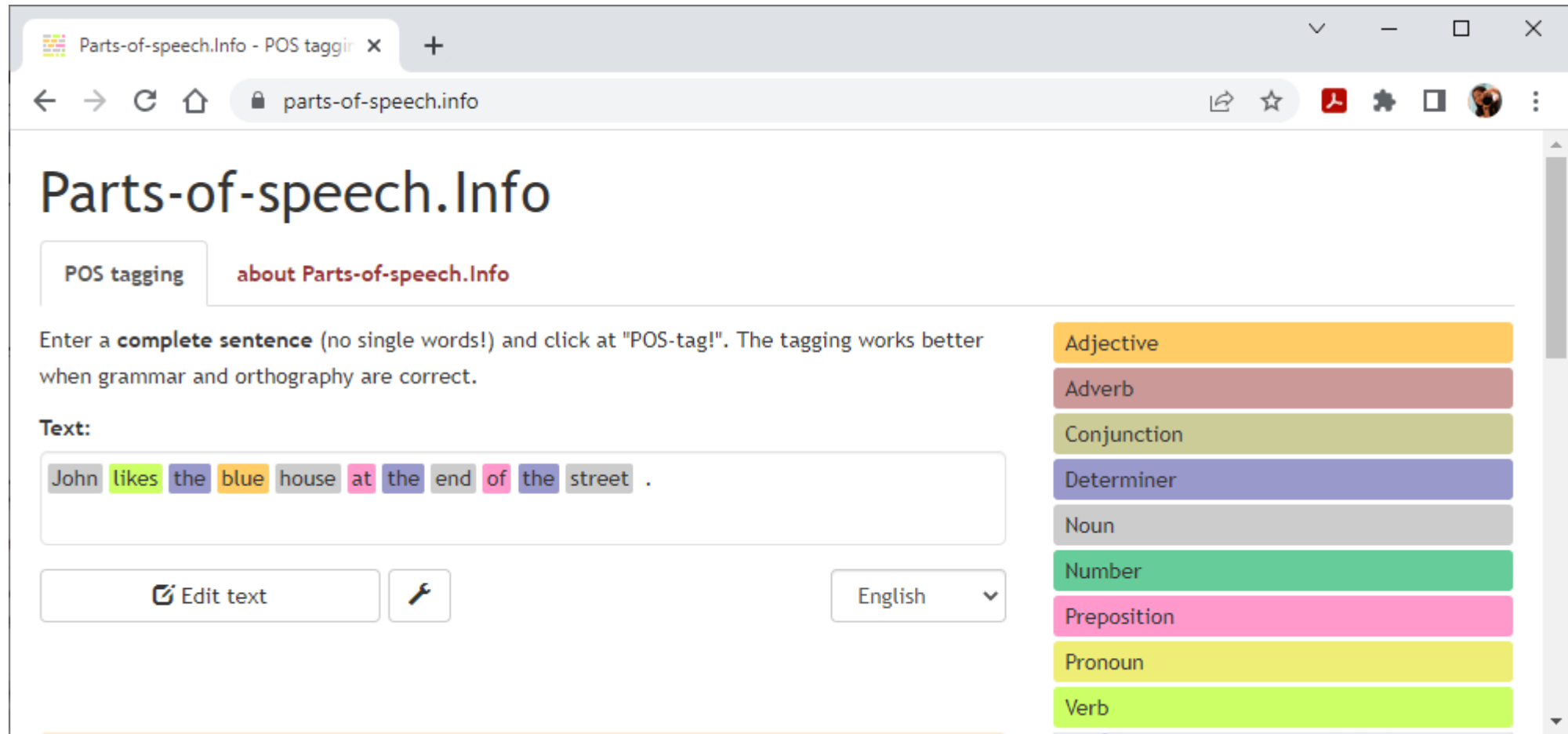
- Some of the features of NLTK according to Real Python
  - **Tokenizing:**
    - It is used to split any text by word or by sentence. This allows the user to work with small pieces of coherent texts.
  - **Filtering Stop Words:**
    - It is used to ignore stop words while processing any text. Common words like in, is, an, etc., are often stop words.
  - **Stemming:**
    - It is used to reduce any word to its root word. It helps the computer to understand the meaning of the word.
  - **Tagging Parts of Speech (POS):**
    - It is used to label word in a sentence according to their parts of speech.
  - **Name Entity Recognition (NER):**
    - It is used to locate named entities in text and determine what type of named entity they are.

# Example of Text tokenization & multiword





# Example of POS Tagging



The screenshot shows a web browser window with the address bar displaying "parts-of-speech.info". The page title is "Parts-of-speech.Info". Below the title, there are two tabs: "POS tagging" (selected) and "about Parts-of-speech.Info".

The main content area contains the following text:

Enter a **complete sentence** (no single words!) and click at "POS-tag!". The tagging works better when grammar and orthography are correct.

**Text:**

John likes the blue house at the end of the street .

Below the text input field, there is an "Edit text" button, a settings icon (wrench), and a language dropdown menu set to "English".

On the right side of the page, there is a vertical list of part-of-speech categories, each with a corresponding colored bar:

- Adjective (orange)
- Adverb (brown)
- Conjunction (olive)
- Determiner (purple)
- Noun (grey)
- Number (teal)
- Preposition (pink)
- Pronoun (yellow)
- Verb (light green)

# Example of Named Entity Recognition

**Person:** Michael Jackson, Oprah Winfrey, Barack Obama, Susan Sarandon

**Location:** Canada, Honolulu, Bangkok, Brazil, Cambridge

**Organization:** Samsung, Disney, Yale University, Google

**Time:** 15.35, 12 PM,

Other categories include Numerical values, Expression, E-Mail Addresses, and Facility.

Apple<sub>ORG</sub> today<sub>DATE</sub> announced the  
second<sub>QUANTITY</sub> generation iPhone SE<sub>COMM</sub>  
a powerful new iPhone<sub>COMM</sub> featuring  
a 4.7-inch<sub>QUANTITY</sub> Retina HD display.

# Examples of NLP Technology



**Smart Assistants**



**Search Results**



**Predictive text**



**Language Translations**



**Digital Phone Calls**



**Text Analytics**



**Virtual Assistants**



**Detecting Duplications**



**Social Media Monitoring**



**Marketing Strategies**

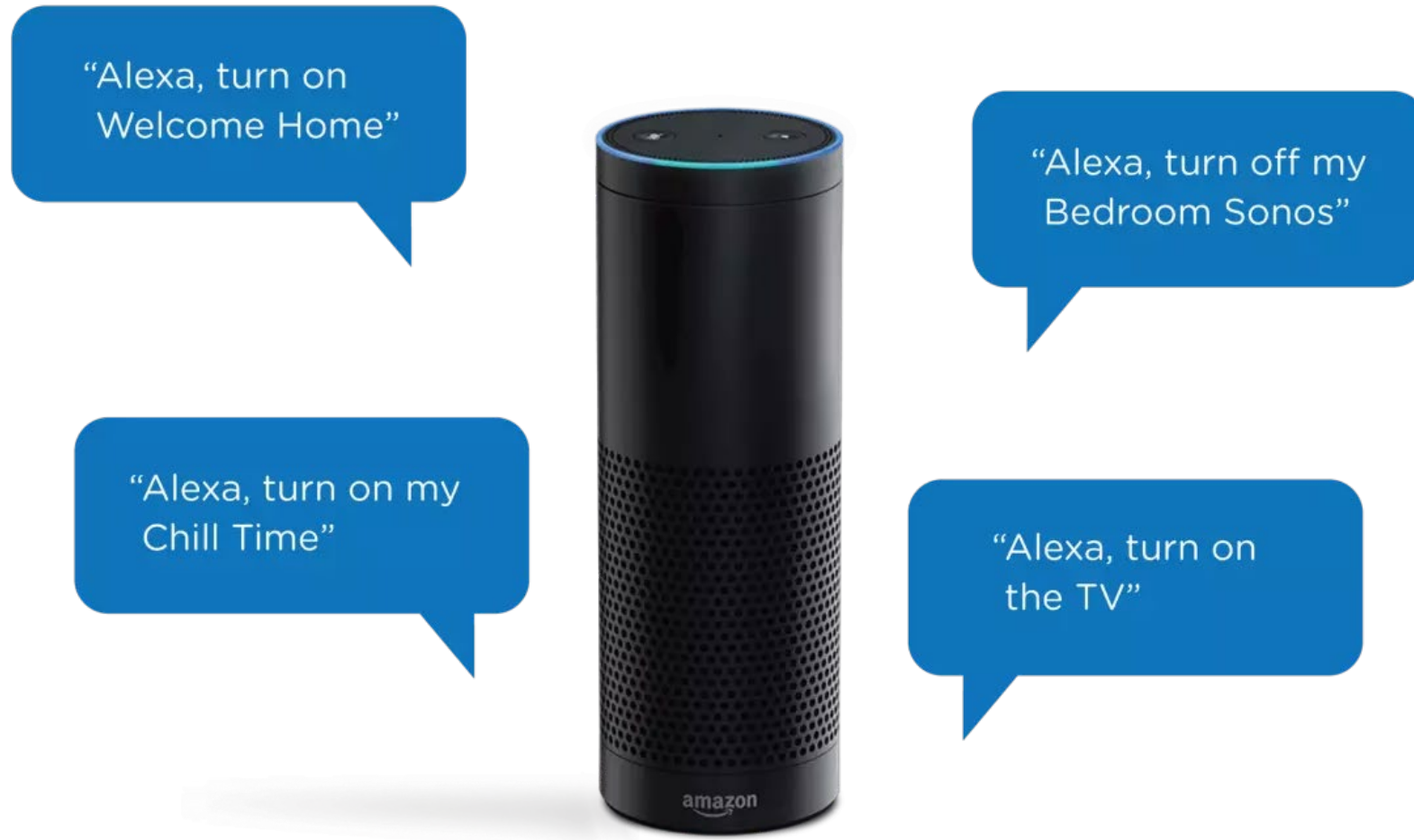


**Descriptive Analytics**



**Automatic Insights**

# Example of Smart Assistants: Amazon Alexa



# How does Alexa work?

- Example of some steps that Amazon do
  - breaks down your “orders” into individual sounds
  - consults a database containing various words’ pronunciations
  - find words most closely correspond to the combination of individual sounds.
  - identifies important words to make sense of the tasks and carry out corresponding functions.
  - Amazon’s servers send the information back to your device. Alexa may speak back.
- Amazon records your words,
  - It is sent to Amazon’s servers to be analyzed more efficiently.



# Homework

Present an Example of NLP in next week