



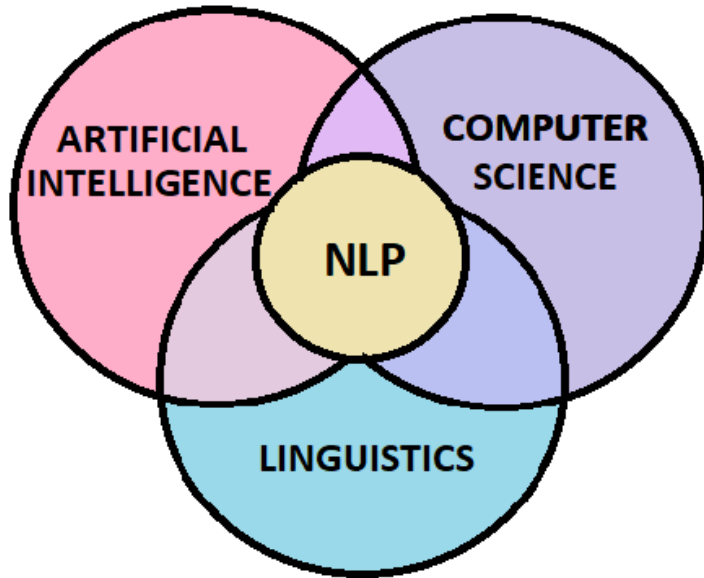
Chapter 1

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?



Reference:
<https://www.analyticsvidhya.com/blog/2021/09/essential-text-pre-processing-techniques-for-nlp/>

- Natural Language Processing (NLP) is a subfield in
 - Artificial Intelligence: รวมความฉลาดของมนุษย์สู่เครื่องจักร,
 - Linguistics: ศึกษาภาษาเพื่อให้เข้าใจระบบของภาษามนุษย์ โดยใช้แนวคิด ทฤษฎี และวิธีการวิจัยที่เป็นวิทยาศาสตร์,
 - Cognitive Science: ศึกษาความคิด การเรียนรู้ การเกิดปัญญา การตัดสินใจ
 - Computer Science: นำความรู้ด้านการคำนวณ การเขียนโปรแกรม เข้าด้วยกันเพื่อใช้แก้ไขปัญหาต่าง ๆ

that 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**).

ตัวอย่างของ DBMS มีอะไรบ้าง???

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>

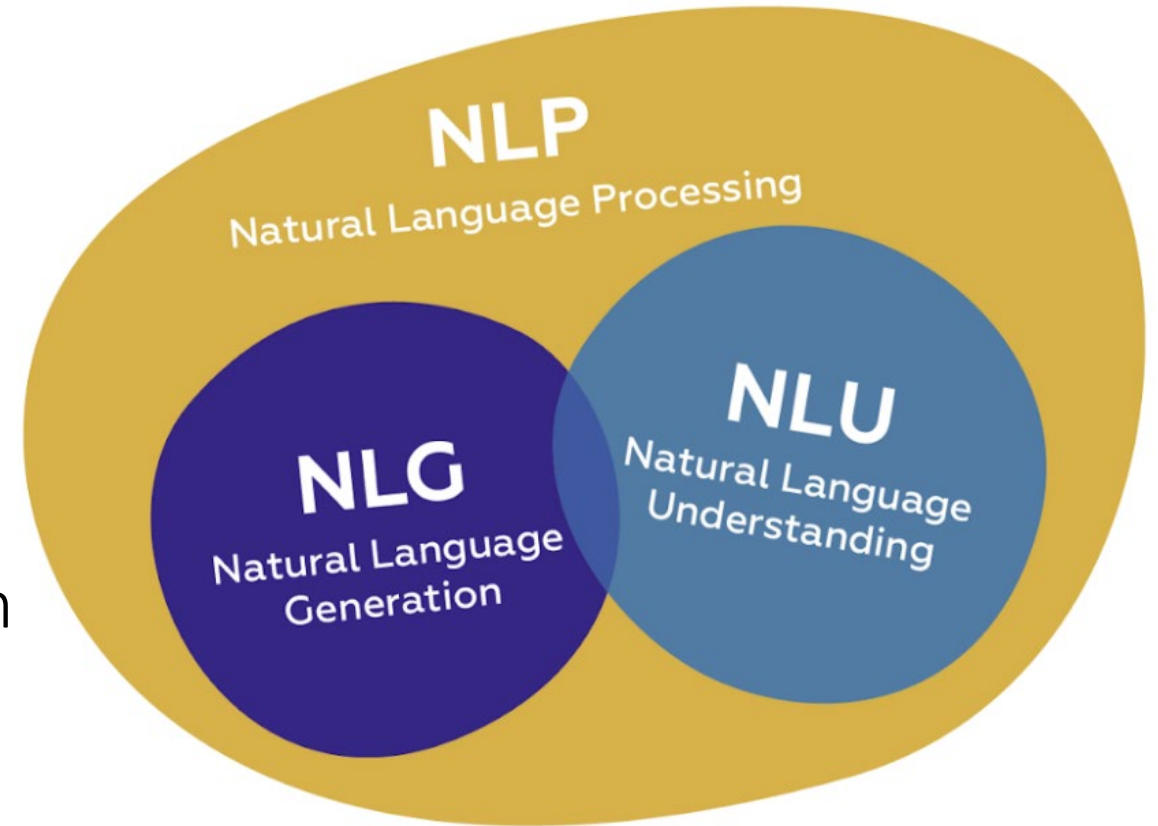
<ITEM>EGGS</>

<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.



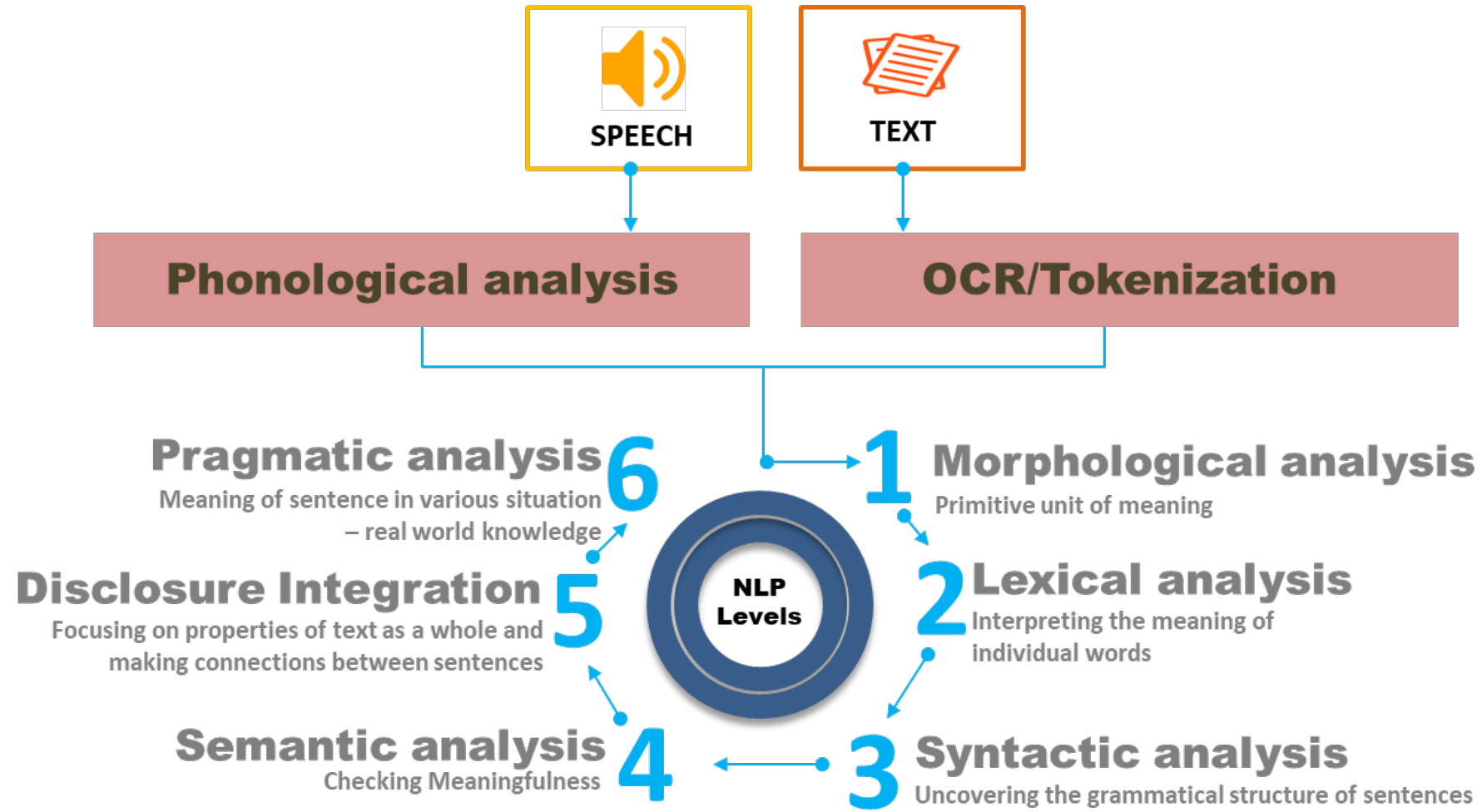
What are NLP Goals?

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



Level of understanding in NLP

- **Phonological Analysis:**

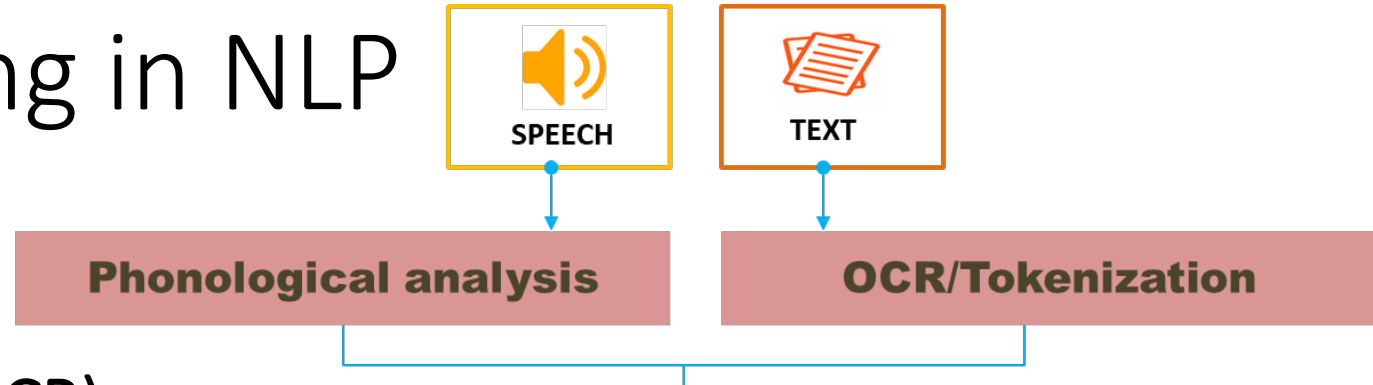
- Interpreting speech sounds.

- **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**.



Pragmatic analysis 6
Meaning of sentence in various situation
– real world knowledge

Disclosure Integration 5
Focusing on properties of text as a whole and
making connections between sentences

Semantic analysis 4
Checking Meaningfulness

NLP Levels

1 Morphological analysis
Primitive unit of meaning

2 Lexical analysis
Interpreting the meaning of
individual words

3 Syntactic analysis
Uncovering the grammatical structure of sentences

- 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- unhappy(adj) unhappiness (n)
 - Root or Stem: happy (adj)
 - Suffix: -ness happiness (n)

https://www.englishclub.com/vocabulary/prefixes.php#google_vignette

Level of understanding in NLP



2 Lexical Analysis:

- Analyzing the **structure of 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. “running” is stemmed as “**run**”, “studies” is stemmed as “**studi**”)
 - **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**” instead of “bet”, “studies” is lemmatized as “**study**”).

Level of understanding in NLP

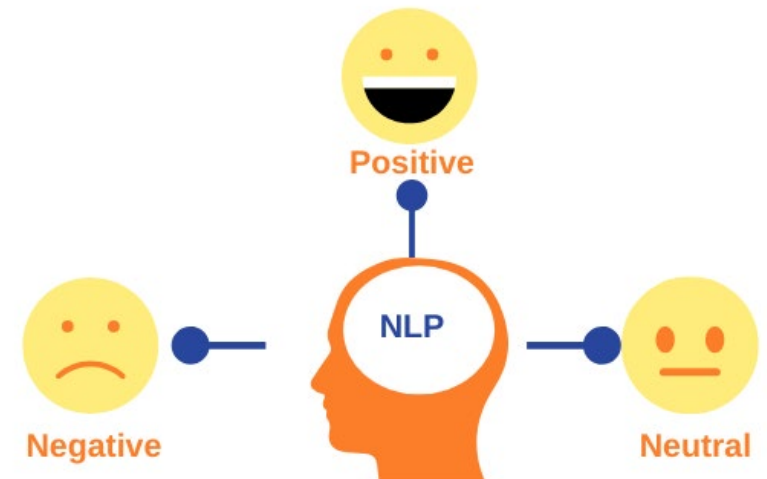


3 Syntactic Analysis (Parsing) (การวิเคราะห์ในเชิงโครงสร้าง):

- 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)

4 Semantic Analysis (การวิเคราะห์ในเชิงความหมาย):

- 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



5 Discourse Analysis (การวิเคราะห์ข้อความ):

- 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



6 Pragmatic Analysis (การวิเคราะห์ในเชิงตีความ):

- 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

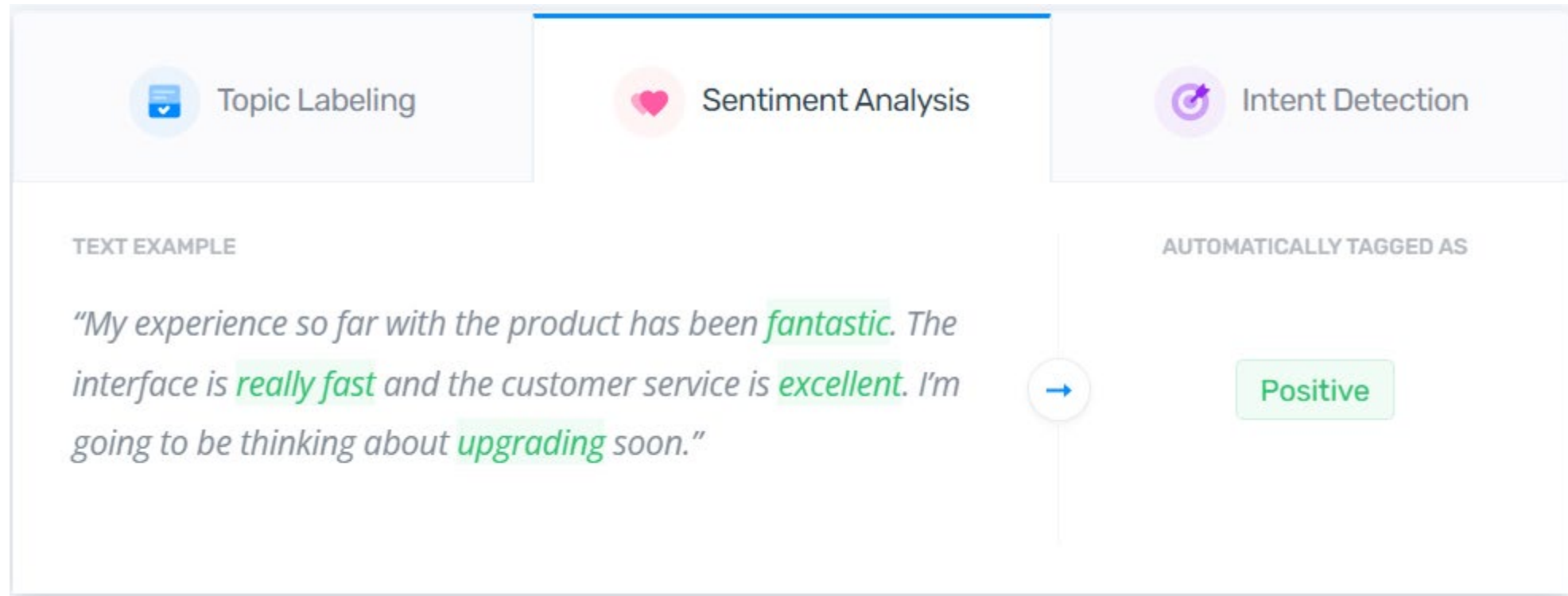
1. [MonkeyLearn](#) | NLP made simple
2. [Aylien](#) | Leveraging news content with NLP
3. [IBM Watson](#) | A pioneer AI platform for businesses
4. [Google Cloud NLP API](#) | Google technology applied to NLP
5. [Amazon Comprehend](#) | An AWS service to get insights from text
6. [NLTK](#) | The most 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/>

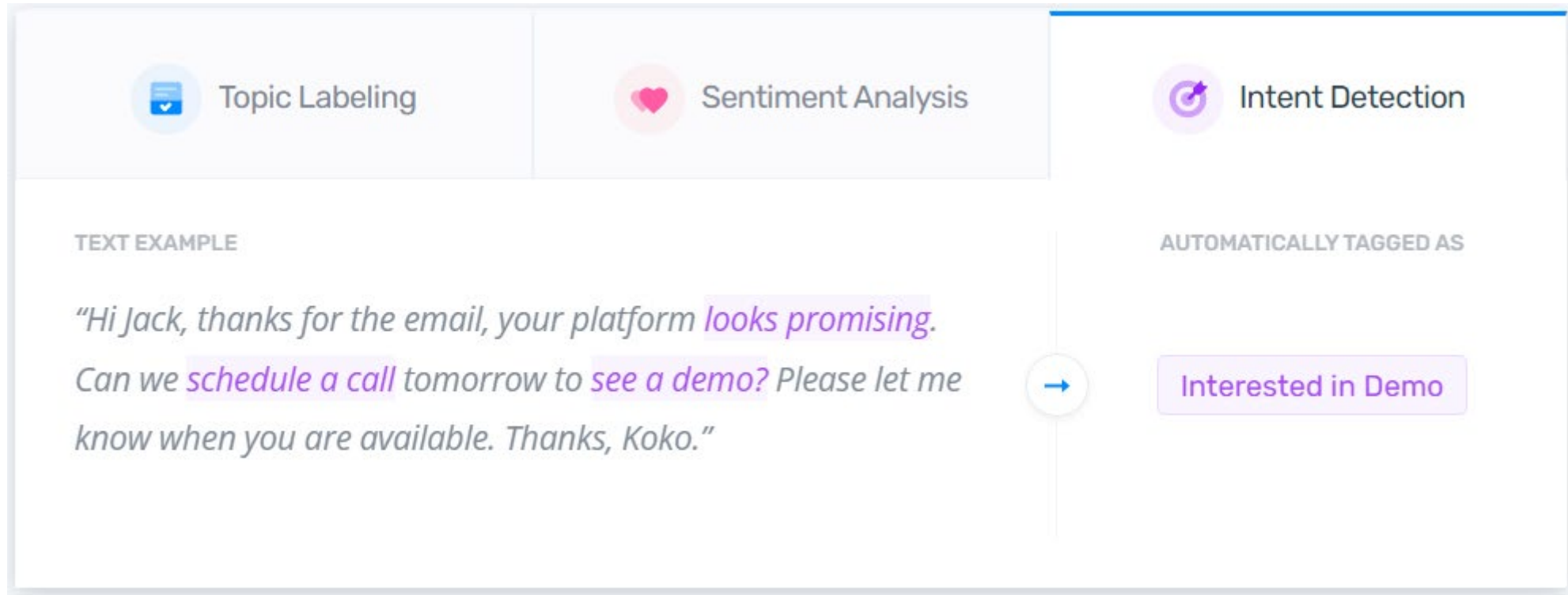
Example of NLP Tools: MonkeyLearn

The screenshot displays the MonkeyLearn NLP interface. At the top, there are three tabs: 'Topic Labeling' (with a document icon), 'Sentiment Analysis' (with a heart icon, currently selected), and 'Intent Detection' (with a target icon). Below the tabs, the interface is split into two main sections. The left section, titled 'TEXT EXAMPLE', 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. The right section, titled 'AUTOMATICALLY TAGGED AS', shows a single tag: 'Order Issue' in a blue box. A blue arrow points from the text example to the tag.

Example of NLP Tools: MonkeyLearn



Example of NLP Tools: MonkeyLearn



The screenshot displays the MonkeyLearn interface with three tool tabs at the top: '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, the 'TEXT EXAMPLE' section contains 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. A blue arrow points from the text to the 'AUTOMATICALLY TAGGED AS' section, which displays a purple button labeled 'Interested in Demo'.

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". The text is color-coded to match the extracted features. To the right, under the heading 'AUTOMATICALLY TAGGED AS', the extracted features are listed in colored boxes with their corresponding categories in smaller boxes: 'Dell' (BRAND), '14"' (SCREEN SIZE), 'Intel Core i5' (CPU), '6GB' (RAM), and '320GB' (HARD DRIVE). A blue arrow points from the text example to the extracted features.

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

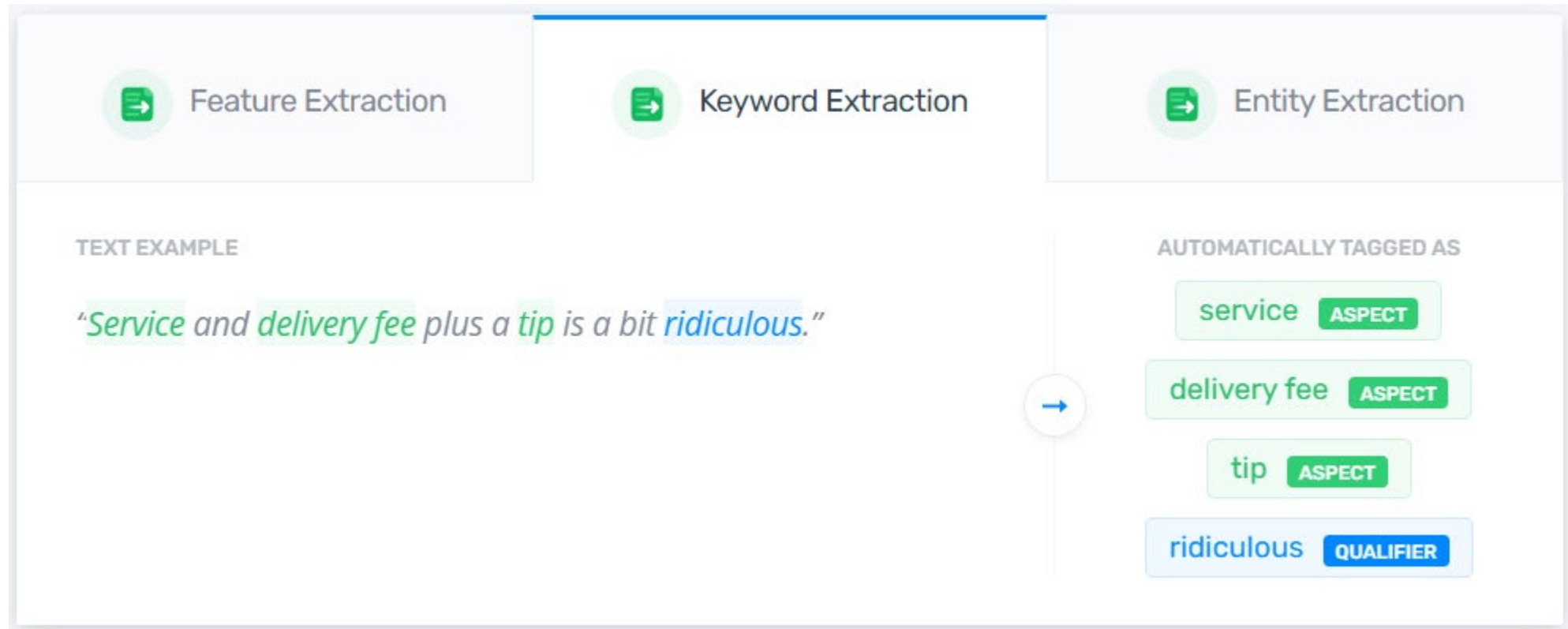
14" SCREEN SIZE

Intel Core i5 CPU

6GB RAM

320GB HARD DRIVE

Example of NLP Tools: MonkeyLearn



Example of NLP Tools: MonkeyLearn

The screenshot displays the MonkeyLearn web interface for text analysis. At the top, there are three tabs: 'Feature Extraction', 'Keyword Extraction', and 'Entity Extraction'. The 'Entity Extraction' tab is currently selected. Below the tabs, a text input area contains a sample paragraph: "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 various entities. On the right side, under the heading 'AUTOMATICALLY TAGGED AS', a list of extracted entities is shown, each with a colored label indicating its category: 'Ron Gilbert' (PEOPLE), 'LucasArts' (COMPANIES), 'Guybrush Threepwood' (PEOPLE), 'Tim Powers' (PEOPLE), and 'Disneyland' (PLACES). A blue arrow points from the text input area to the entity extraction results.

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."

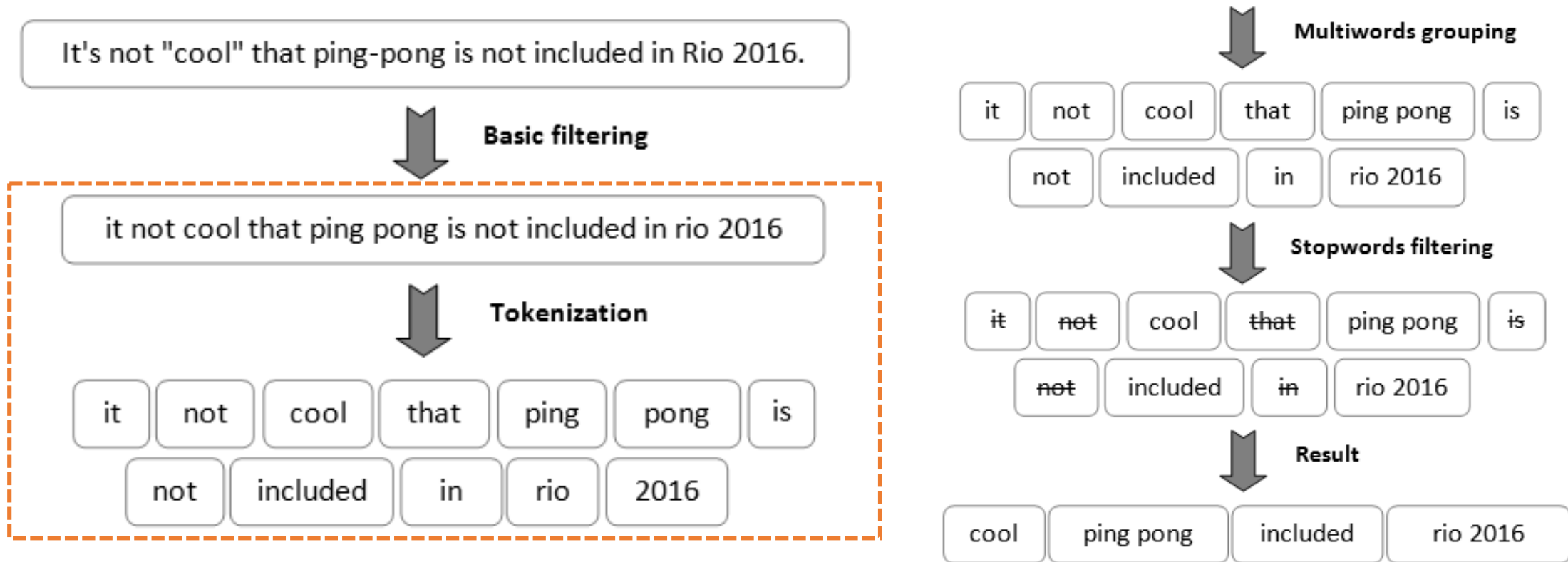
AUTOMATICALLY TAGGED AS

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

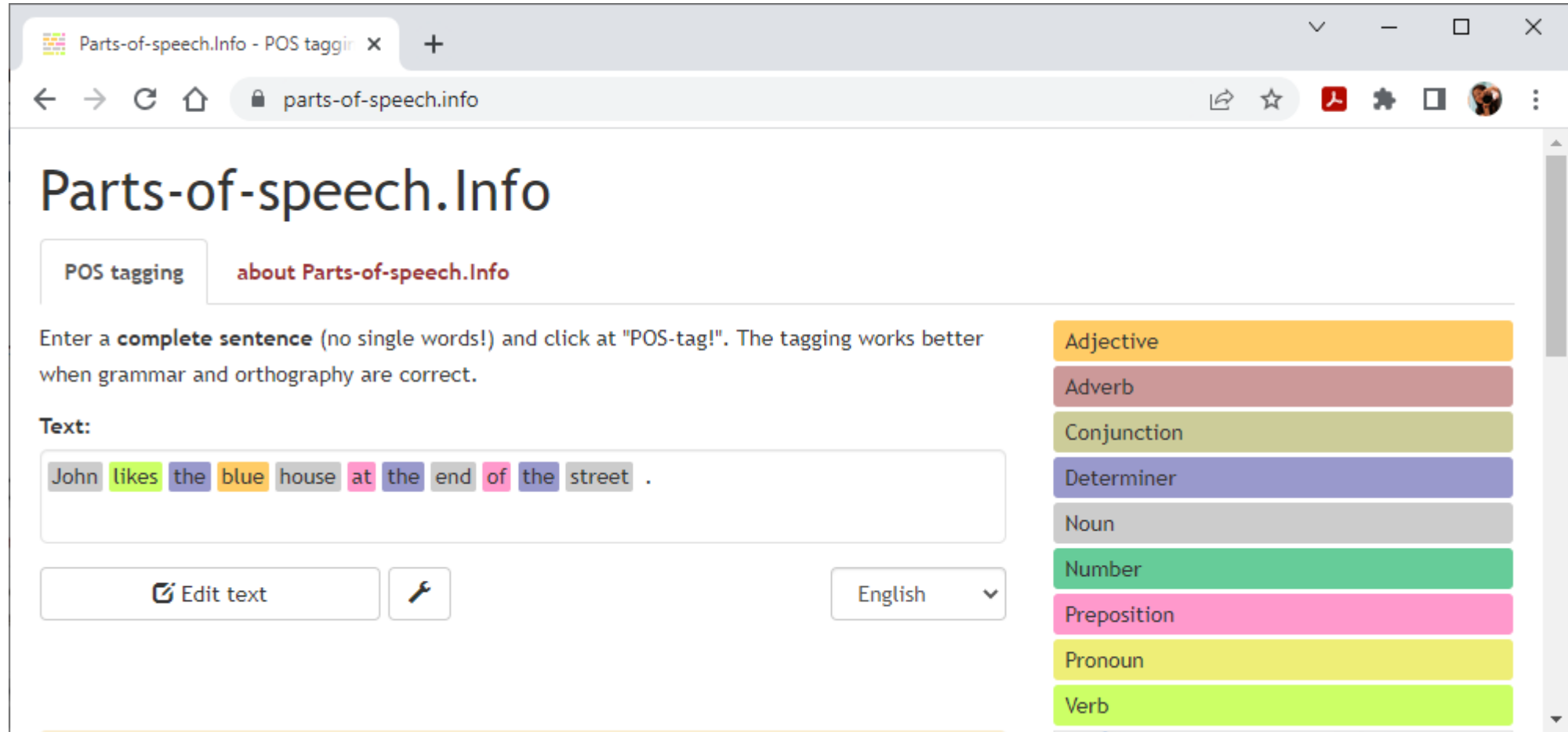
NLP Libraries: NLTK

- 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_{ORG} today_{DATE} announced the
second_{QUANTITY} generation iPhone SE_{COMM}
a powerful new iPhone_{COMM} featuring
a 4.7-inch_{QUANTITY} 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

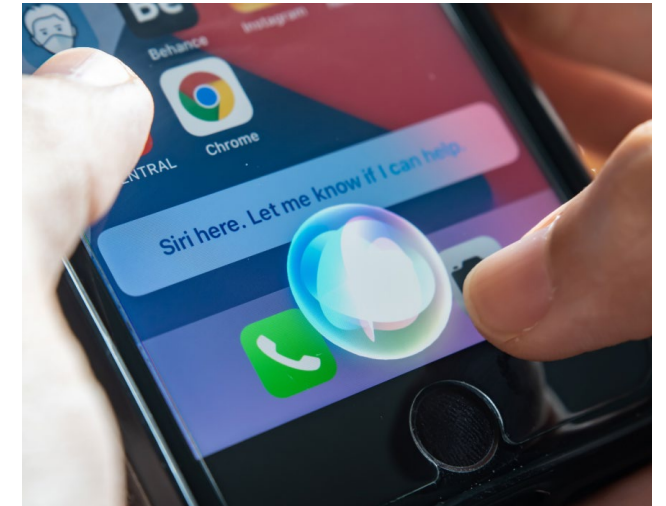
Example of Smart Assistants:



Amazon Alexa



Google Assistant



Siri



Homework

How does [Amazon Alexa](#) work?
[Google Assistant](#)
[Siri](#)