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 Science

  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).

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 →

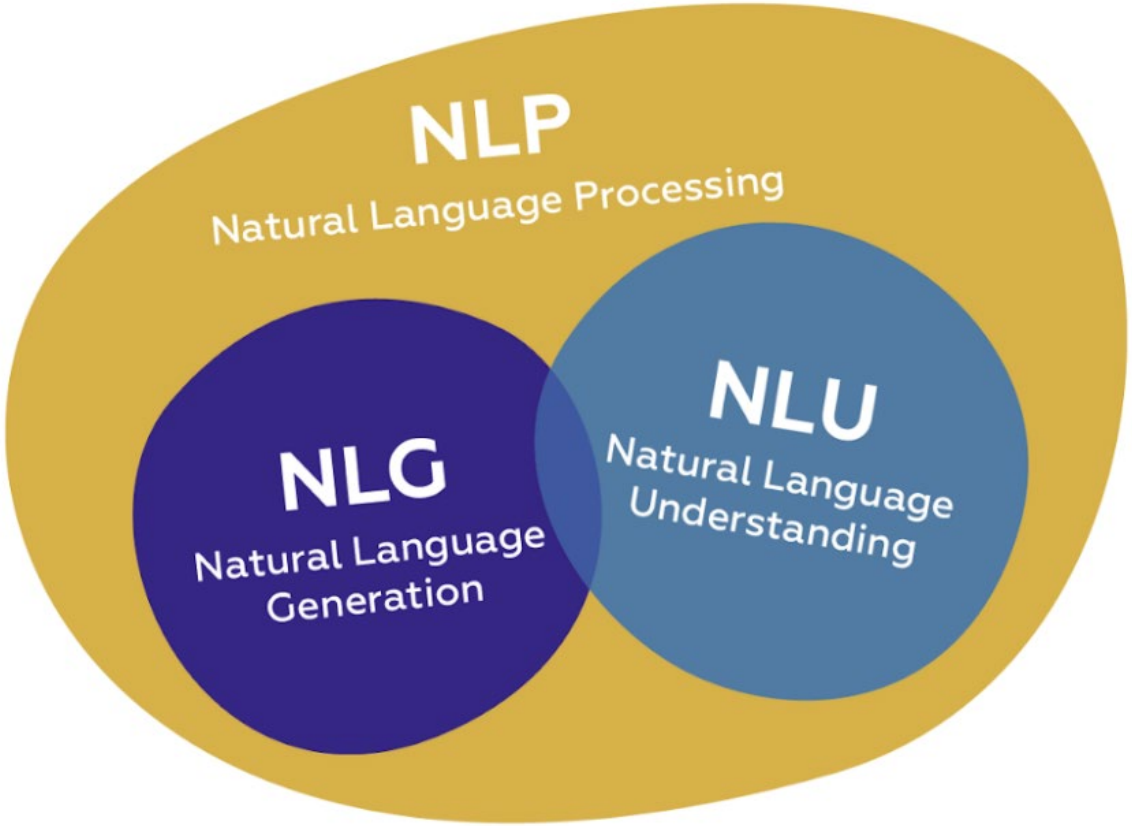
← NLG

- STRUCTURED -

<SHOPPIN LIST>
<ITEM>EGGS</>
<ITEM>MILK</>
</>

NLU vs NLG

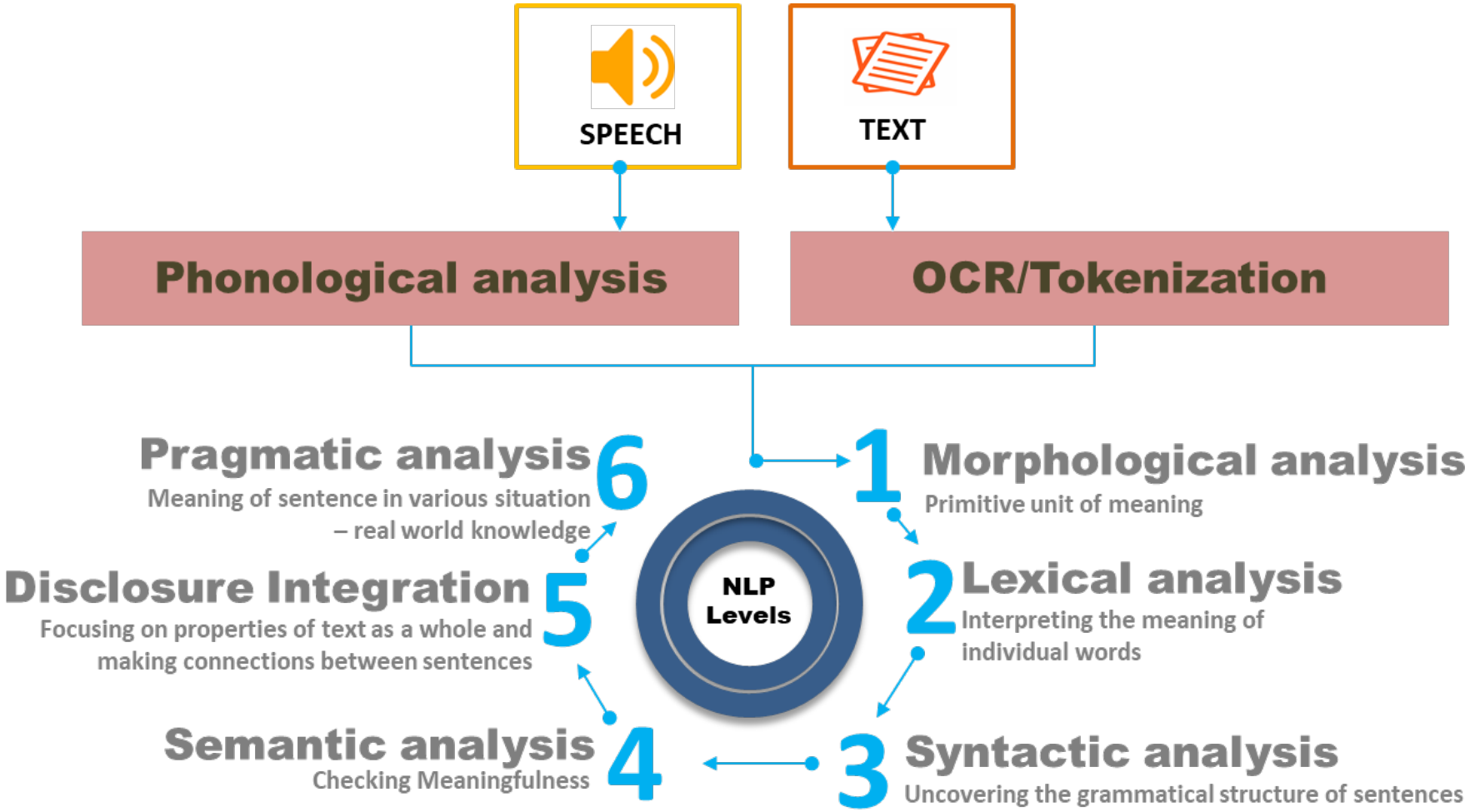
- Both are main branches of the NLP.
- NLU involves transforming human language into a machine-readable format.
- NLG involves the processing and conversation of the information from the computer language to the understandable human language.



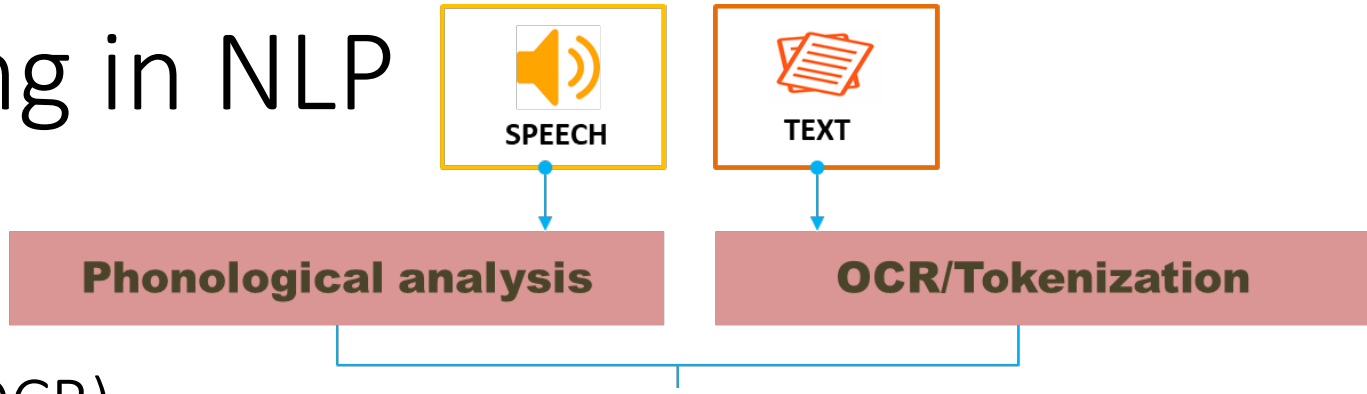
NLP Goals

- 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

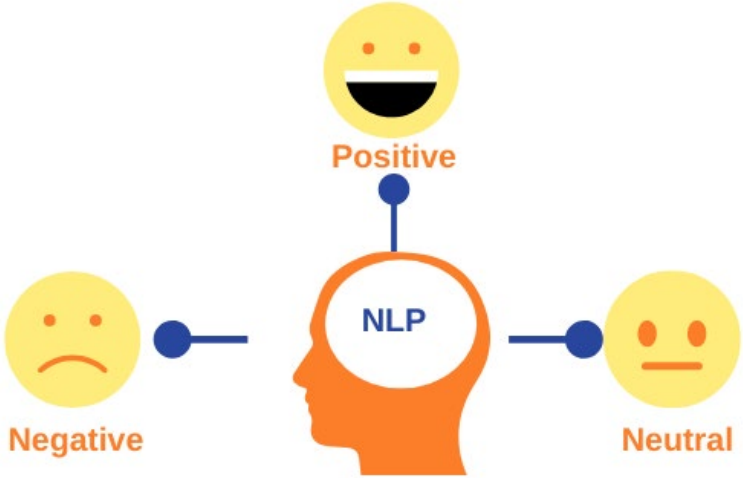
- Morphological Analysis:
  - It studies and understanding the structure of words.
  - It identifies how a word is produced through the use of morphemes that is the smallest units of meanings.
  - It can broken down words into three morphemes (prefix, stem, and suffix). , e.g., the word: “unhappiness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

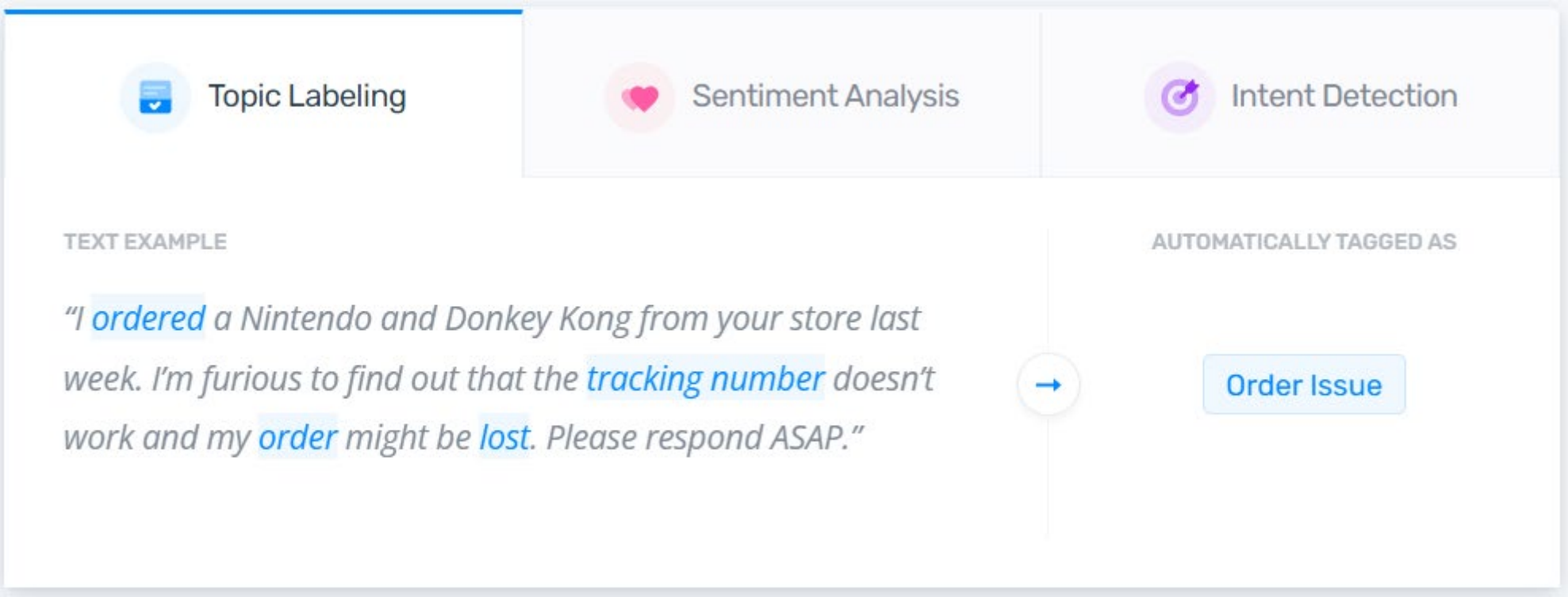
- Pragmatic Analysis:
  - It analyze what the given text basically means.
  - It explains how extra meaning is read in text.
  - This requires much world knowledge (i.e., the understanding of intentions, plans, and goals).
  - For examples:
    - “Close the window?"
    - “Do you have a watch?”

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

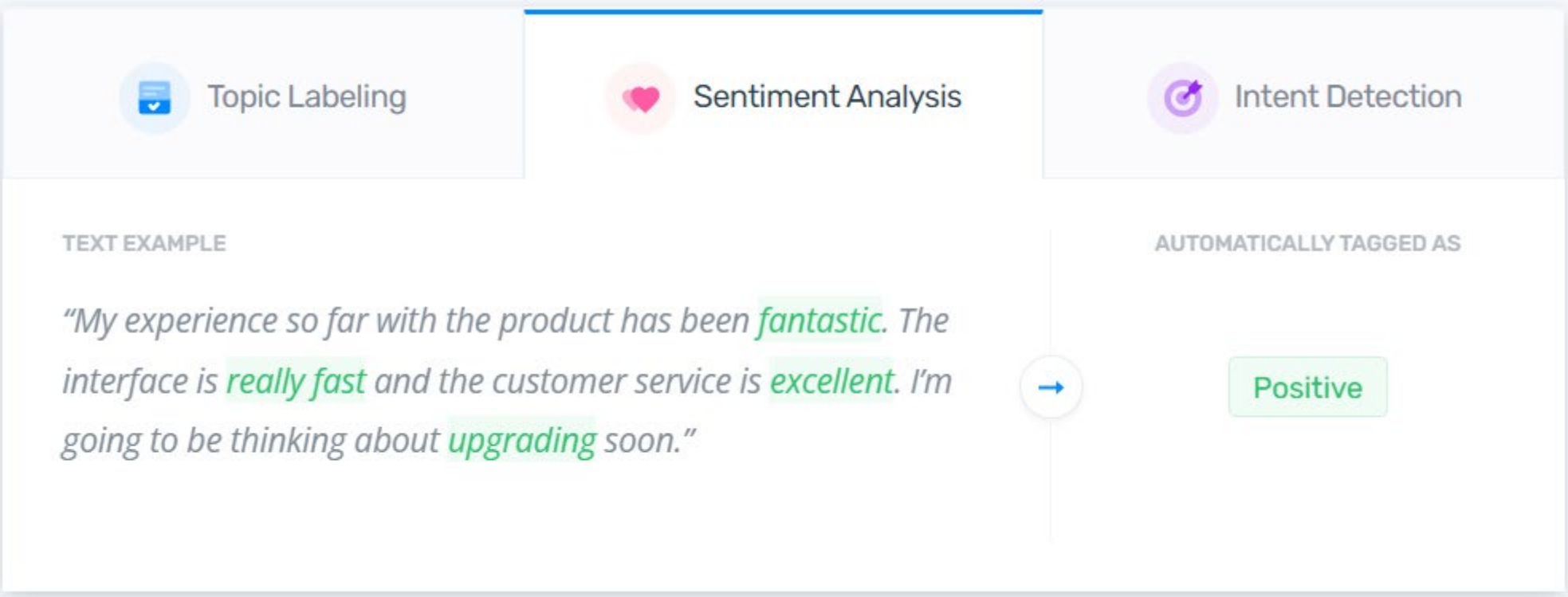
https://monkeylearn.com/blog/natural-language-processing-tools/

Example of NLP Tools: MonkeyLearn



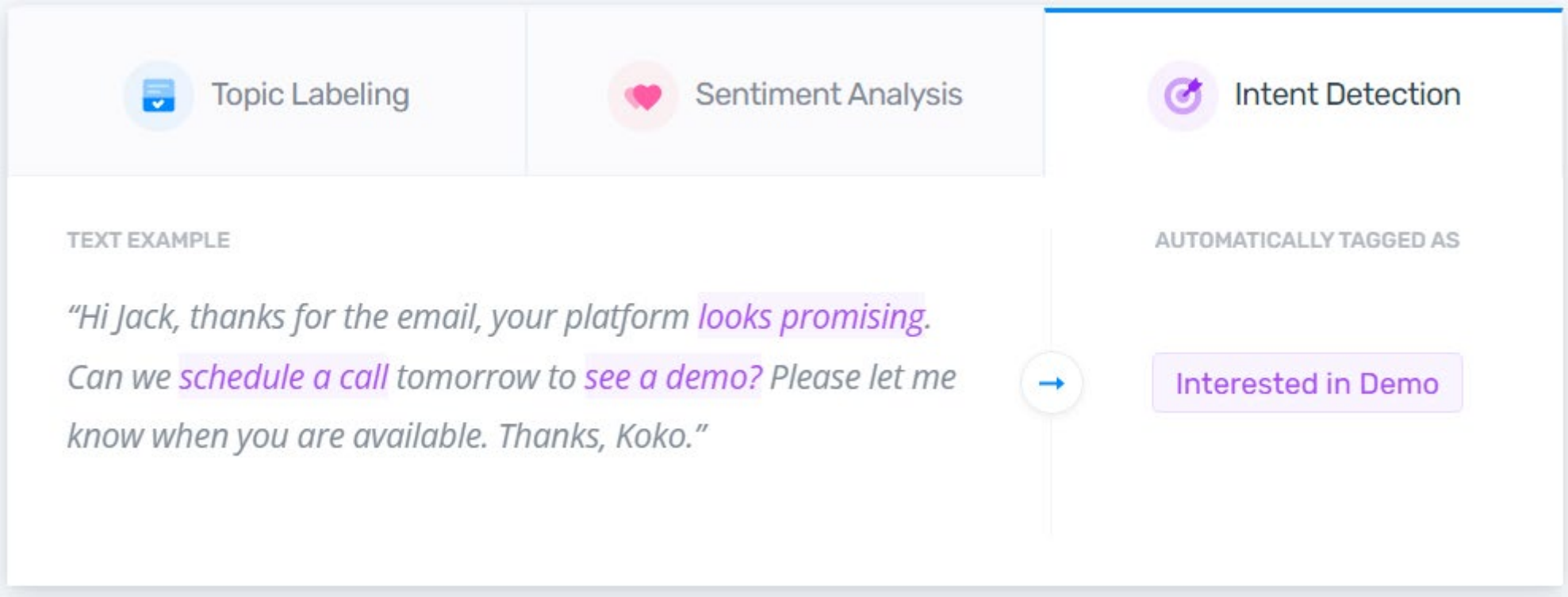
https://monkeylearn.com/text-classifiers/

Example of NLP Tools: MonkeyLearn



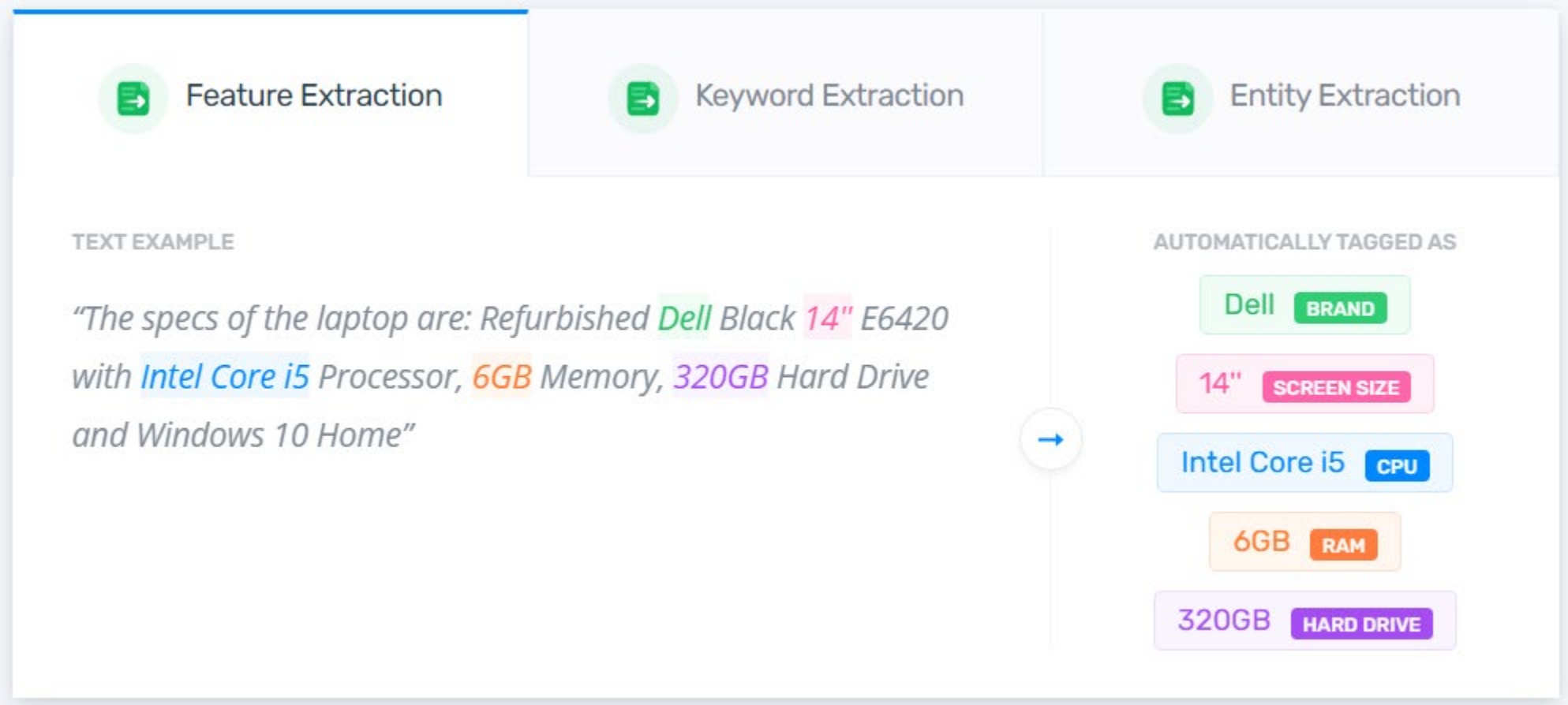
https://monkeylearn.com/text-classifiers/

Example of NLP Tools: MonkeyLearn



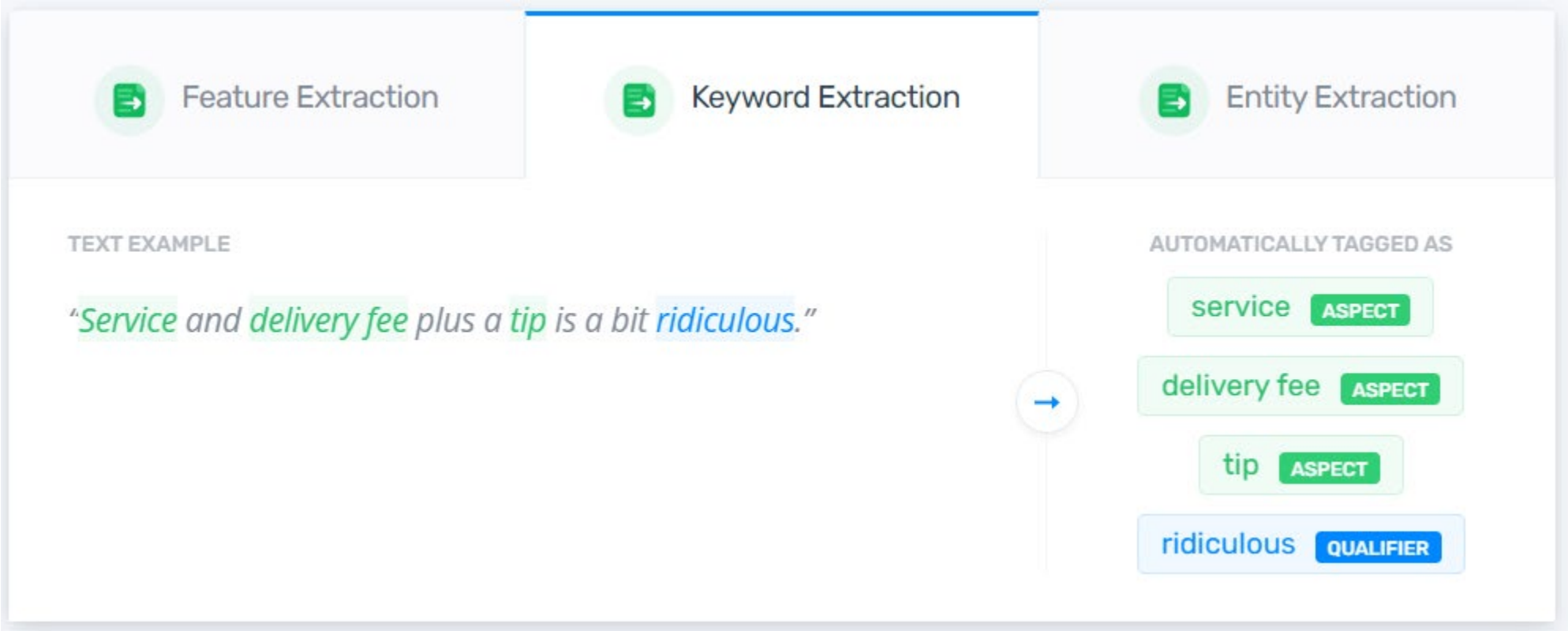
https://monkeylearn.com/text-classifiers/

Example of NLP Tools: MonkeyLearn



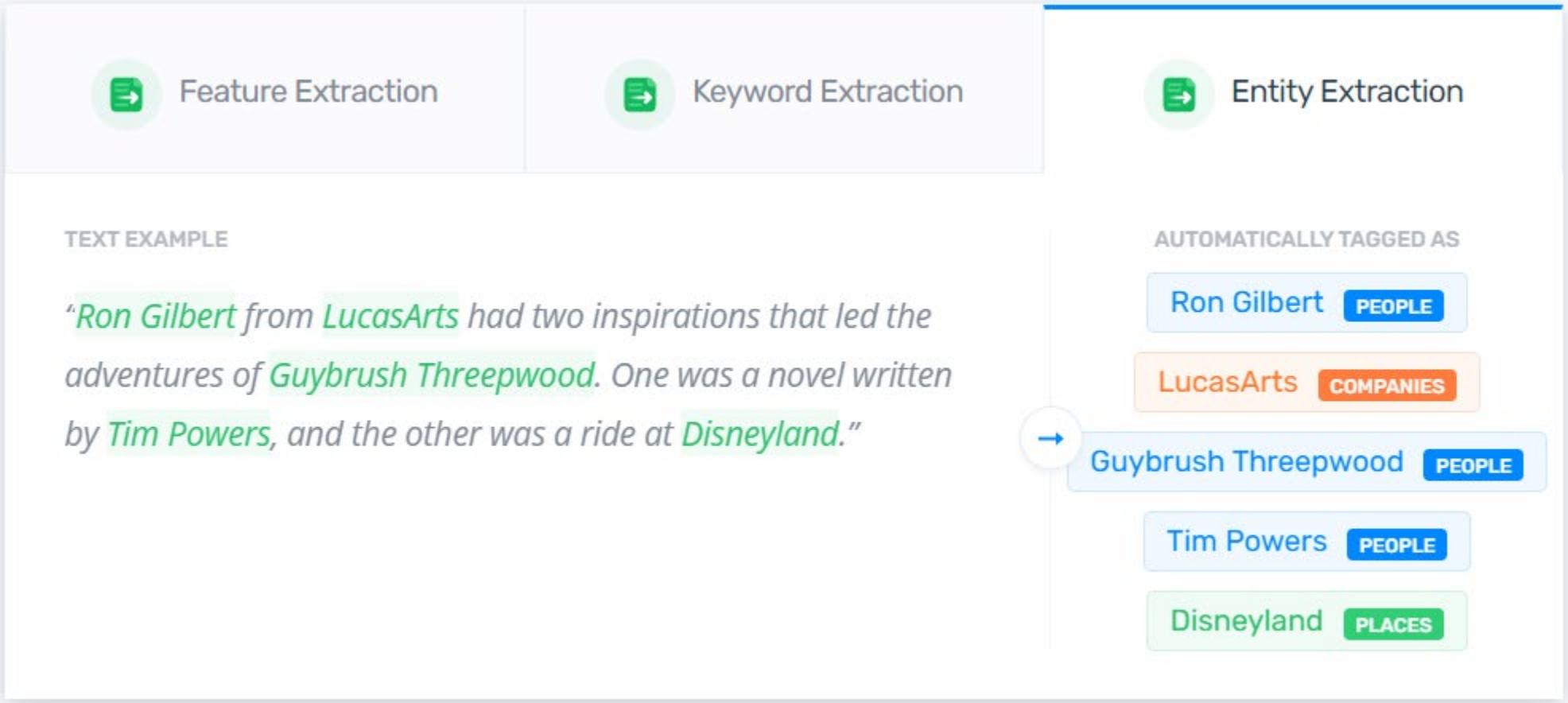
https://monkeylearn.com/text-extractors/

Example of NLP Tools: MonkeyLearn



https://monkeylearn.com/text-extractors/

Example of NLP Tools: MonkeyLearn



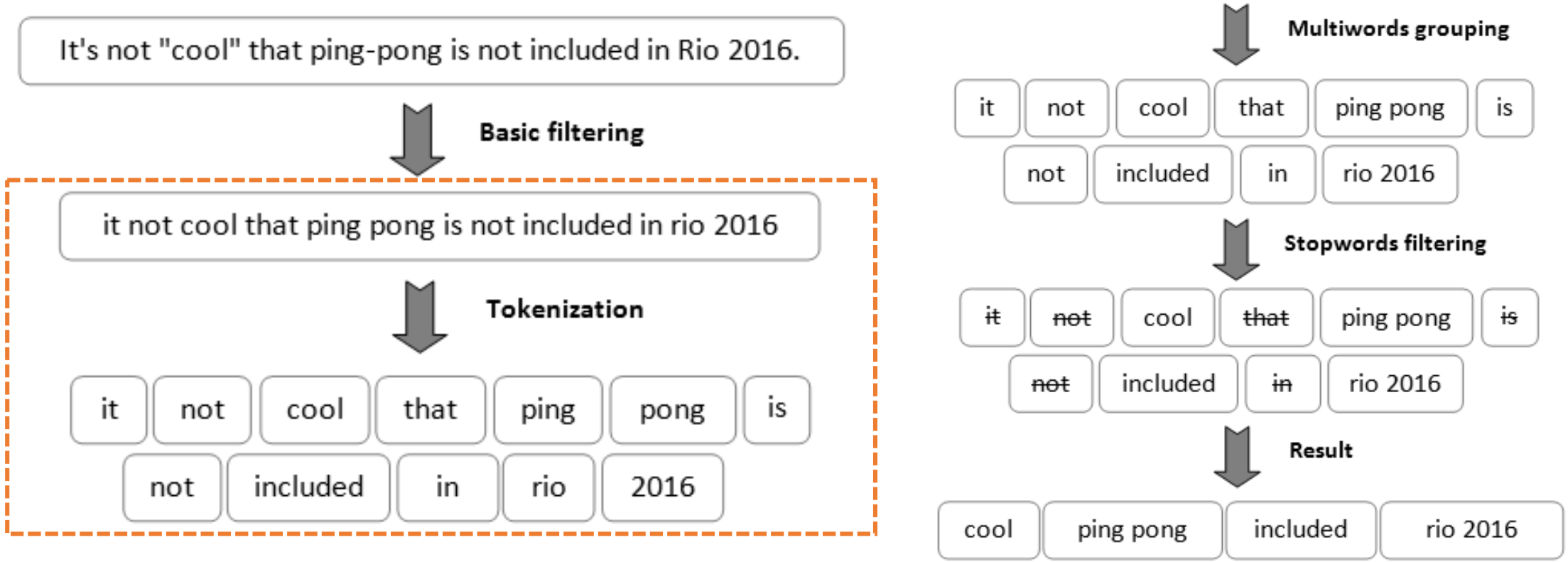
https://monkeylearn.com/text-extractors/

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.

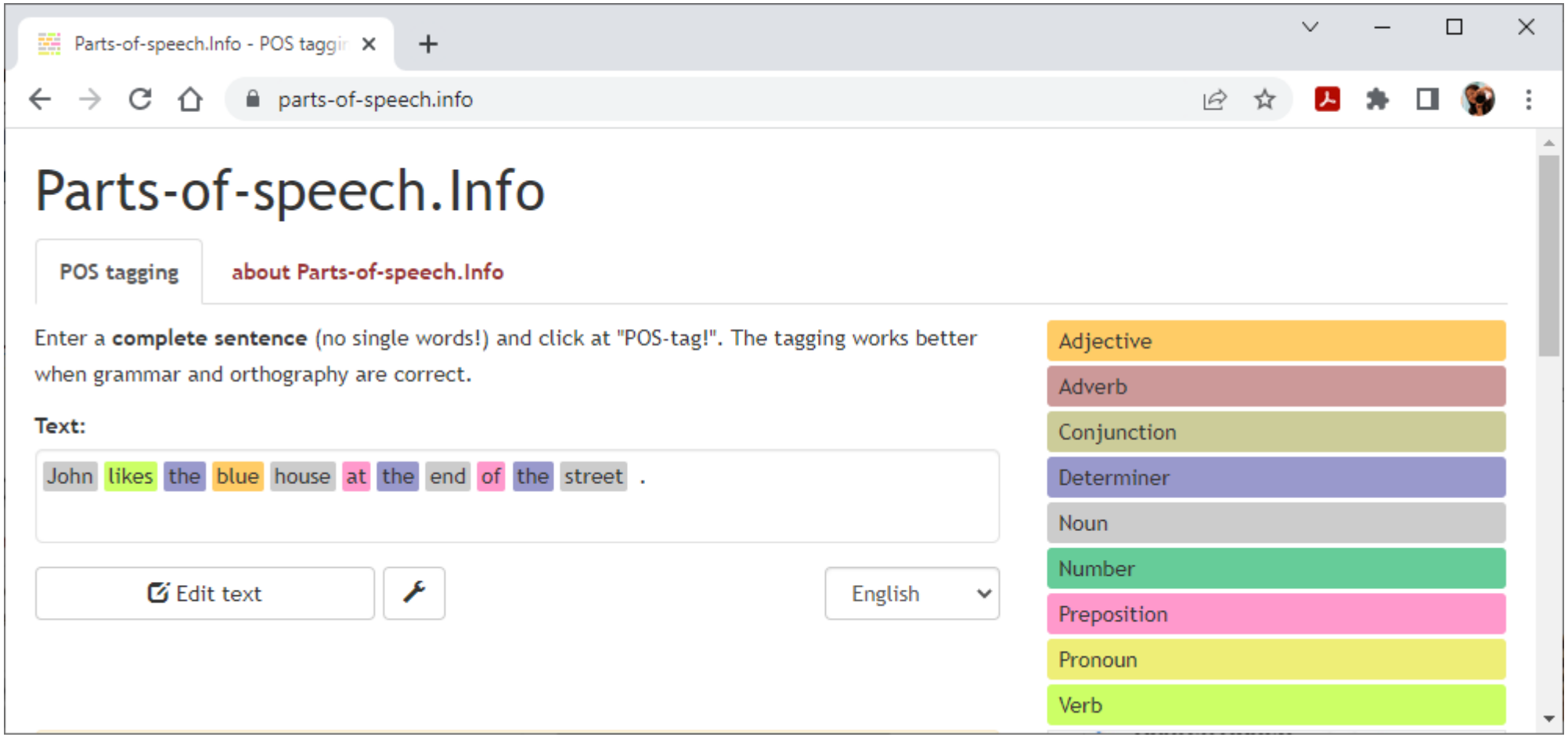
https://www.nltk.org/

Example of Text tokenization & multiword



https://www.meaningcloud.com/developer/resources/doc/models/1.0/models/text-tokenization-multiwords

Example of POS Tagging



https://parts-of-speech.info/

Example of Named Entity Recognition

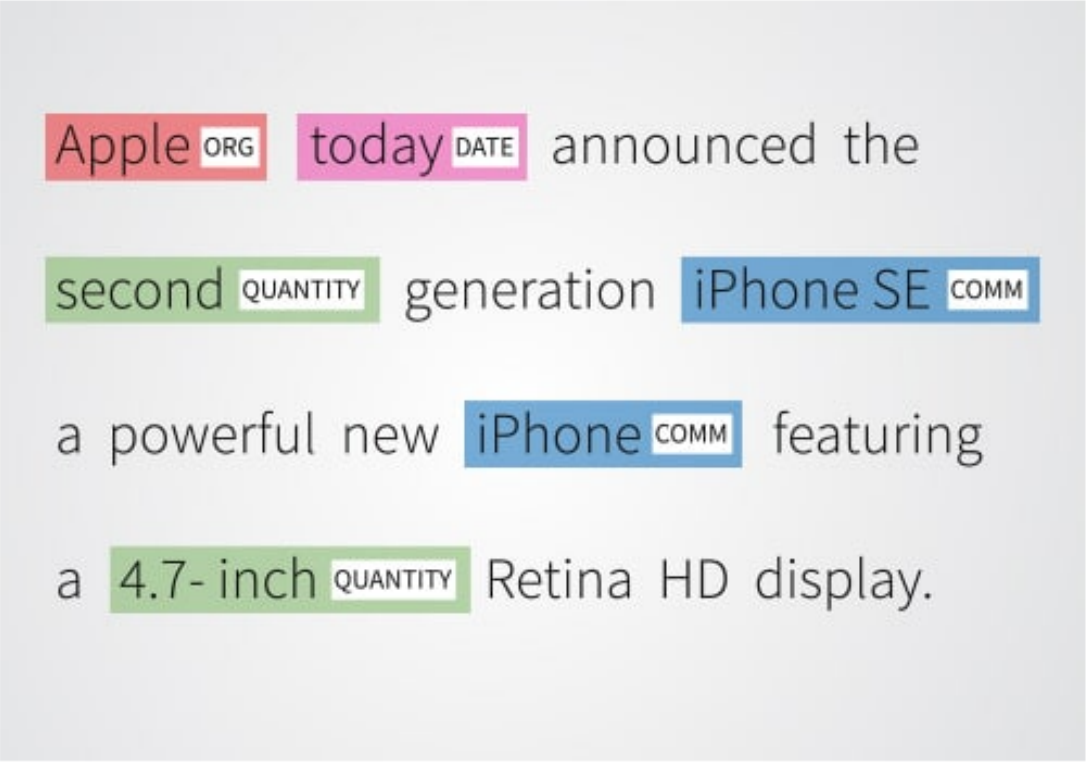
Person: Michael Jackson, Oprah Winfrey, Barack Obama, Susan Sarandon

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.



https://www.shaip.com/blog/named-entity-recognition-and-its-types/

Examples of NLP Technology

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