simplilearn | PURDUE UNIVERSITY

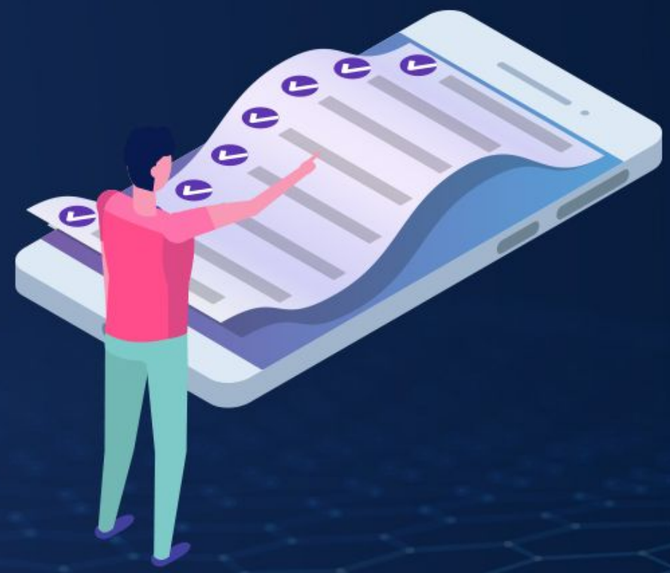
Natural Language Processing

Introduction to Natural Language Processing

Learning Objectives

By the end of this lesson, you will be able to:

- Describe natural language processing and its components
- Explain the different applications of NLP
- Define and demonstrate text processing



DATA AND
ARTIFICIAL INTELLIGENCE

Introduction to NLP

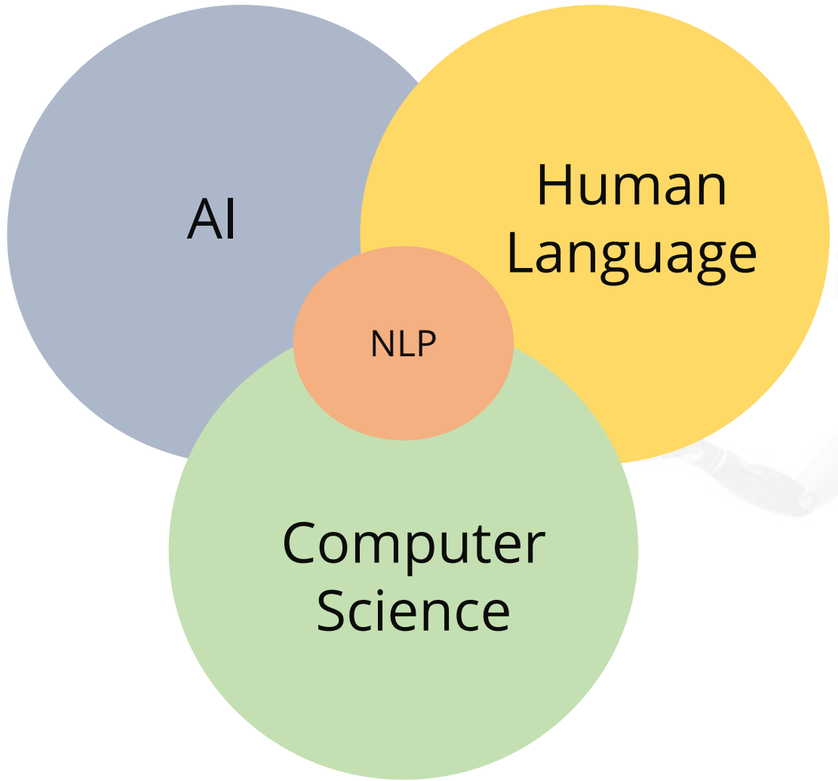
What Is NLP?

Natural Language Processing (NLP) is a branch of AI.

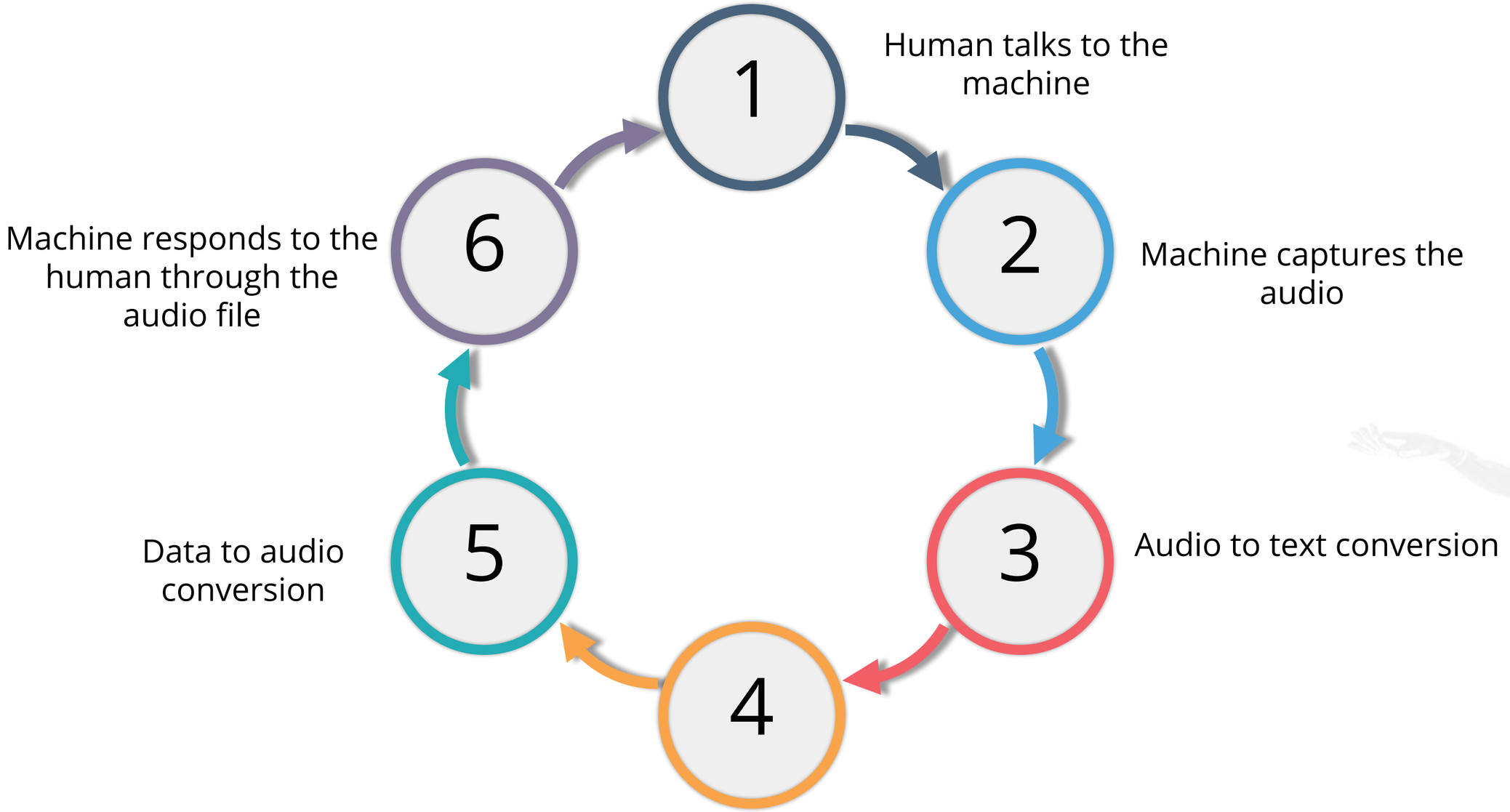
It helps machine to deal with human languages.

It helps machine to understand, interpret, and manipulate human languages.

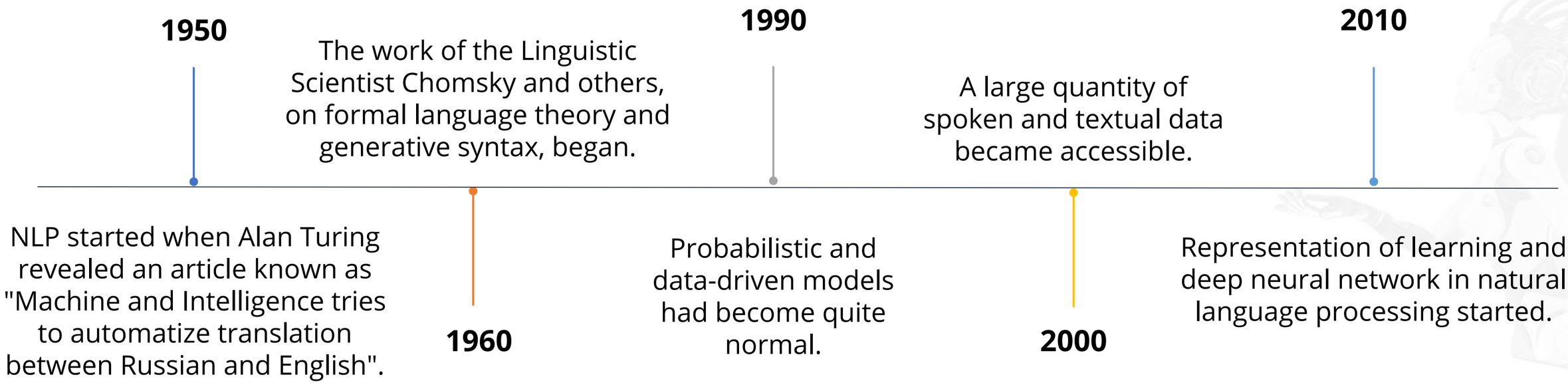
Most of the Natural Language Processing techniques depend on machine learning to derive meaning from human languages.



Interaction between Humans and Machines Using NLP



History of NLP



Human Language

To understand NLP, let us first understand the human language.

Alphabets

Word: Combination of Alphabets

Sentence: Combination of Words

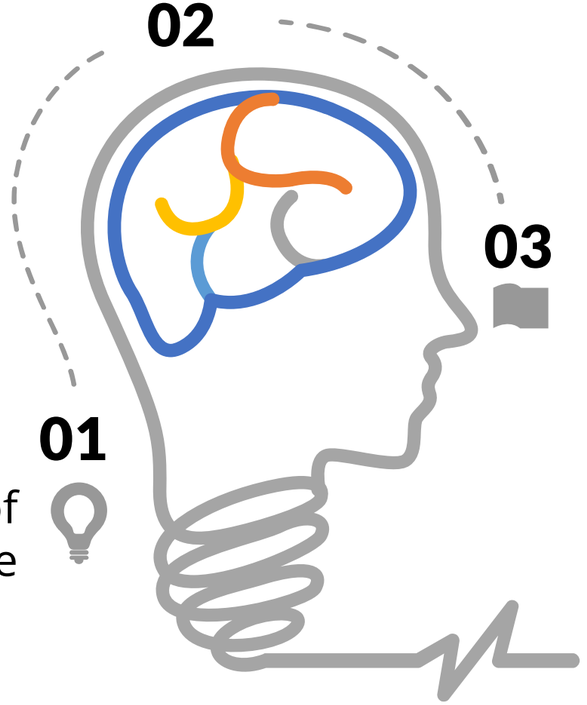
Apply Grammar Rules

Meaningful Sentence

Why NLP

02

Text is the largest repository of human knowledge and is growing quickly



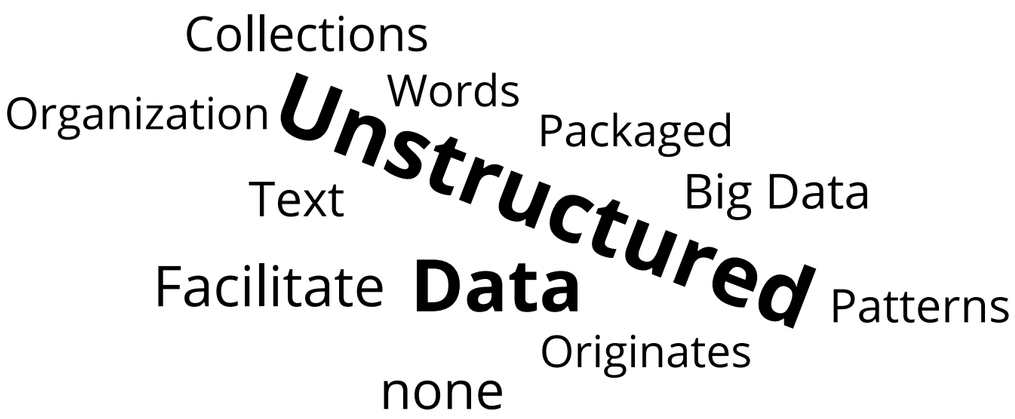
03

NLP represents computer programs that understand text or speech

01

NLP is the hallmark of human intelligence

Need for NLP



Data Sources

How to analyze this unstructured data?
Use Text Mining

Understanding Text Mining

It is also called text analysis.

Process of deriving insights from natural language text

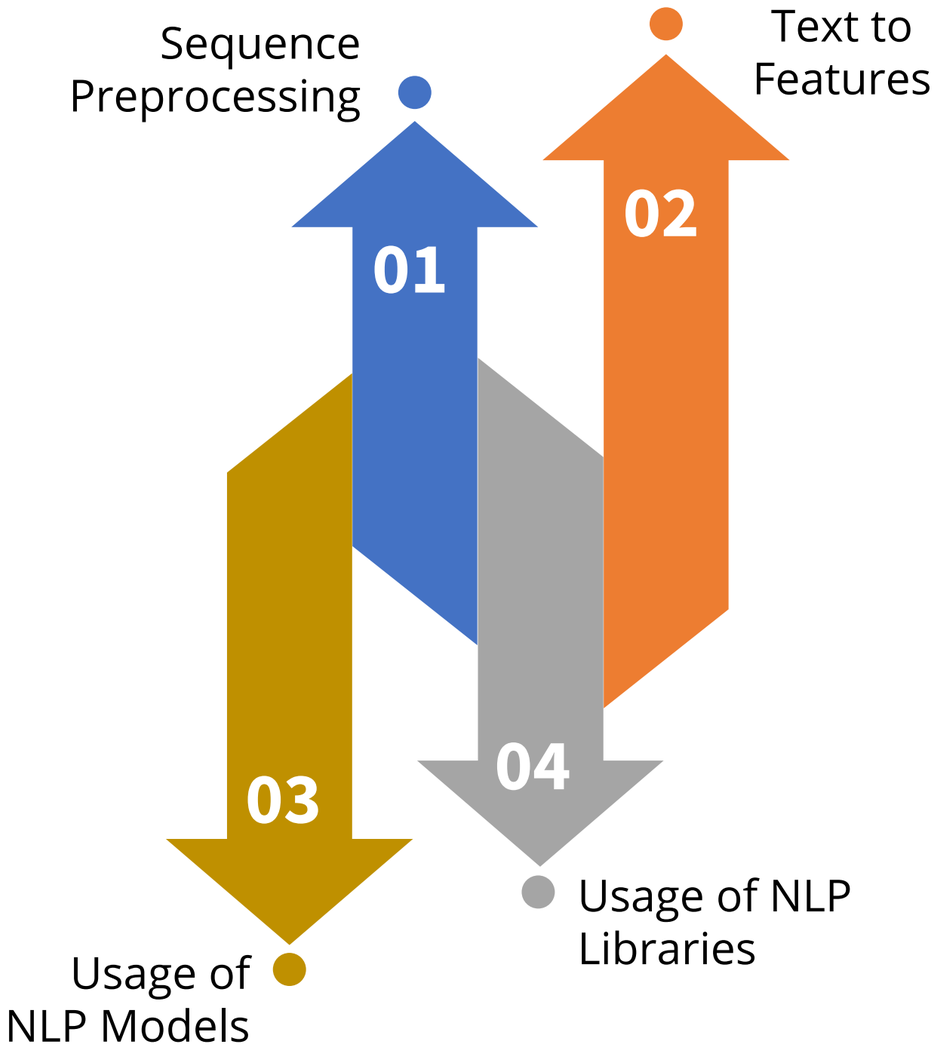
This is where NLP helps

Structure the input text

Derive pattern

Evaluate output

How NLP Works



Different Aspects of NLP

01	NLP is the ability of a computer to analyze, understand, and generate human languages.
02	A language is a system, a set of rules or set of symbols.
03	Symbols are combined and used for conveying information or broadcasting the information.
04	In NLP, rules of grammar are used for handling symbols.
05	NLP is a component of artificial intelligence.

Categories of NLP

Rule-Based NLP

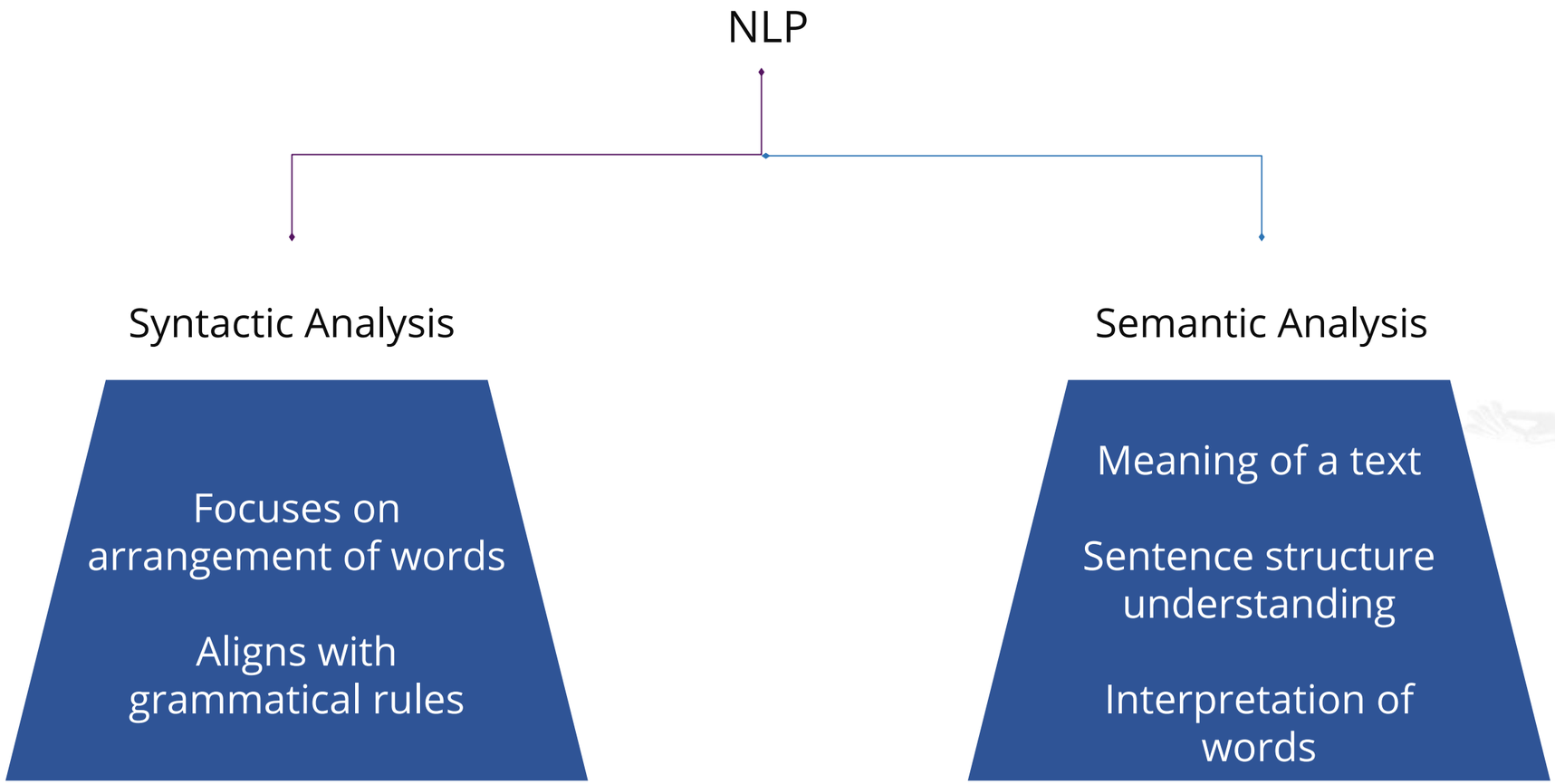
- Designed by creating a set of rules
- Developed by heuristic rules

Statistical Revolution

Statistical NLP

- Relies heavily on machine learning
- Applies automatic learning procedure

Techniques Used in NLP



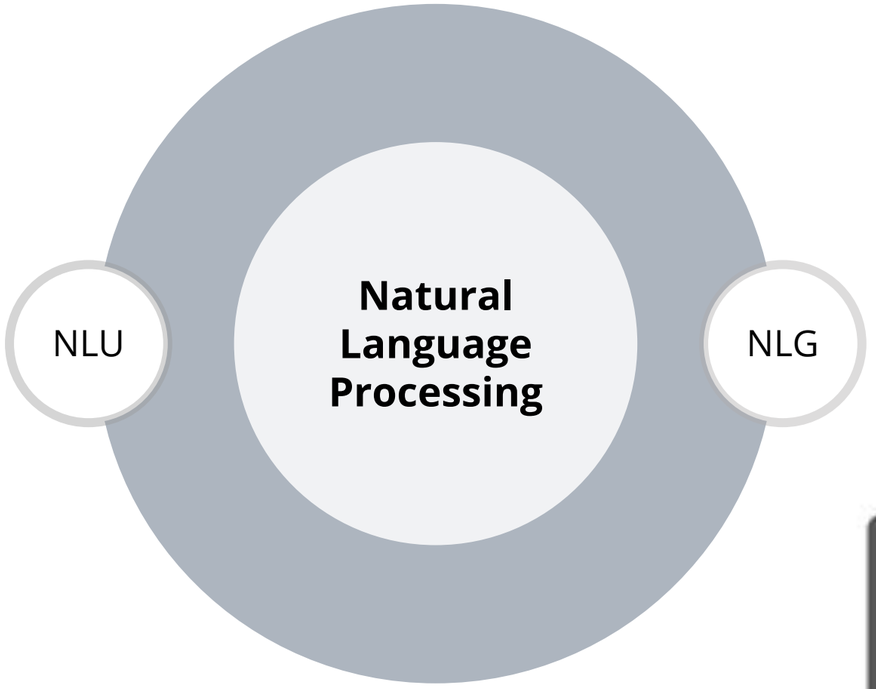
DATA AND
ARTIFICIAL INTELLIGENCE

Components of Natural Language Processing

Components of Natural Language Processing

1

Natural Language Understanding (NLU)



Natural Language Generation (NLG)

2

Components: Natural Language Understanding (NLU)

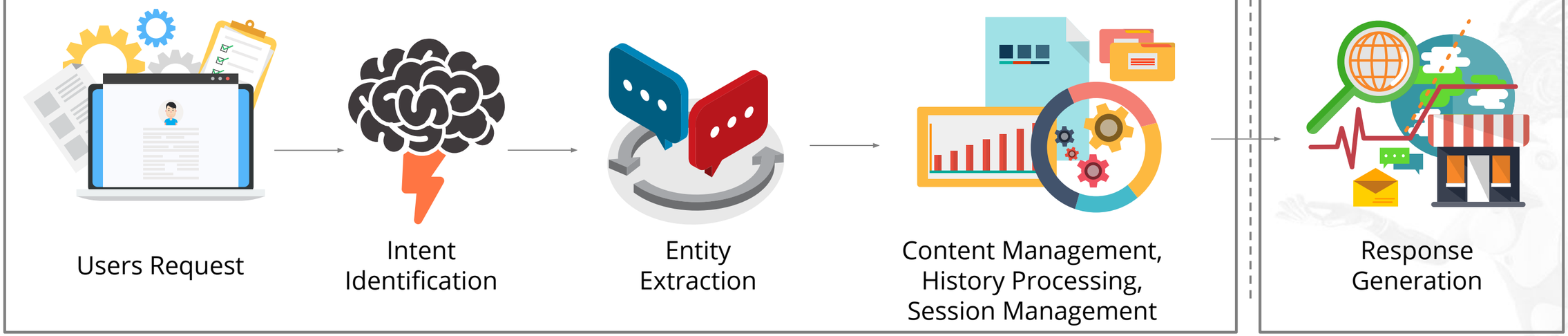
Taking some sentences and finding out what they mean

Components: Natural Language Generation (NLG)

1. Taking some formal representation of what you want to say and working out a way to express it in a natural language
2. Mapping the given input in the natural language with a useful representation
3. Producing output in the natural language from some internal representation
4. Different level of analysis: morphological analysis, syntactic analysis, semantic analysis, and discourse analysis

Uses of NLP

Use of NLP in conversational bot in each step:



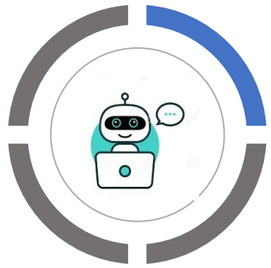
DATA AND
ARTIFICIAL INTELLIGENCE

Applications of Natural Language Processing

NLP in Real-Life

Speech Recognition

Machine Translation



Chatbot

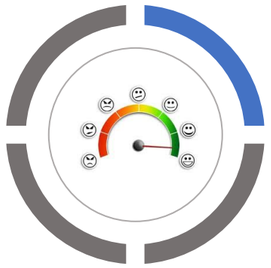
Information Retrieval

NLP in Real Life

Information Extraction

Question Answering

Spell Check



Sentiment Analysis

NLP in Real-Life: Business Usage

Improve user experience

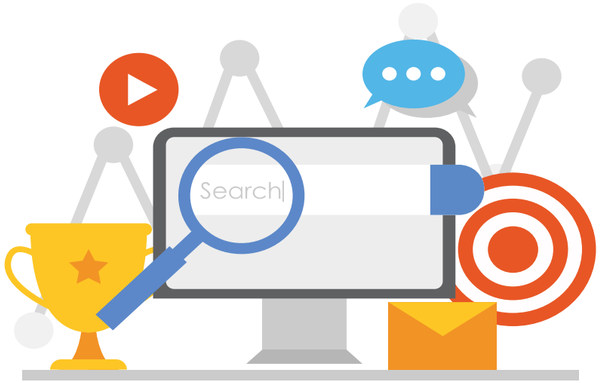
- Spellcheck
- Autocomplete
- Autocorrect

Automate support

- Chatbot
- Product ordering

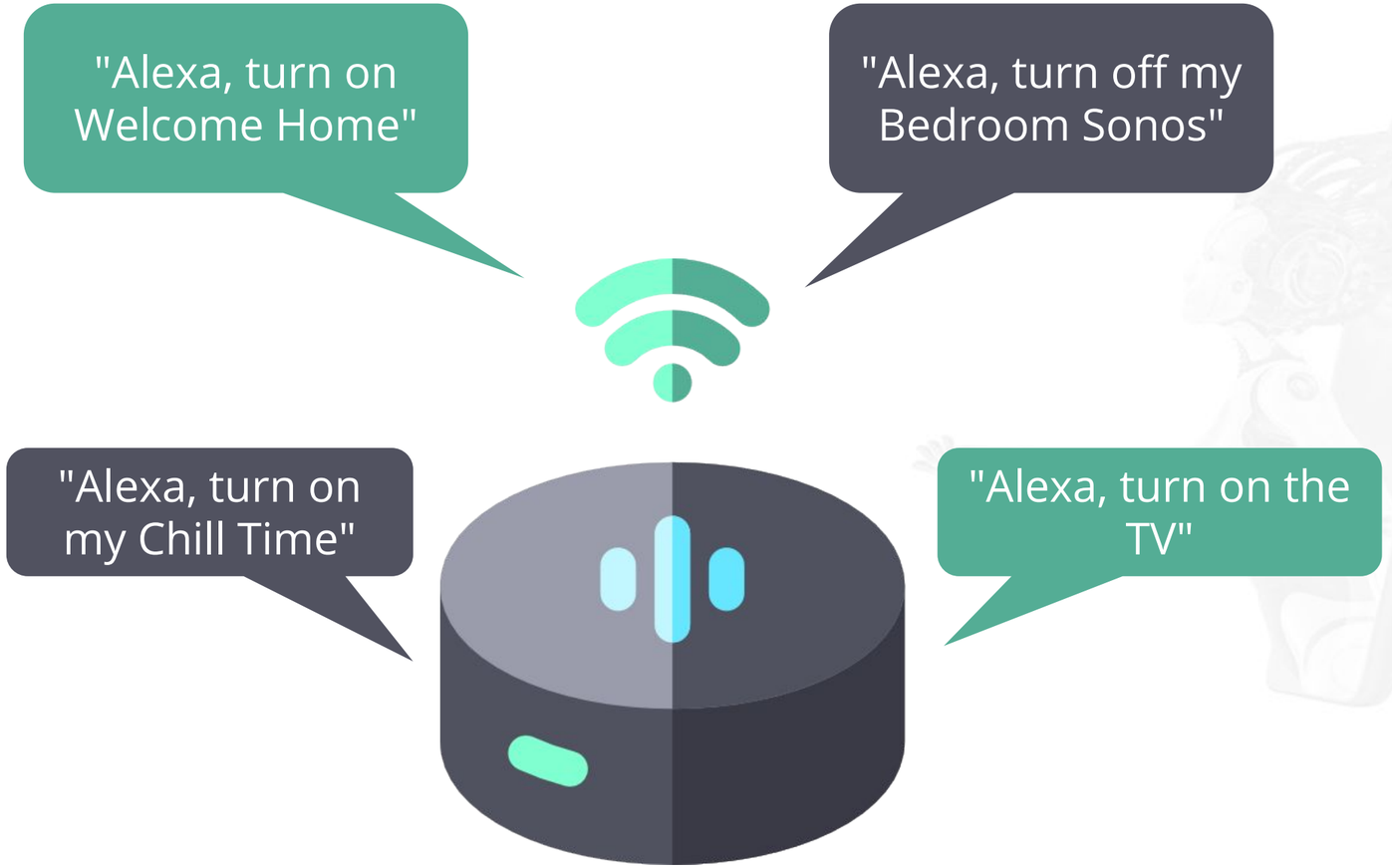
Monitor and analyze feedback

- Generate actionable insight from huge amount of review or feedback

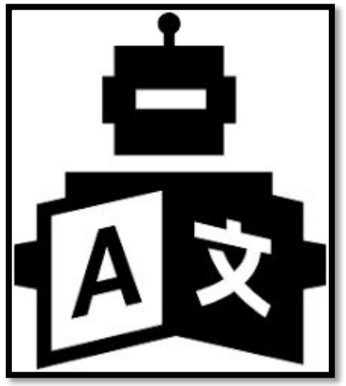


NLP in Real-Life: Speech Recognition

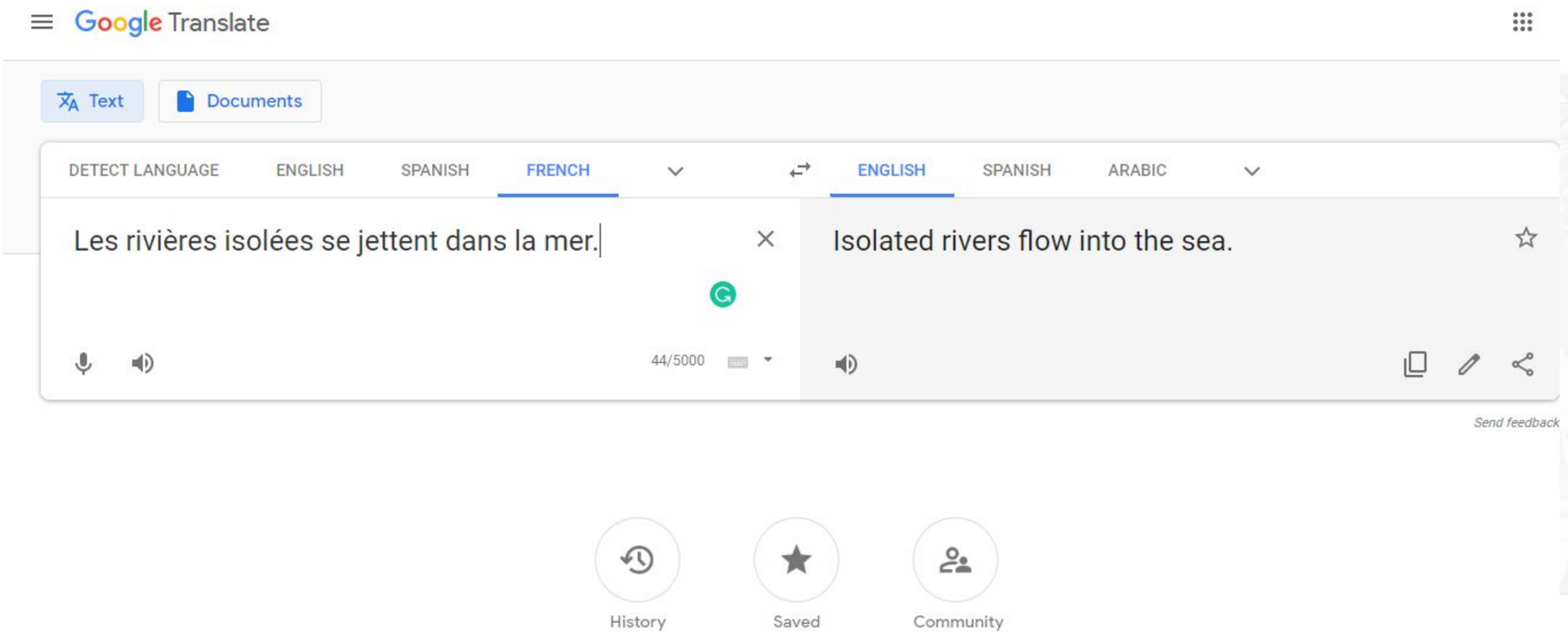
- Google Assistant
- Siri
- Alexa
- Cortana



NLP in Real-Life: Machine Translation

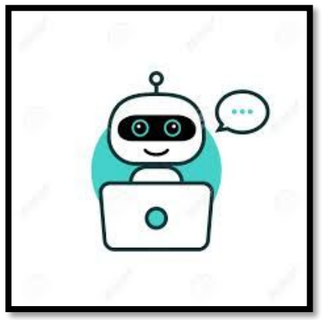


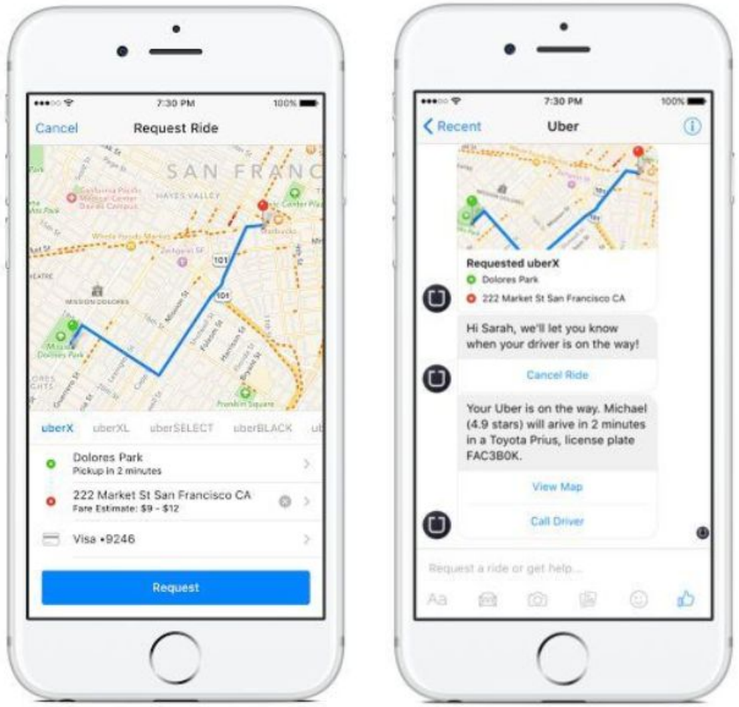
Google translator

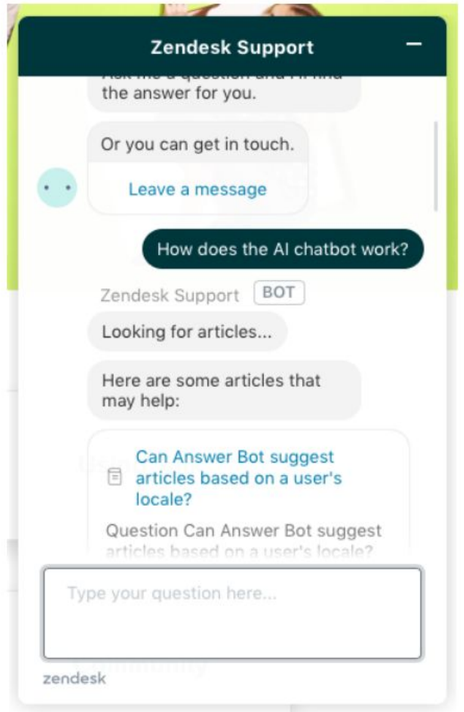


NLP in Real-Life: Chatbots

Uber, Facebook Messenger, and Zendesk are some of the companies who have implemented chatbots using NLP.







Source

NLP in Real-Life: Information Retrieval

Find information according to the given query

Collections Audio Video Words Organization Packaged Involve Sentences Text Data Big Data Facilitate Unstructured Patterns none Originates

NLP techniques used in IR are:

- Stemming
- Part-of-Speech Tagging
- Compound Recognition
- Decompounding
- Chunking
- Word-Sense Disambiguation

Google finds relevant and similar results using Information Retrieval.

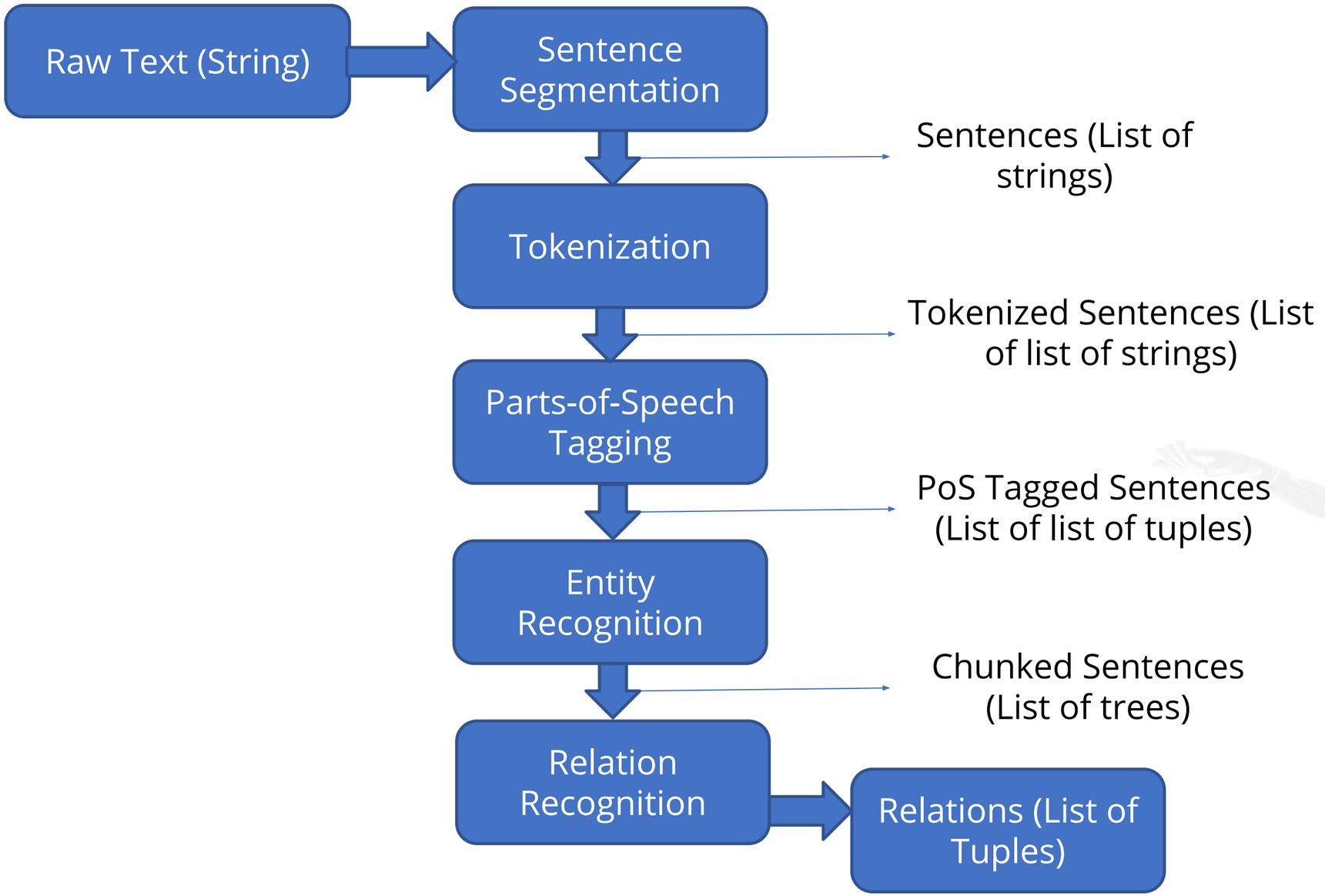
NLP in Real-Life: Information Extraction

Automatic extraction of structured information from unstructured or semi-structured machine-readable documents

Gmail structures events from e-mails

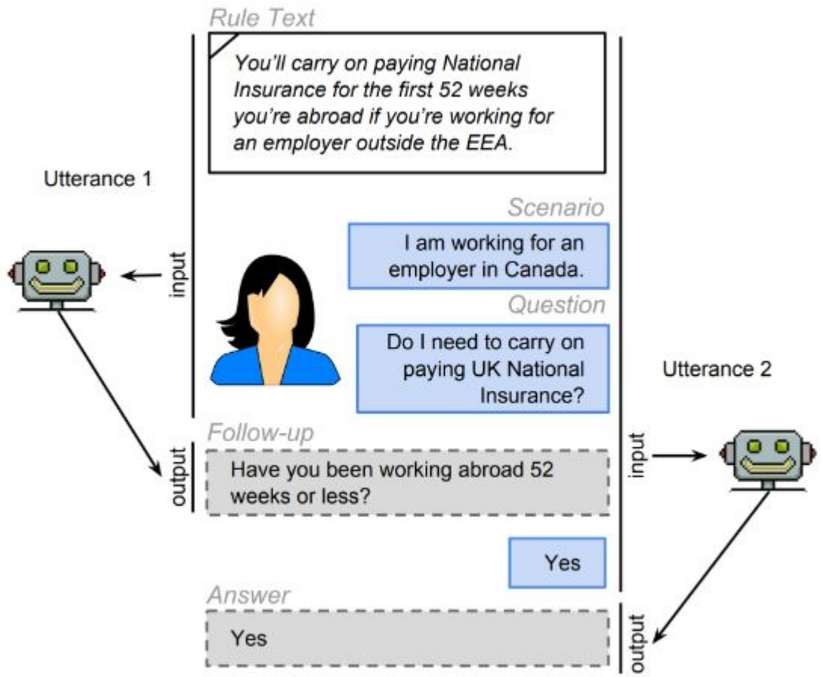
Source

NLP in Real-Life: Information Extraction



NLP in Real-Life: Question Answering

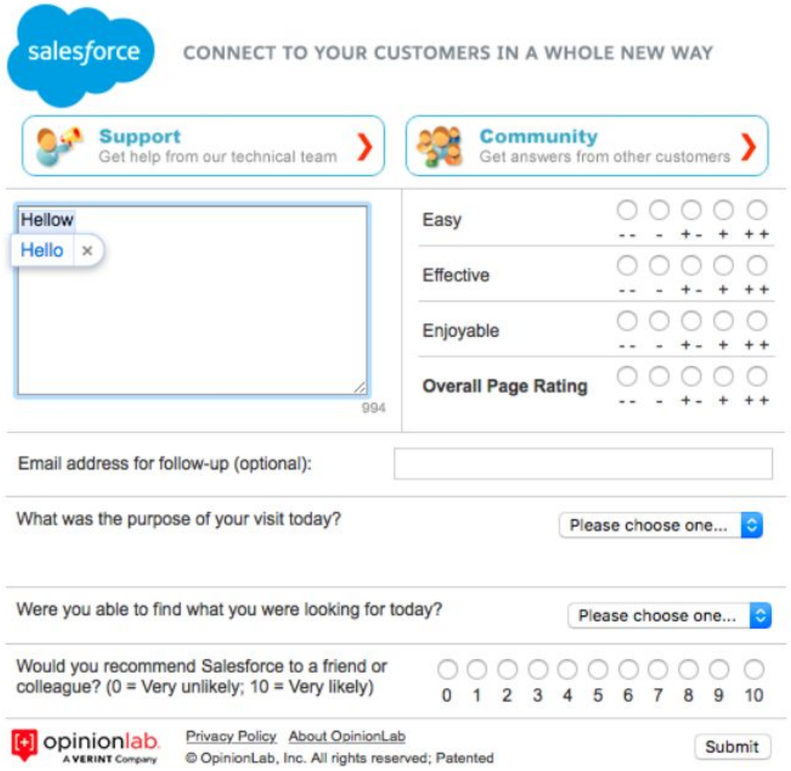
System that automatically answers questions



Source

NLP in Real-Life: Spell Check

Salesforce implemented spell check in the contact forms using NLP.



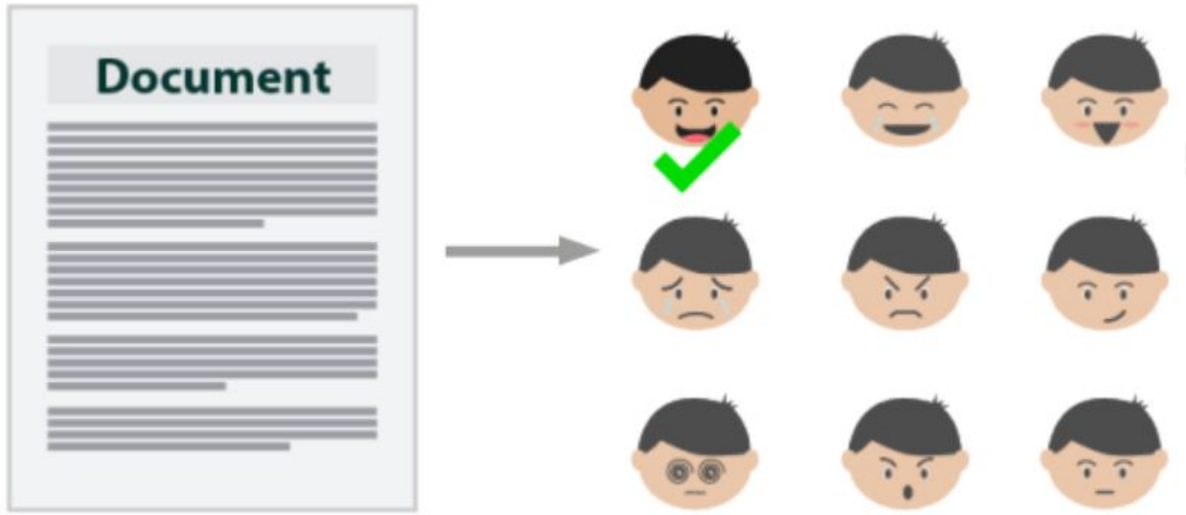
Source

Grammarly: A grammar checking SW built on NLP

NLP in Real-Life: Sentiment Analysis

To extract subjective information from a piece of text
Example: Whether an author is being subjective or objective or even positive or negative

NLP is used here



Source

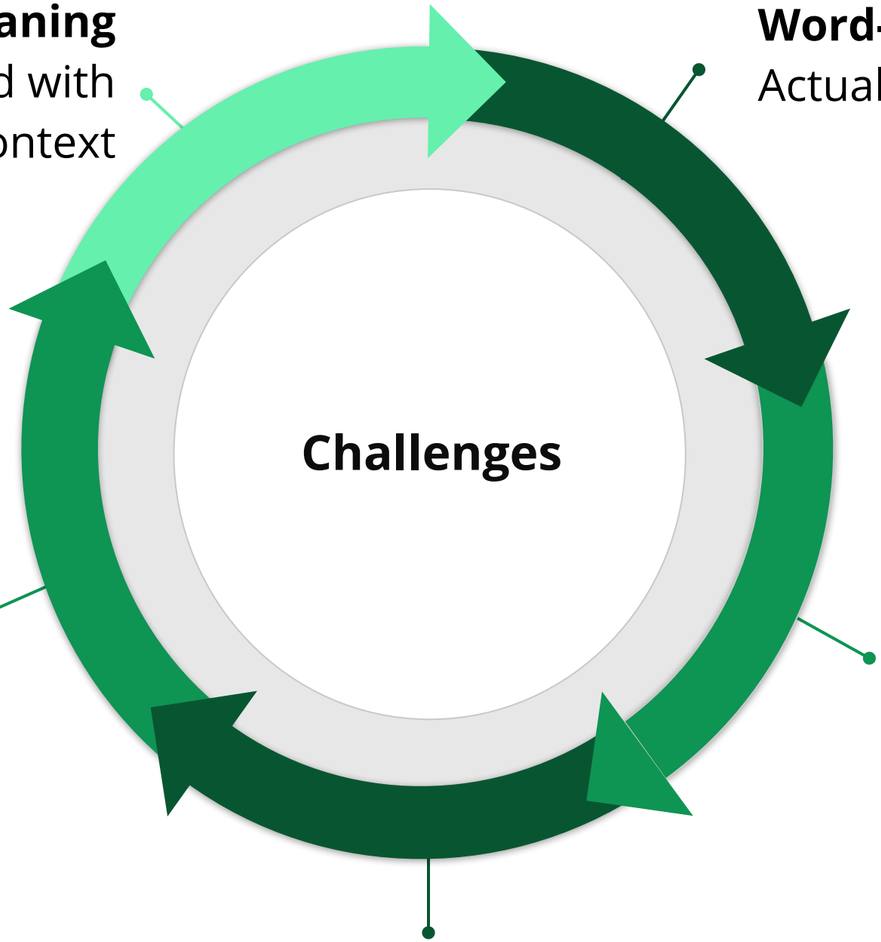
DATA AND
ARTIFICIAL INTELLIGENCE

Challenges and Scope

Why NLP Is Difficult

Nature of the human language	Rules that dictate the passing of information using natural languages are not easy for computers to understand.
Human language is unstructured data	Only 21% of the data is structured data, and a lot of information in the world is unstructured.
Tough to extract meaning from text	Process of reading and understanding English is very complex.

Challenges and Scope



Semantic Meaning
Understanding a word with respect to its context

Word-Sense Disambiguation
Actual context of the text

Entity Extraction
Extracting the unknown entities

Multiple Intents
User speaking many things in single text

Anaphora Resolution
Absence of important entity in text conversation

Challenges and Scope: Semantic Meaning

There are many good properties available on HDFC Red portal.

I want to purchase a red carpet from a store.

Word RED has different meanings in these contexts.

Challenges and Scope: Understanding Entities

A 2M solution of $CaCl_2$ consists of 221.82g of $CaCl_2$ dissolved in enough water to make one liter of solution.

Understanding and extraction of $CaCl_2$ as entity in this context is complex.

Challenges and Scope: Anaphora Resolution

Peter and Greg are NLP developers.
He is living in Pune.

The word “He” used in 2nd sentence does not specify which person to refer.

Challenges and Scope: Multiple Intents

My bank account is functional. Please provide me resolution process and I want to buy Laptop from Flipkart.

The word “He” used in 2nd sentence does not specify which person to refer.

NLU Challenges: Ambiguity

Lexical Ambiguity

More than 1 meaning of a word in a sentence

Example: The fisherman went to the **bank**.

Syntactic or Grammatical Ambiguity

More than 1 meaning of a sentence

Example : **Visiting relatives** can be boring.

Referential Ambiguity

Reference of a pronoun

The boy told the father about the theft. **He** was very upset.

DATA AND
ARTIFICIAL INTELLIGENCE

Data Formats

Data Formats

To apply NLP on data, we need to have the data which is available on different kinds of sources in different formats.

Below are the types of data formats:

Structured

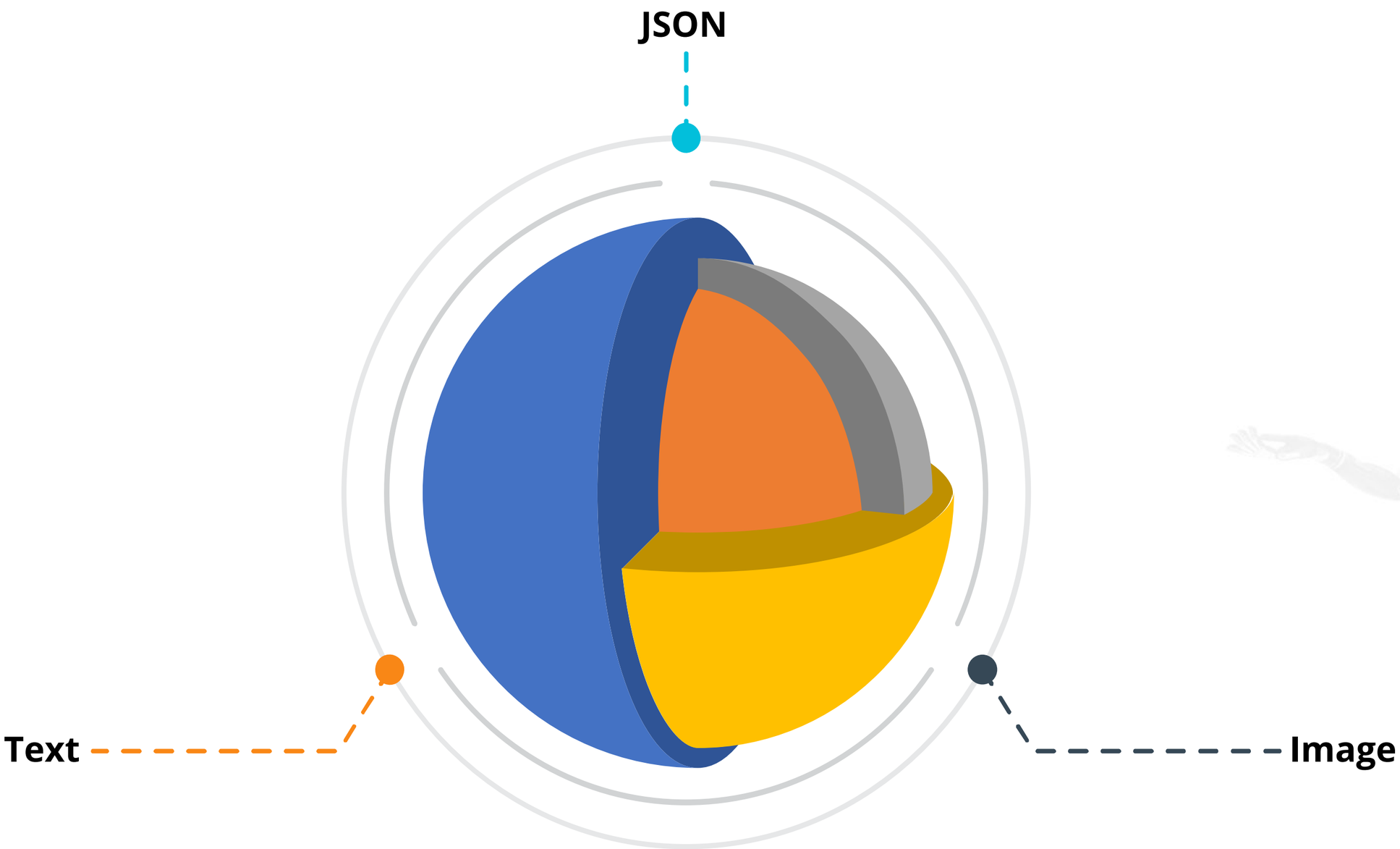
Unstructured

Semi-Structured

Data Formats: Structured

	SQL Data	

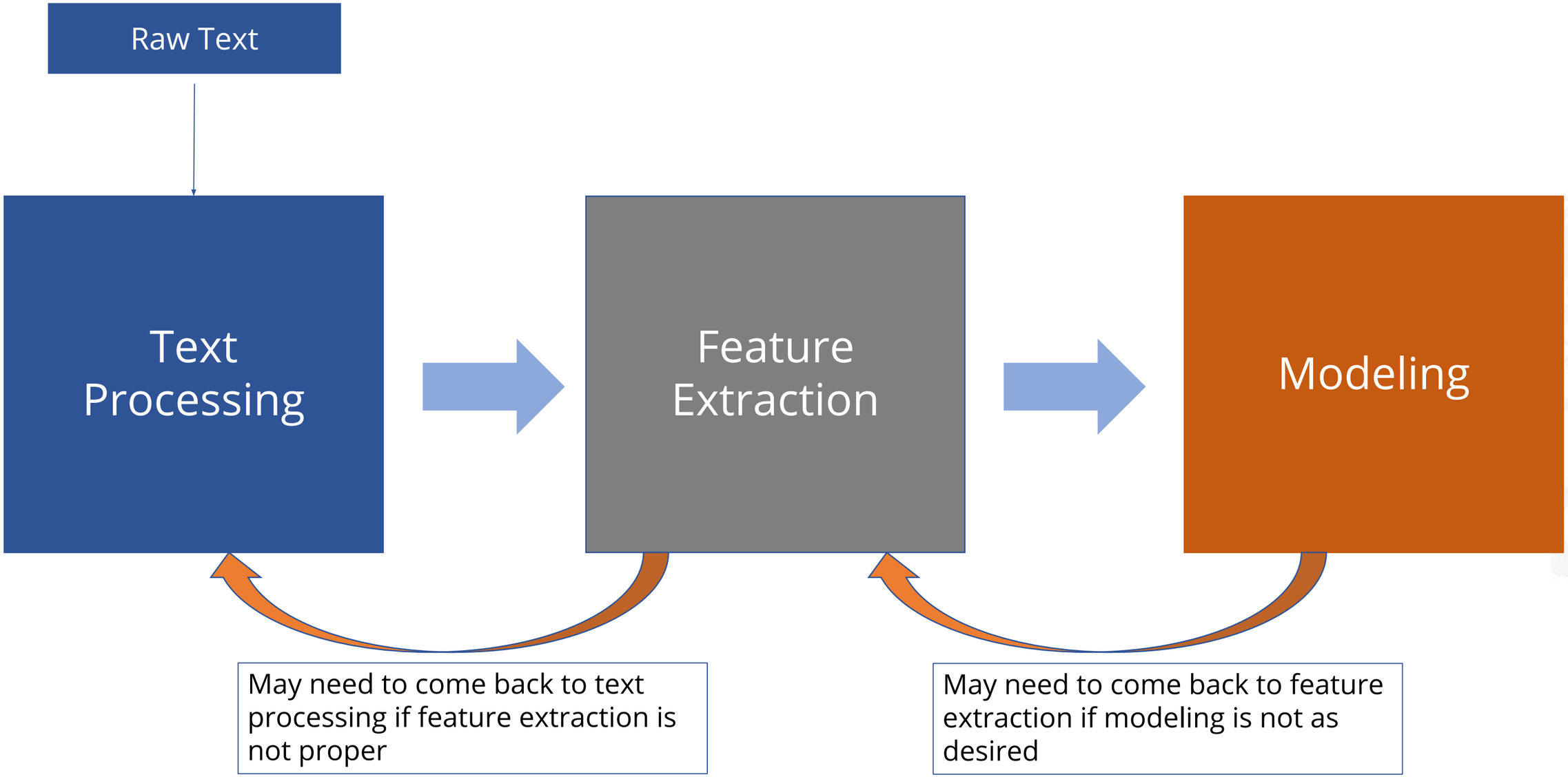
Data Formats: Unstructured and Semi-Structured



DATA AND
ARTIFICIAL INTELLIGENCE

NLP Pipeline

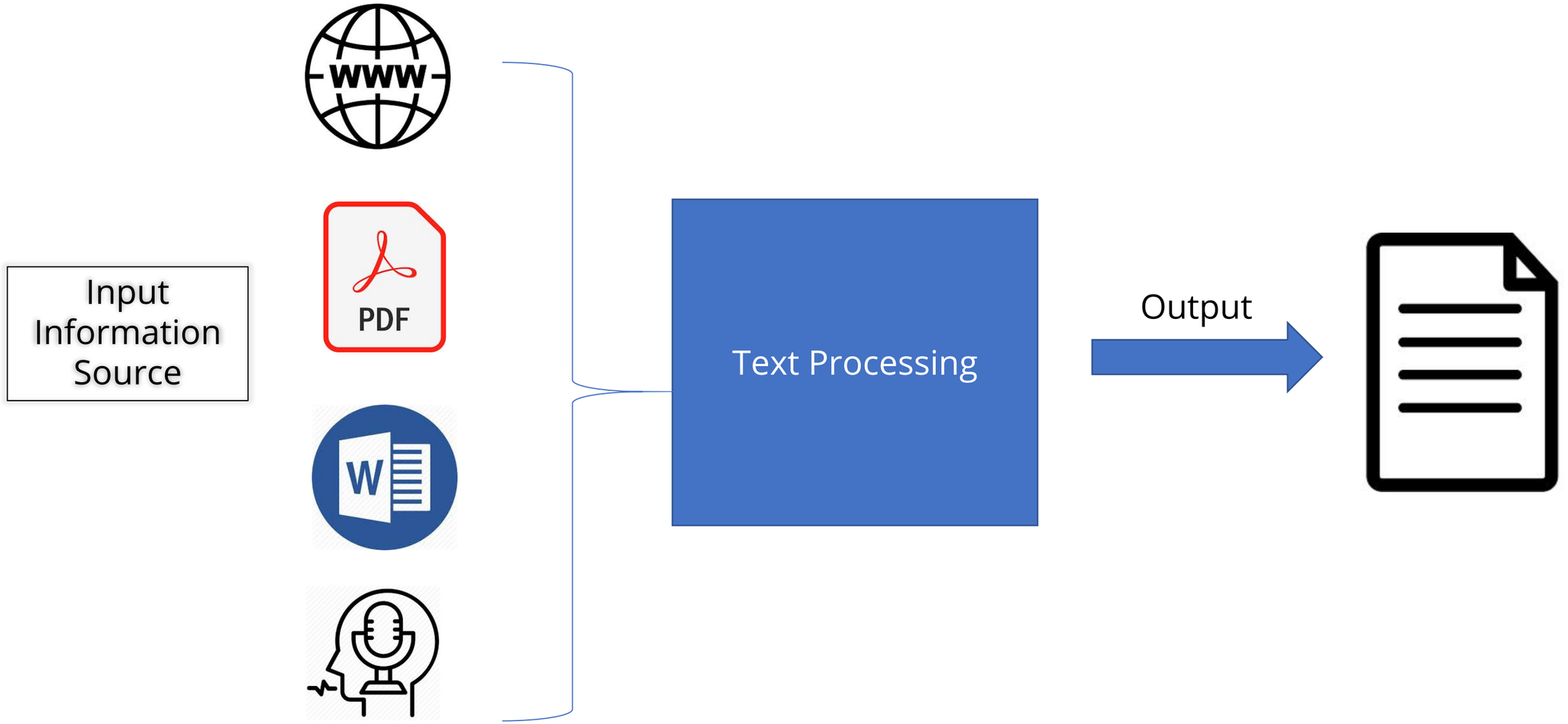
NLP Pipeline



DATA AND
ARTIFICIAL INTELLIGENCE

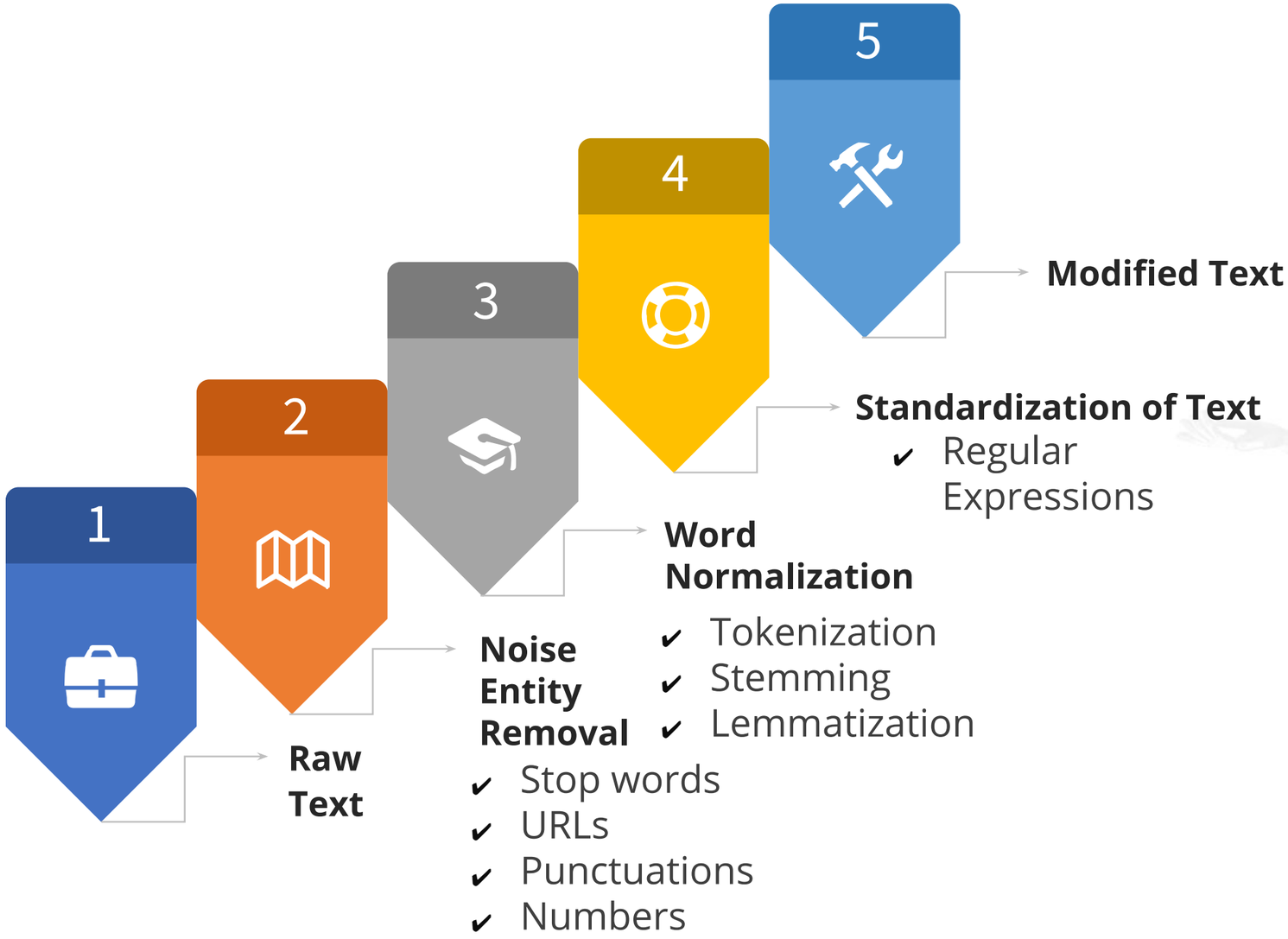
Text Processing

Text Processing

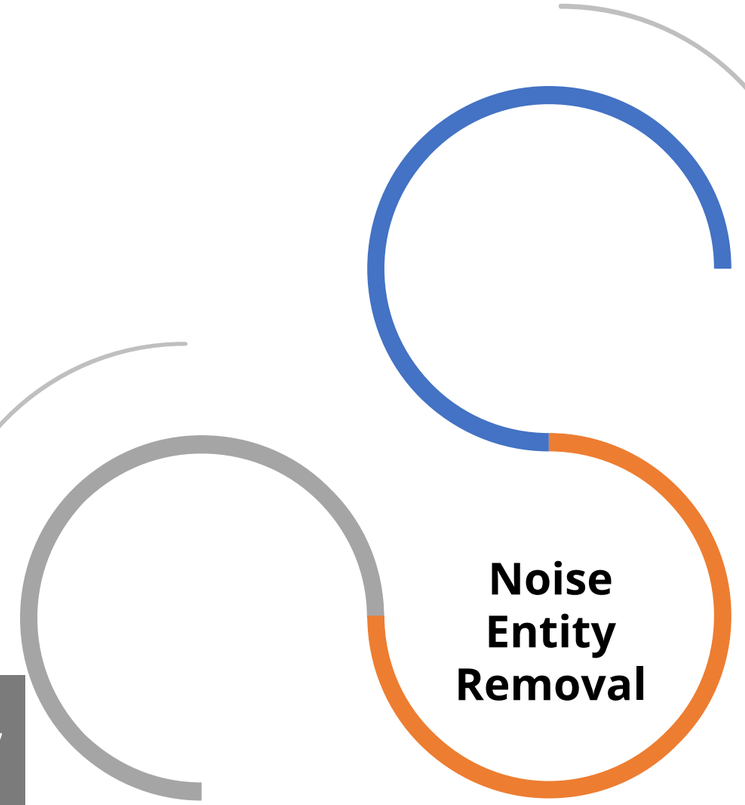


Sequence or Text Processing

Sequence or Text Processing has the following steps:-



Sequence or Text Processing: Noise Entity Removal



Convert all the words of document into lowers case

For efficient processing, we must keep all words in same casing to avoid case sensitivity of text.

Eliminating unnecessary punctuations, stop words, numbers, and URLs

These types of things do not contribute for better result. They only increase the size of texts and decrease the efficiency of algorithms.

Sequence or Text Processing: Tokenization

Tokenization

- Break the sentence into separate words.
- These words are called tokens.
- Split words whenever there is a space between them.
- Treat punctuation marks as separate tokens since punctuation also has meaning.

Example:

Sentence	Word
London is the capital and the most populous city of England and the United Kingdom	“London”, “is”, “ the”, “capital”, “and”, “ the”, “most”, “populous”, “city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”

Sequence or Text Processing: Stemming

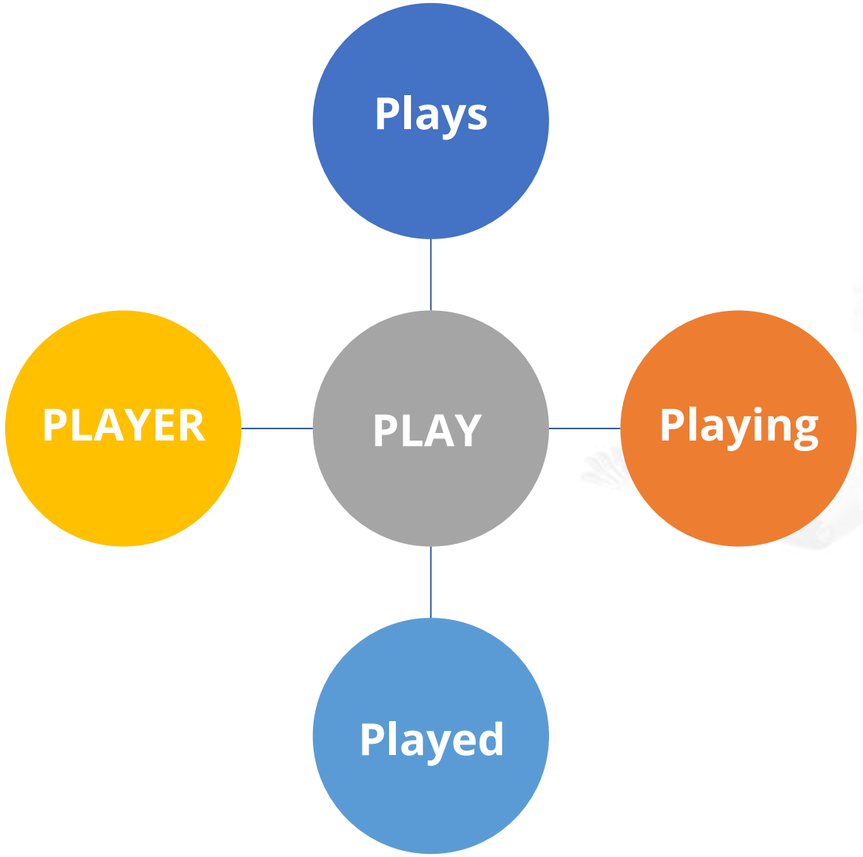
Stemming:

It takes the root of the words.

It removes the last few words or suffix of a word where it misspelt or incorrect words.

Example:

Word	Suffix	Stem
studies	-es	studi
ninez	-ez	nin



Sequence or Text Processing: Lemmatization

Lemmatization:

It converts the text to meaningful base form by considering its context.

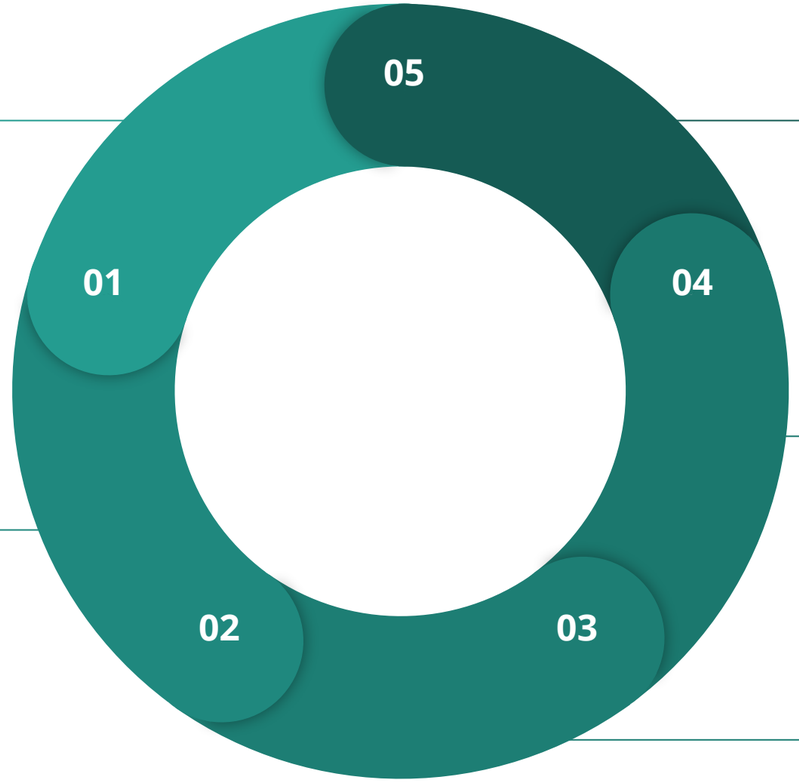
Example:

Word	Morphological Information	Lemma
Studying	Gerund of the word study	Study
Ninez	Singular number of nine	Ninez

Sequence or Text Processing: Chunking and Chinking

It is simplest technique used for entity detection

Chinking a way to remove a chunk from chunk

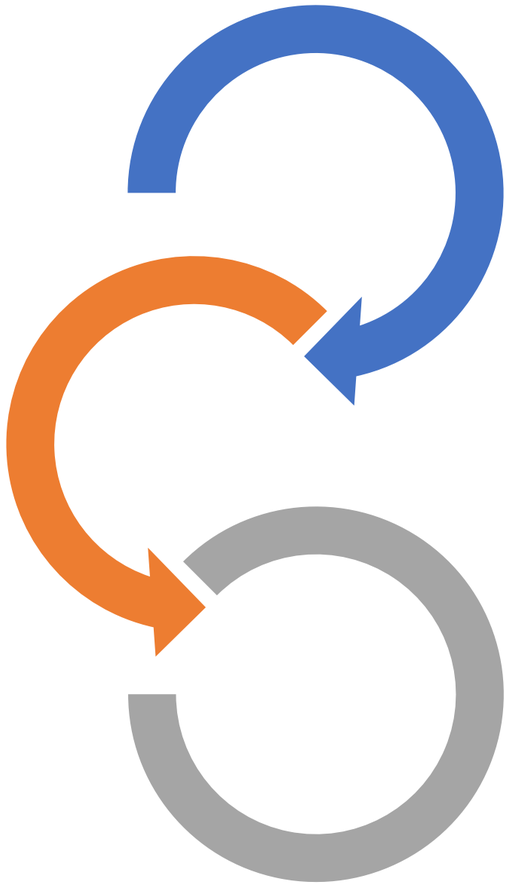


It is the process of extracting meaningful short phrases from sentences by analyzing the parts of speech

Words or patterns can also be defined. These should not be a part of chunk and such words are known as chinks

Chunk pattern are made by normal regular expression which are designed and modified to match the part of speech tags

Sequence or Text Processing: Regular Expression



Object Standardization:

- Some words or symbols which are not present in standard dictionary are also not recognized by any search processes.
- Examples: hashtags, acronyms, and colloquial slangs

Note: With the help of regular expression, we can remove these things.

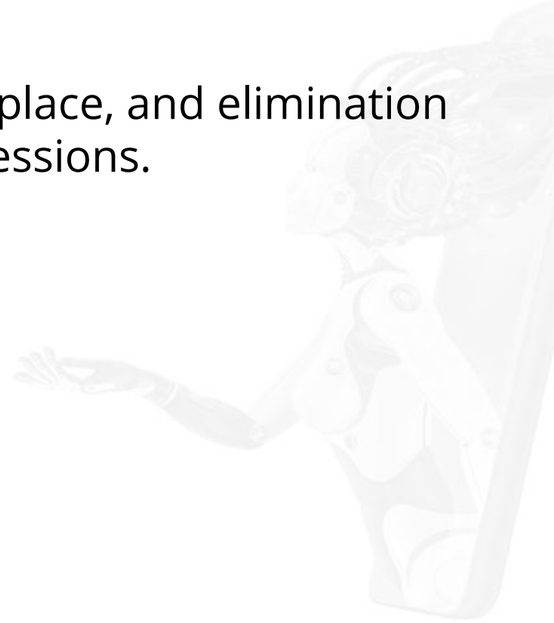
Sequence or Text Processing: Regular Expression

Regular Expression (Regex):

It is a sequence of characters that define pattern-matching, search-and-replace, and elimination functions. All type of noises can be removed with the help of regular expressions.

Regex Examples:

Expression	Description
[abc]	Find any character between the brackets
[^abc]	Find any character that is not between the brackets
[0-9]	Find any character between the brackets (digit)



NLTK

NLTK: Introduction

1. This tool is used for manipulation or understanding text or speech by any software or machine.
2. This is one of the most usable and mother of all NLP libraries.
3. It is a platform used for building Python programs that work with human language data for application in statistical Natural Language Processing (NLP).

NLTK: Introduction

Following are text processing libraries:

- Tokenization
- Lemmatization
- Parsing
- Classification
- Stemming
- Tagging
- Semantic Reasoning

NLTK: Syntax and library

System Requirement:

Operating System:
macOS / OS X · Linux · Windows (Cygwin, MinGW, Visual Studio)

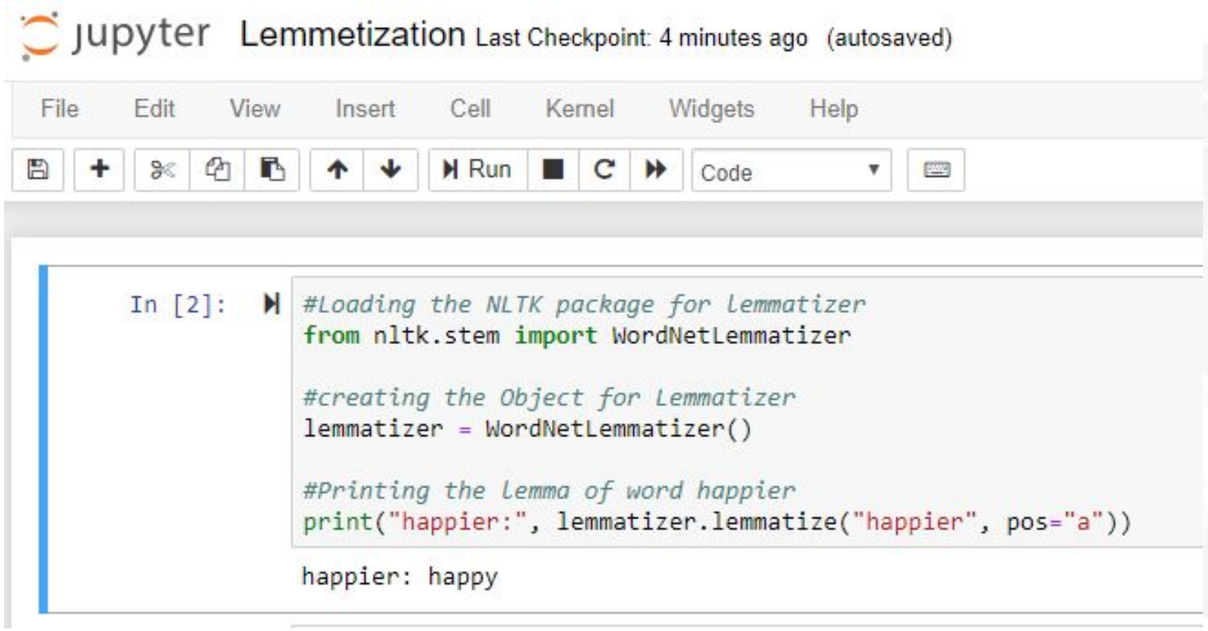
Python Version:
Python 2.7, 3.5+ (only 64 bit)

```
>> import nltk
```

NLTK: Lemmatization

For grammatical purpose, documents are going to use different forms of a word, for example:

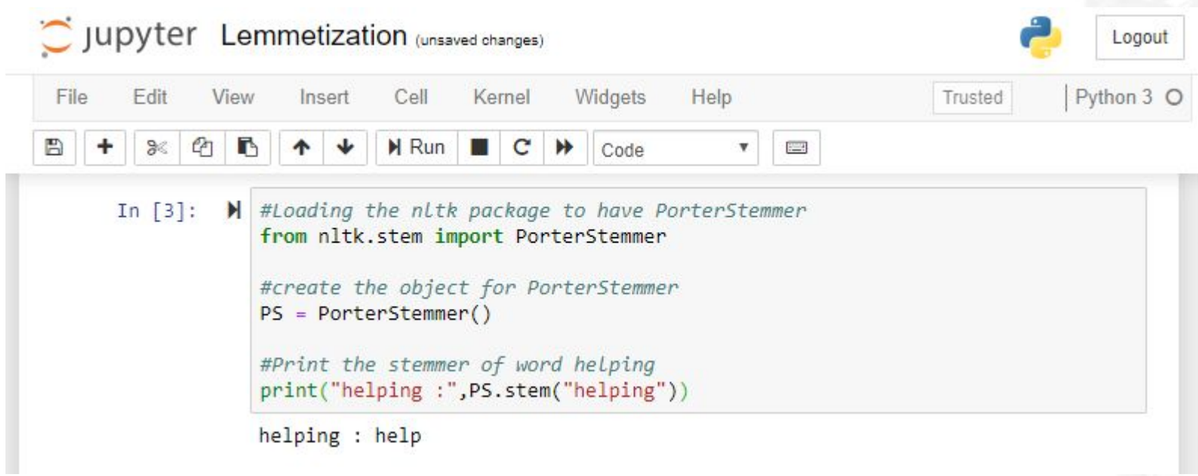
```
#Loading the NLTK package for
lemmatizer
from nltk.stem import WordNetLemmatizer

#creating the Object for Lemmatizer
lemmatizer = WordNetLemmatizer()

#Printing the lemma of word happier
print("happier:",
lemmatizer.lemmatize("happier",
pos="a"))
```



Output: happier: happy

NLTK: Stemming

```
#Loading the nltk package to have
PorterStemmer
from nltk.stem import PorterStemmer

#create the object for PorterStemmer
PS = PorterStemmer()

#Print the stemmer of word helping
print("helping :",PS.stem("helping"))
```



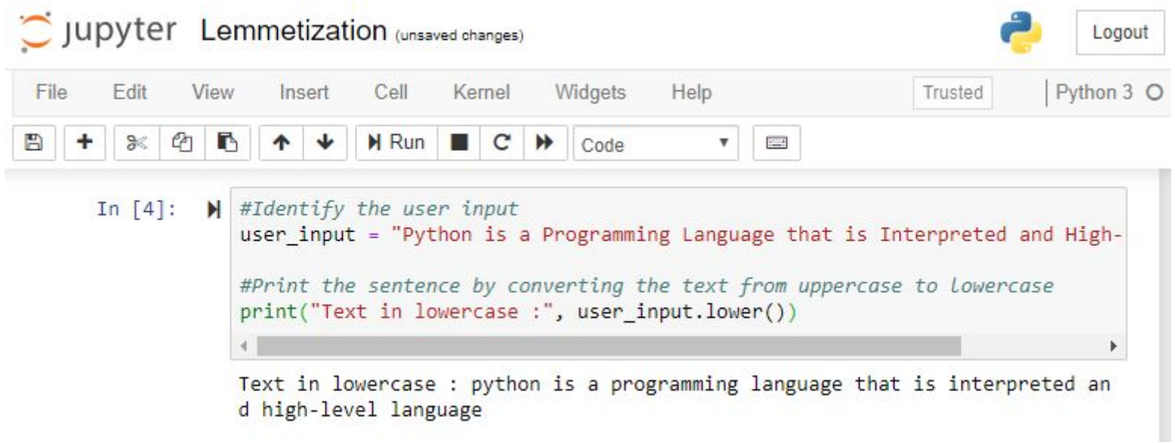
Output: helping: help

NLTK: Processing Raw Text

Text processing includes:
Converting all letters to lower or upper case

```
#Identify the user input
user_input = "Python is a Programming Language that is Interpreted and
High-Level language"

#Print the sentence by converting the text from uppercase to lowercase
print("Text in lowercase :", user_input.lower())
```

Output: Text in lowercase : python is a programming language that is interpreted and high-level language.



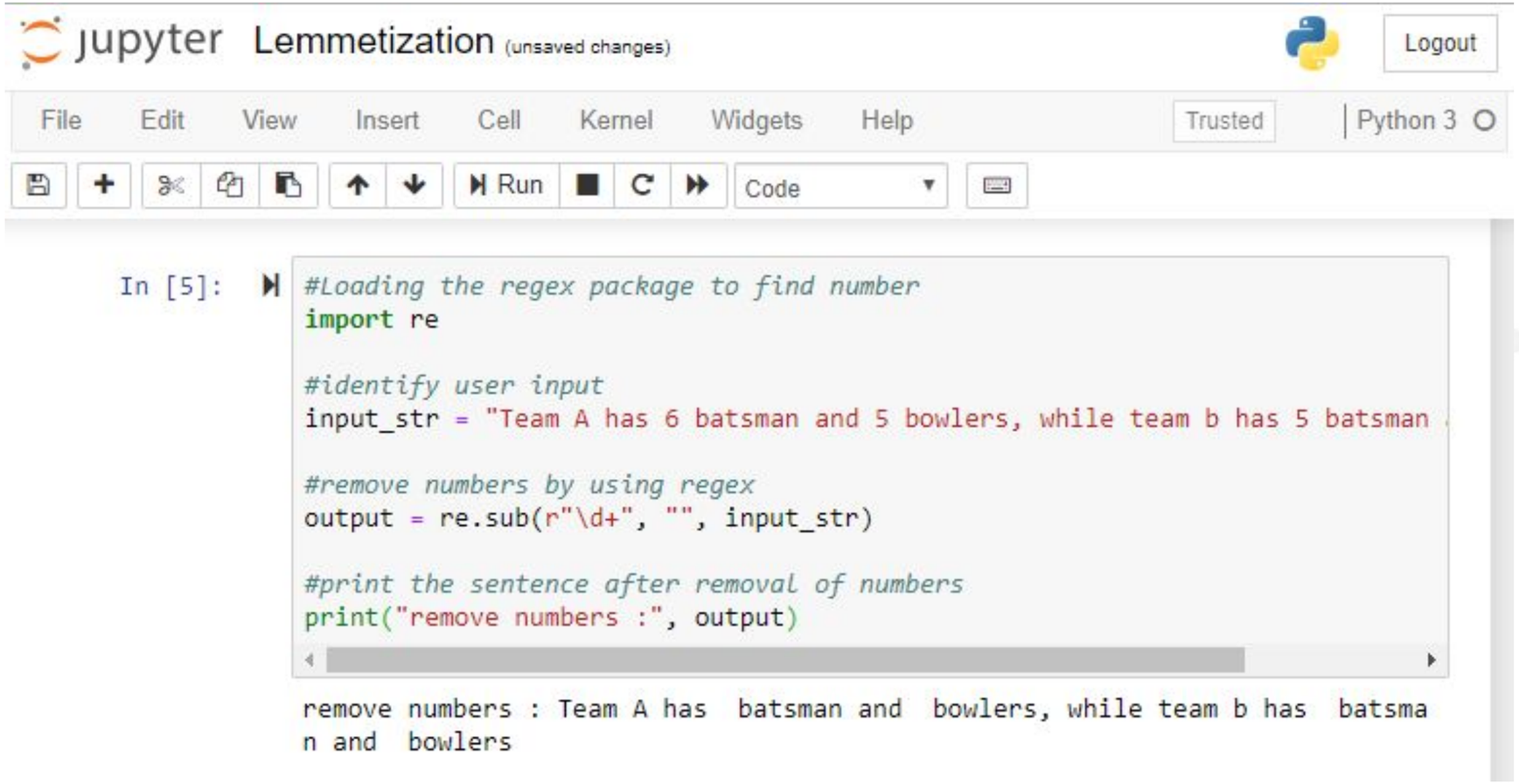
NLTK: Processing Raw Text

Converting numbers into words or removing numbers

```
#Loading the regex package to find number
import re

#identify user input
input_str = "Team A has 6 batsman and 5 bowlers, while team b has
5 batsman and 6 bowlers"

#remove numbers by using regex
output = re.sub(r"\d+", "", input_str)

#print the sentence after removal of numbers
print("remove numbers :", output)
```

NLTK: Processing Raw Text

Output: remove numbers : Team A has batsman and bowlers, while team b has batsman and bowlers



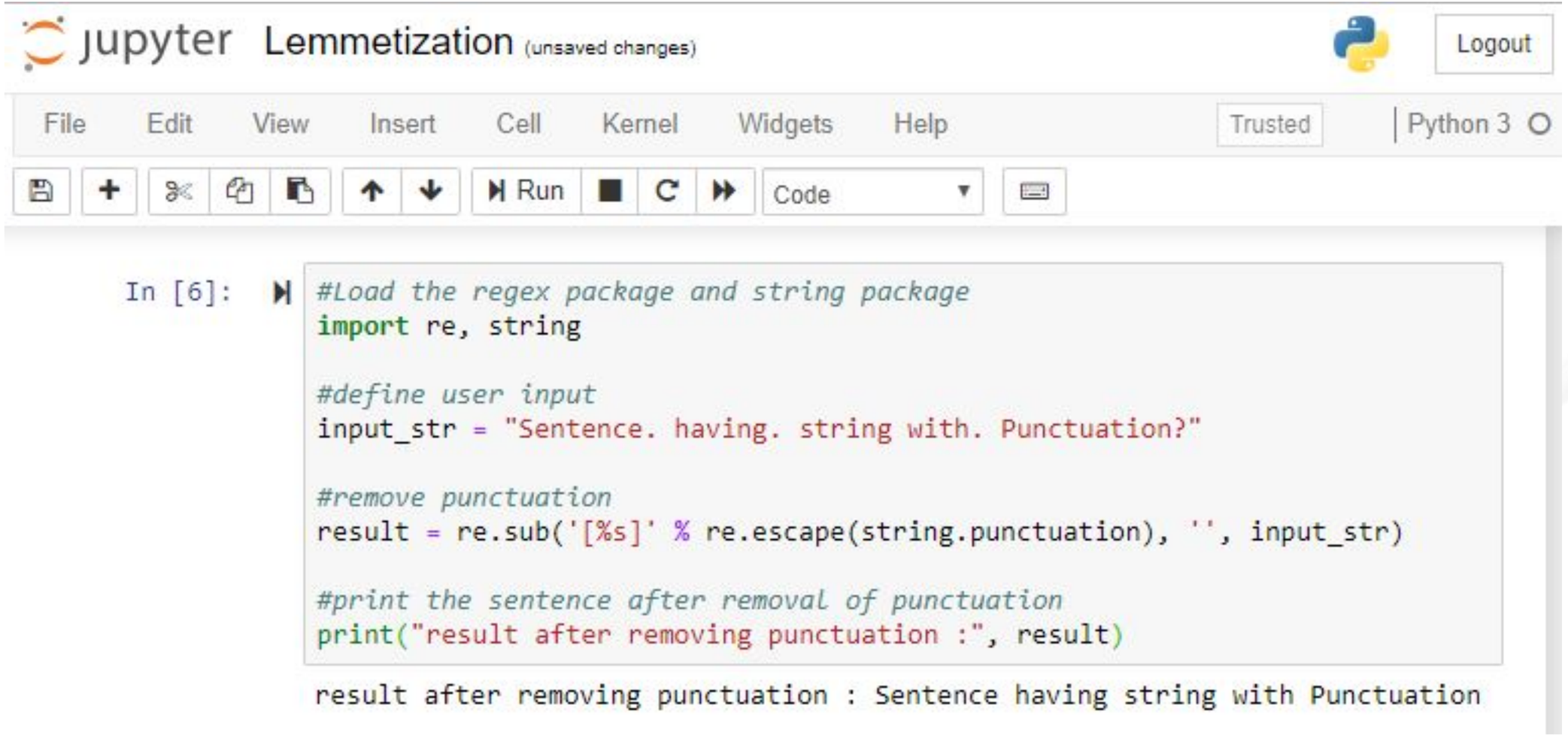
NLTK: Processing Raw Text

Removing accent, punctuations marks, and other diacritics

```
#Load the regex package and string package
import re, string

#define user input
input_str = "Sentence. having. string with. Punctuation?"

#remove punctuation
result = re.sub('[%s]' % re.escape(string.punctuation), '', input_str)

#print the sentence after removal of punctuation
print("result after removing punctuation :", result)
```

NLTK: Processing Raw Text

Output: result after removing punctuation : Sentence having string with Punctuation



NLTK: Processing Raw Text

Removing white spaces:

```
#Load the regex and string package
import re

#define input from user
input_str = 'pythonis programming    language \t\n\r\tHello \t'

#Print the sentence after removing the spaces
print('Remove spaces using regex :', re.sub(r"\s+", "", input_str),"\n", sep='')

#Print the sentence after removing the landing spaces
print('Remove landing spaces using regex :', re.sub(r"^\s+", "",
input_str),"\n", sep='')

#Print the sentence after removing the trailing spaces
print('Remove trailing spaces using regex :', re.sub(r"\s+$", "",
input_str),"\n", sep='')

#Print the sentence after removing the leading and trailing spaces
print('Remove landing spaces using regex :', re.sub(r"^\s+|\s+$", "",
input_str),"\n", sep=‘’)
```

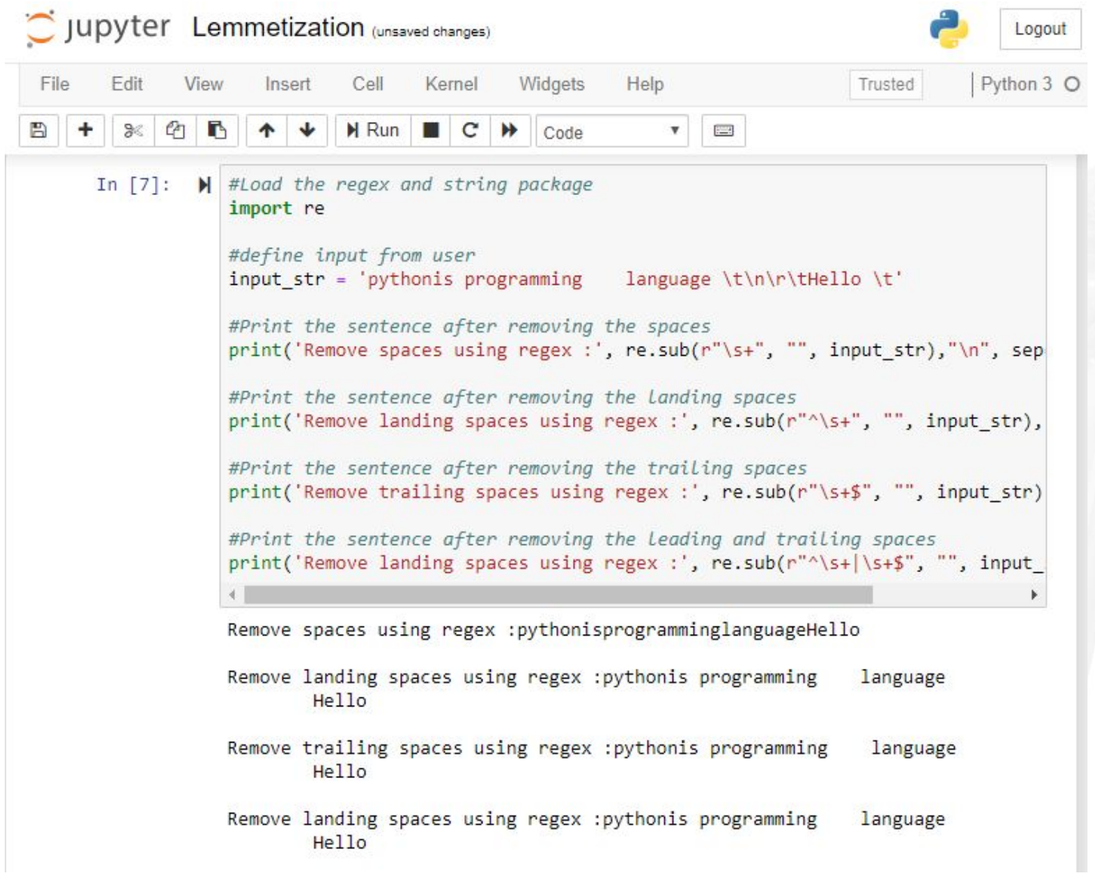
NLTK: Processing Raw Text

Output:
Remove spaces using regex
:pythonisprogramminglanguageHello

Remove landing spaces using regex :pythonis programming language
Hello

Remove trailing spaces using regex :pythonis programming language
Hello

Remove landing spaces using regex :pythonis programming language
Hello

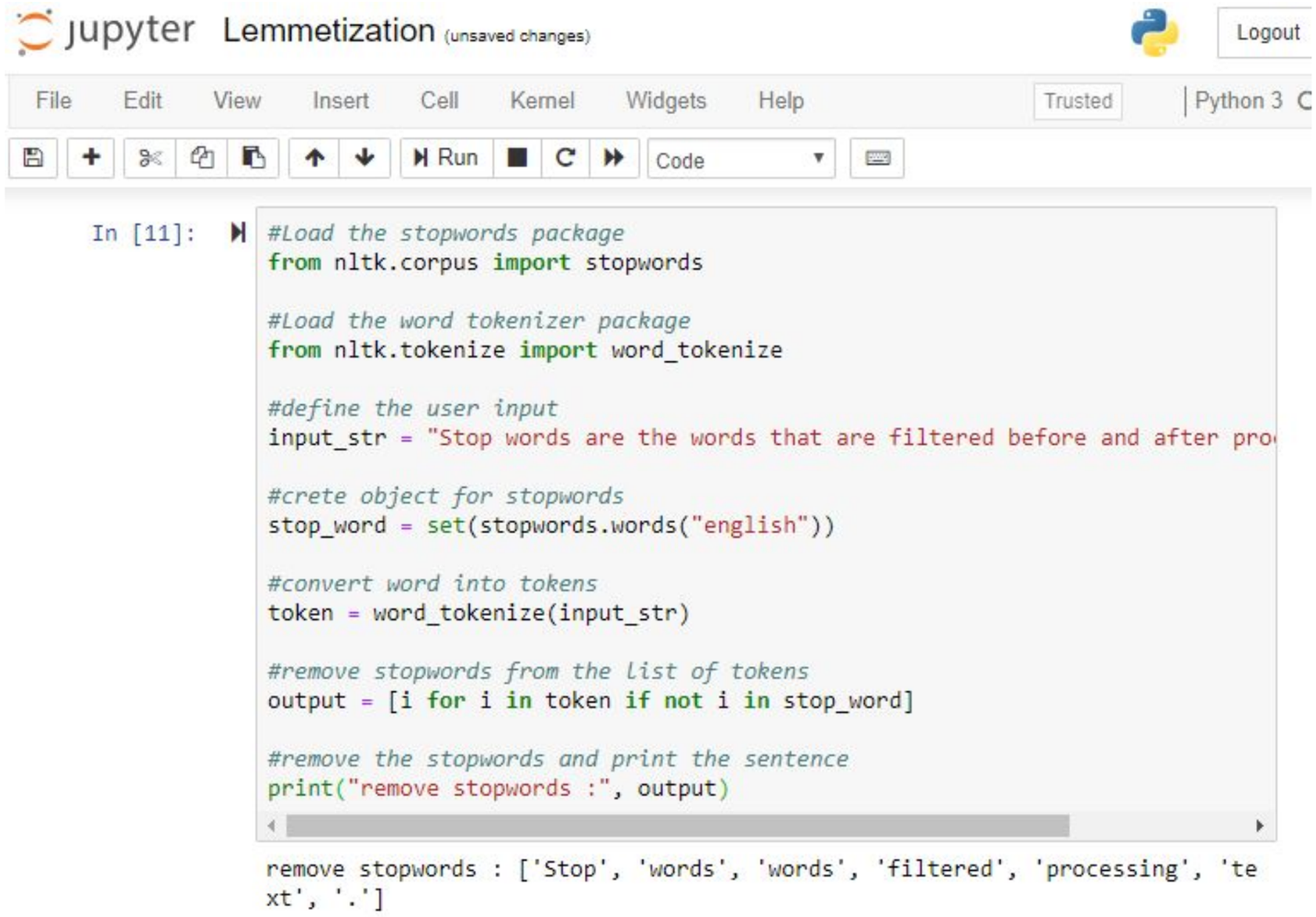


NLTK: Stopwords

```
#Load the stopwords package
from nltk.corpus import stopwords

#Load the word tokenizer package
from nltk.tokenize import word_tokenize

#define the user input
input_str = "Stop words are the words that are filtered before
and after processing of text."

#crete object for stopwords
stop_word = set(stopwords.words("english"))

#convert word into tokens
token = word_tokenize(input_str)

#remove stopwords from the list of tokens
output = [i for i in token if not i in stop_word]

#remove the stopwords and print the sentence
print("remove stopwords :", output)
```

NLTK: Stopwords

Output: remove stopwords : ['Stop', 'words', 'words', 'filtered', 'processing', 'text', '.']

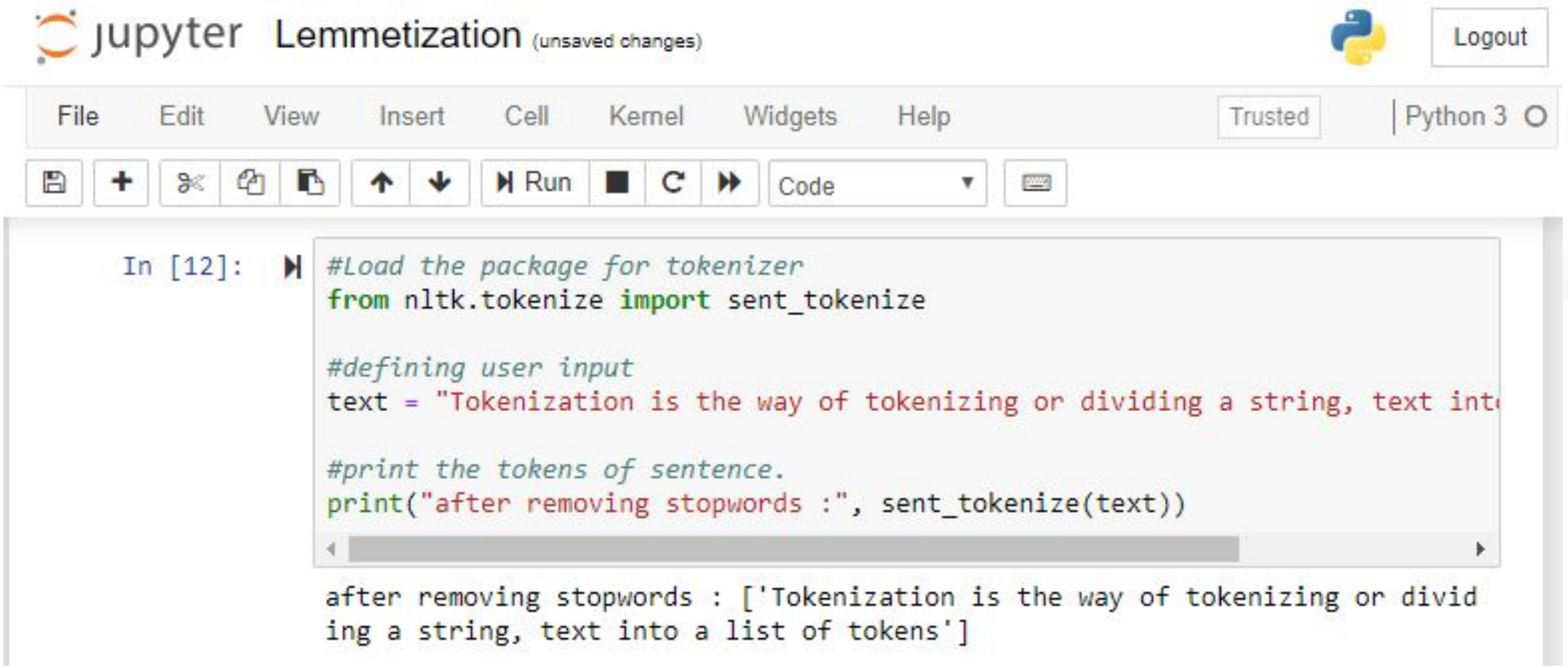


NLTK: Tokenizers

```
#Load the package for tokenizer
from nltk.tokenize import sent_tokenize

#defining user input
text = "Tokenization is the way of tokenizing or dividing
a string, text into a list of tokens"

#print the tokens of sentence.
print("after removing stopwords :", sent_tokenize(text))
```

NLTK: Tokenizers

Output: after removing stopwords : ['Tokenization is the way of tokenizing or dividing a string, text into a list of tokens']



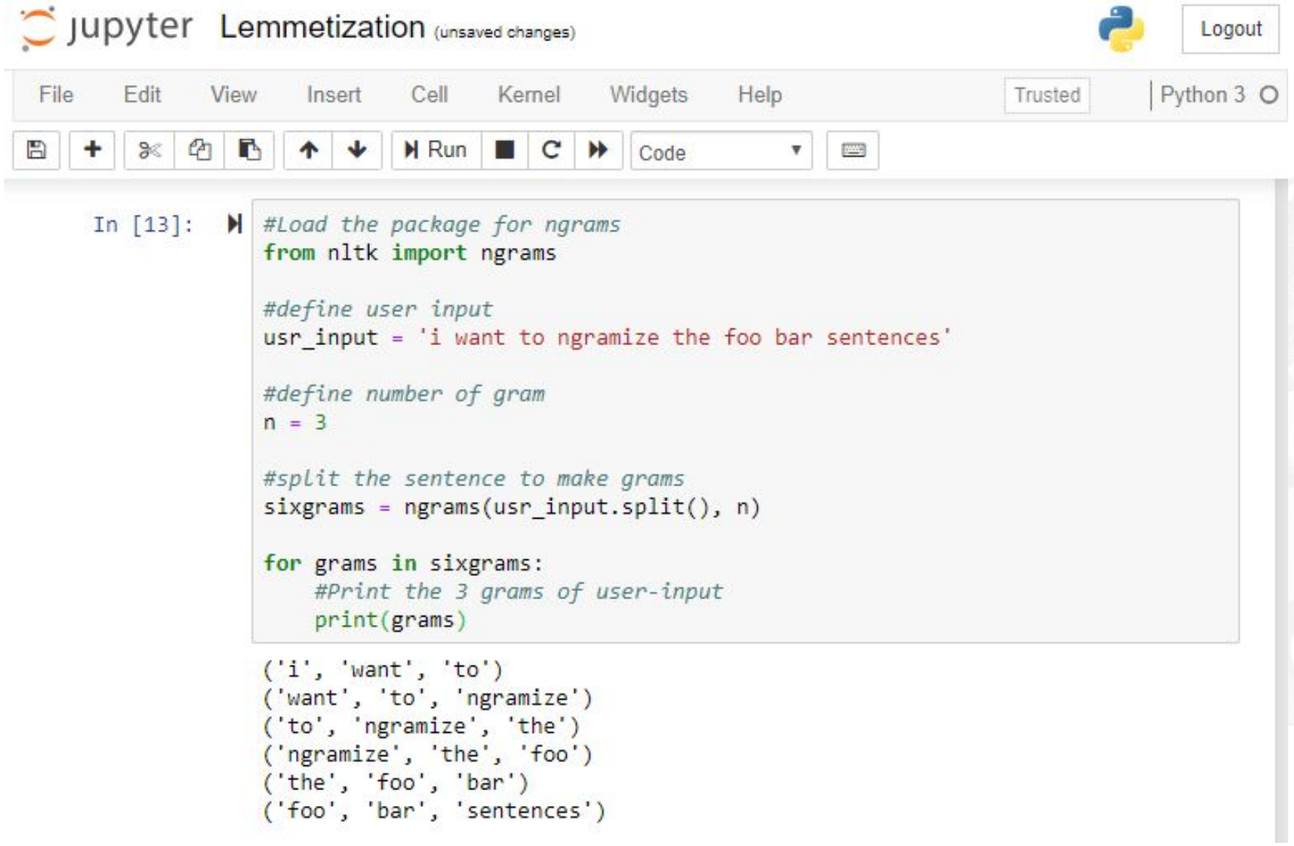
NLTK: Ngram

```
#Load the package for ngrams
from nltk import ngrams

#define user input
usr_input = 'i want to ngramize the foo bar
sentences'

#define number of gram
n = 3

#split the sentence to make grams
sixgrams = ngrams(usr_input.split(), n)

for grams in sixgrams:
    #Print the 3 grams of user-input
    print(grams)
```

NLTK: Ngram

Output:

('i', 'want', 'to')
('want', 'to', 'ngramize')
('to', 'ngramize', 'the')
('ngramize', 'the', 'foo')
('the', 'foo', 'bar')
('foo', 'bar', 'sentences')



NLTK: Limitations

1. Does not support word vectors
2. Is slow
3. Not for production purpose
4. Good only for English and difficult for other languages

DATA AND
ARTIFICIAL INTELLIGENCE

Re

Re: Introduction

- Re is an inbuilt library which comes with python.
- It uses a set of symbols to identify the patterns from the text.
  Example: email address **^([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})$**
- It is used in information retrieval: **import nltk**

```
import re
```

Text Processing Using Stemming and Regular Expression

Problem Statement: Demonstrate text processing using stemming and regular expression.

Access: Click on the **Practice Labs** tab on the left side panel of the LMS. Copy or note the username and password that is generated. Click on the **Launch Lab** button. On the page that appears, enter the username and password in the respective fields, and click **Login**.

ASSISTED PRACTICE

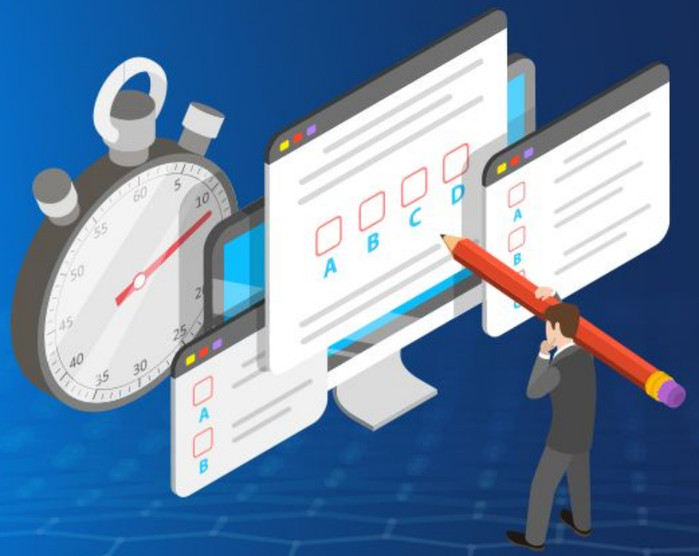
Tweets Cleanup and Analysis Using Regular Expressions

Objective: Use regular expressions to work with messy tweets data: clean up the data, extract hashtags, analyze the most popular hashtags that occur along with a target hashtag (#economy).

Problem Statement: Social media is a gold mine of information. Brands, governments, or anyone can leverage their business with the help of the information contained. It can be information on the sentiments for a brand, or the themes being spoken about , or the associated trends for a particular hashtag. In this project, we will work on the tweets on Twitter. We will find other hashtags that occur frequently with our target hashtag. This will give us an understanding of which other topics people are associating this hashtag with.

LESSON-END PROJECT

DATA AND
ARTIFICIAL INTELLIGENCE



Knowledge Check

Knowledge Check

1

One of the main challenges of NLP is _____.

a. Handling ambiguity of sentences

b. Handling tokenization

c. Both a and b

d. None of the above

Knowledge Check

1

One of the main challenge of NLP is _____.

a. Handling ambiguity of sentences

b. Handling tokenization

c. Both a and b

d. None of the above

The correct answer is **a.**

One of the main challenges of NLP is handling ambiguity of sentences.

Knowledge Check

2

Regular expression is used for_____.

a. Information retrieval

b. Finding the pattern

c. Database management

d. Both a and b

Knowledge Check

2

Regular expression is used for_____.

a. Information retrieval

b. Finding the pattern

c. Database management

d. Both a and b

The correct answer is **d.**

Regular expression is used for information retrieval and finding the pattern.

Knowledge Check

3

NLP is the technique of interpretation of all types of languages which includes _____.

a. Human Language

b. Assembly Language

c. Machine Language

d. Binary Data

Knowledge Check

3

NLP is technique for interpretation of all type of languages which includes _____.

a. Human Language

b. Assembly Language

c. Machine Language

d. Binary Data

The correct answer is **a.**

NLP has its focus on understanding the human spoken or written language and converting that interpretation into machine understandable language.

Knowledge Check

4

Natural Language Processing (NLP) is a field of _____.

a. Computer Science

b. Artificial Intelligence

c. Linguistics

d. All of the above

Knowledge Check

4

Natural Language Processing (NLP) is a field of _____.

a. Computer Science

b. Artificial Intelligence

c. Linguistics

d. All of the above

The correct answer is **d.**

Natural Language Processing is a field of computer science, artificial intelligence, and linguistics.

Knowledge Check

5

Which of the following techniques can be used for the purpose of keyword normalization?
1- Lemmatization 2- Levenshtein 3- Stemming 4- POS

a. 1 and 2

b. 2 and 4

c. 1 and 3

d. 1,2, and 3

Knowledge Check

5

Which of the following techniques can be used for the purpose of keyword normalization?
1- Lemmatization 2- Levenshtein 3- Stemming 4- POS

a. 1 and 2

b. 2 and 4

c. 1 and 3

d. 1,2, and 3

The correct answer is **c.**

Lemmatization and stemming are the techniques of keyword normalization.

Key Takeaways

You are now able to:

- Describe natural language processing and its components
- Explain the different applications of NLP
- Define and demonstrate text processing