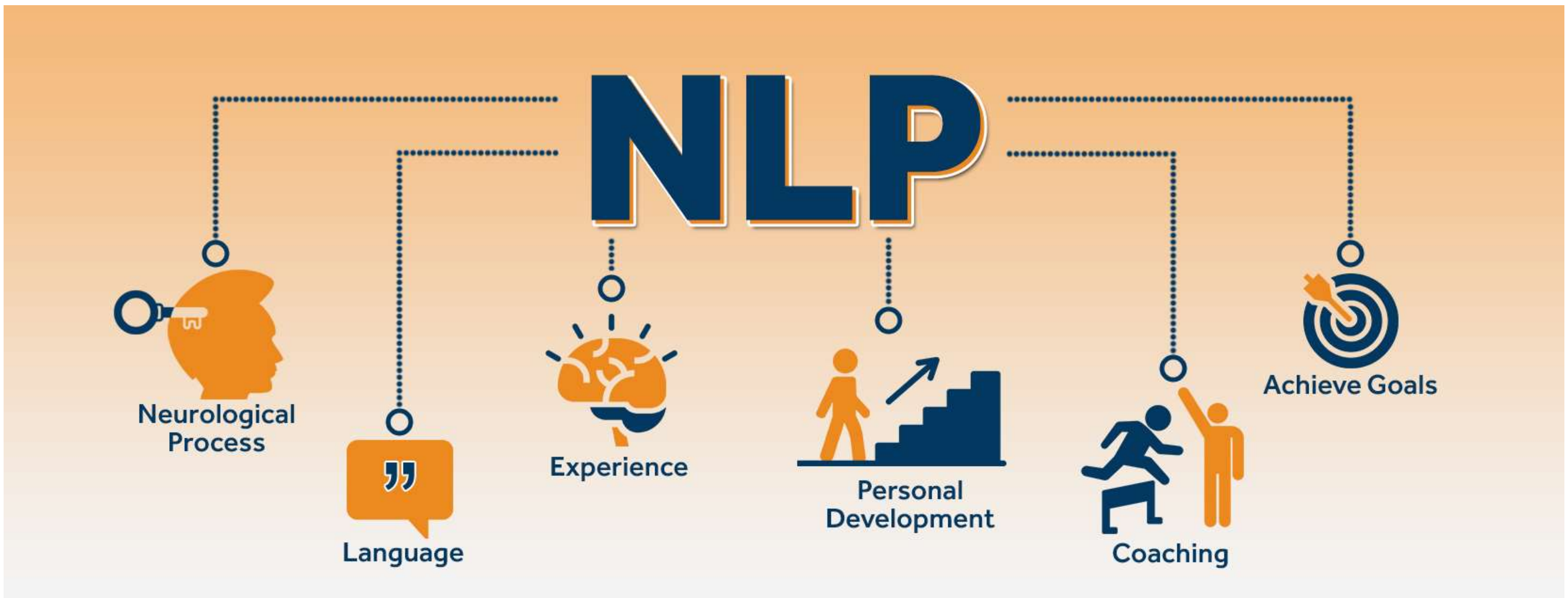
*Auther : −Prince. Kumar. sharma*



## 🪄*Text − Preprocessing − 1*

--------------------------------------------------------------------------------------------$--------------------------------------------------------------------------------------------

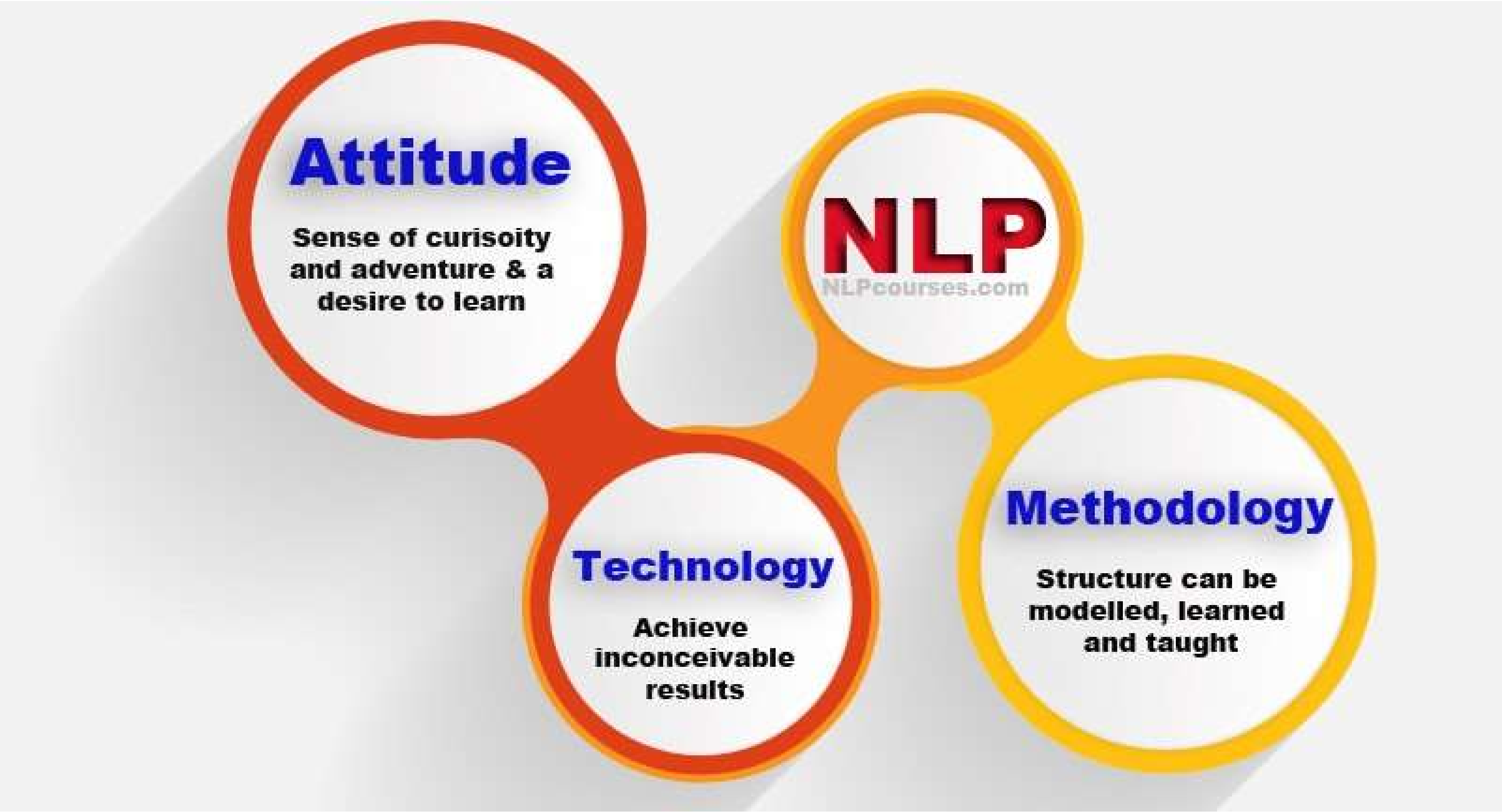
**Text preprocessing is the process of cleaning and transforming unstructured text data to prepare it for analysis. It involves techniques such as tokenization, lowercasing, normalization, stop-word removal, and part-of-speech tagging**

--------------------------------------------------------------------------------------------*OR*--------------------------------------------------------------------------------------------

Text preprocessing is an essential step in natural language processing (NLP) that involves cleaning and transforming unstructured text data to prepare it for analysis12345. The various preprocessing steps that are involved include:

1. Tokenization
2. Stemming
3. Lemmatization
4. Stop-word removal
5. Part-of-speech tagging

1. **Tokenization** :- It is a techinque that break the sentence into word or convert the document into a sentences.
2. **Stemming** :- Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers.
3. **Lemmatization** :-Lemmatization is simmilar to stemming but there is few difference to each other Lemmatization Provides root word and meaningfull word and his time complexity is High compare to stemming.
4. **Stop word** :- In Natural Language Processing (NLP), stop words are words that are commonly used in a language but do not carry much meaning and are usually ignored during text analysis
5. **Part of speech Tagging** :- Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP) that involves assigning a grammatical tag to each word in a sentence based on its definition and context


Attitude
Sense of curisoity and adventure & a desire to learn
NLP
NLPcourses.com
Technology
Achieve inconceivable results
Methodology
Structure can be modelled, learned and taught

# *Importing.Useful.Library*

In [29]:
```
import pandas as pd
import numpy as np
import scipy.stats as st
from sklearn.feature_extraction.text import CountVectorizer
from nltk import word_tokenize,sent_tokenize
from nltk.stem import PorterStemmer,WordNetLemmatizer
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer,TfidfTransformer
import re
from sklearn.compose import ColumnTransformer
```

# ✨*TextPreprocessingwiththehelpofNaturalLanguageProcessing*

In [30]:
```
x ="""Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines
the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions,
variations in sentence structure—these just a few of the irregularities of human language that take humans years to learn,
but that programmers must teach natural language-driven applications to recognize and understand accurately from the start,
if those applications are going to be useful.

Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting.
Some of these tasks include the following:

Speech recognition, also called speech-to-text, is the task of reliably converting voice data into text data. Speech
recognition is required for any application that follows voice commands or answers spoken questions. What makes speech
recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and
intonation, in different accents, and often using incorrect grammar.
Part of speech tagging, also called grammatical tagging, is the process of determining the part of speech of a particular
word or piece of text based on its use and context. Part of speech identifies ‘make’ as a verb in ‘I can make a paper plane,’
and as a noun in ‘What make of car do you own?’
Word sense disambiguation is the selection of the meaning of a word with multiple meanings  through a process of semantic
analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps
distinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a bet’ (place).
Named entity recognition, or NEM, identifies words or phrases as useful entities. NEM identifies ‘Kentucky’ as a location
or ‘Fred’ as a man's name.
Co-reference resolution is the task of identifying if and when two words refer to the same entity. The most common example
is determining the person or object to which a certain pronoun refers (e.g., ‘she’ = ‘Mary’),  but it can also involve
identifying a metaphor or an idiom in the text  (e.g., an instance in which 'bear' isn't an animal but a large hairy person).
Sentiment analysis attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.
Natural language generation is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of
putting structured information into human language."""
```

In [31]:
```
x
```

Out[31]:
```
"Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines\nthe
intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions,\nvariat
ions in sentence structure—these just a few of the irregularities of human language that take humans years to learn, \nbut that
programmers must teach natural language-driven applications to recognize and understand accurately from the start,\nif those ap
plications are going to be useful.\n\nSeveral NLP tasks break down human text and voice data in ways that help the computer mak
e sense of what it's ingesting. \nSome of these tasks include the following:\n\nSpeech recognition, also called speech-to-text,
is the task of reliably converting voice data into text data. Speech\nrecognition is required for any application that follows
voice commands or answers spoken questions. What makes speech\nrecognition especially challenging is the way people talk—quickl
y, slurring words together, with varying emphasis and \nintonation, in different accents, and often using incorrect grammar.\nP
art of speech tagging, also called grammatical tagging, is the process of determining the part of speech of a particular\nword
or piece of text based on its use and context. Part of speech identifies ‘make’ as a verb in ‘I can make a paper plane,’ \nand
as a noun in ‘What make of car do you own?’\nWord sense disambiguation is the selection of the meaning of a word with multiple
meanings  through a process of semantic \nanalysis that determine the word that makes the most sense in the given context. For
example, word sense disambiguation helps \ndistinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a
bet’ (place).\nNamed entity recognition, or NEM, identifies words or phrases as useful entities. NEM identifies ‘Kentucky’ as a
location\nor ‘Fred’ as a man's name.\nCo-reference resolution is the task of identifying if and when two words refer to the sam
e entity. The most common example \nis determining the person or object to which a certain pronoun refers (e.g., ‘she’ = ‘Mar
y’),  but it can also involve \nidentifying a metaphor or an idiom in the text  (e.g., an instance in which 'bear' isn't an ani
mal but a large hairy person).\nSentiment analysis attempts to extract subjective qualities—attitudes, emotions, sarcasm, confu
sion, suspicion—from text.\nNatural language generation is sometimes described as the opposite of speech recognition or speech-
to-text; it's the task of\nputting structured information into human language."
```

In [7]:
```
from nltk import sent_tokenize
from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer,WordNetLemmatizer
import re
```

In [8]:
```
x = sent_tokenize(x)
```

In [12]:
```
x[2]
```

Out[12]:
```
"Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting."
```

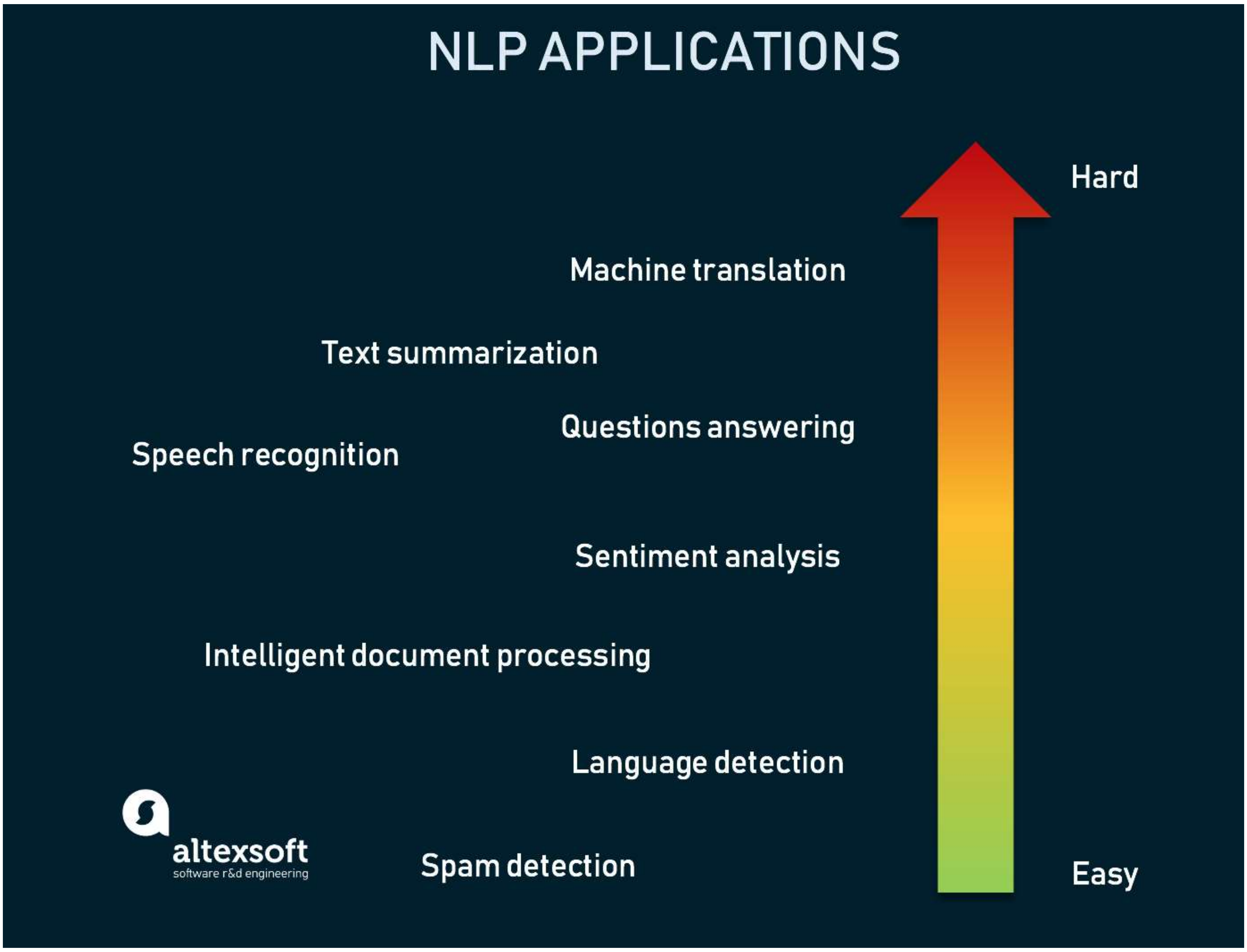
In [17]:
```
lis = []
for i in x:
    word = re.sub("[^a-zA-Z]"," ",i)  #removing the Punchuations from this text
    word = re.sub(" +"," ",word)  # removing extra spaces from the text
    word = word.lower()  # converting the text into a lower word
    #word = "".join(word)
    lis.append(word)  # adding all text into a new word
```

In [18]:
```
lis
```

Out[18]:
```
['human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines the
intended meaning of text or voice data ',
 'homonyms homophones sarcasm idioms metaphors grammar and usage exceptions variations in sentence structure these just a few o
f the irregularities of human language that take humans years to learn but that programmers must teach natural language driven
applications to recognize and understand accurately from the start if those applications are going to be useful ',
 'several nlp tasks break down human text and voice data in ways that help the computer make sense of what it s ingesting ',
 'some of these tasks include the following speech recognition also called speech to text is the task of reliably converting vo
ice data into text data ',
 'speech recognition is required for any application that follows voice commands or answers spoken questions ',
 'what makes speech recognition especially challenging is the way people talk quickly slurring words together with varying emph
asis and intonation in different accents and often using incorrect grammar ',
 'part of speech tagging also called grammatical tagging is the process of determining the part of speech of a particular word
or piece of text based on its use and context ',
 'part of speech identifies make as a verb in i can make a paper plane and as a noun in what make of car do you own word sense
disambiguation is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that det
ermine the word that makes the most sense in the given context ',
 'for example word sense disambiguation helps distinguish the meaning of the verb make in make the grade achieve vs make a bet
place ',
 'named entity recognition or nem identifies words or phrases as useful entities ',
 'nem identifies kentucky as a location or fred as a man s name ',
 'co reference resolution is the task of identifying if and when two words refer to the same entity ',
 'the most common example is determining the person or object to which a certain pronoun refers e g she mary but it can also in
volve identifying a metaphor or an idiom in the text e g an instance in which bear isn t an animal but a large hairy person ',
 'sentiment analysis attempts to extract subjective qualities attitudes emotions sarcasm confusion suspicion from text ',
 'natural language generation is sometimes described as the opposite of speech recognition or speech to text it s the task of p
utting structured information into human language ']
```

In [24]:
```
for i in lis:
    x = word_tokenize(i)
    word = [WordNetLemmatizer().lemmatize(ww,pos="v") for ww in x if ww not in set(stopwords.words("english"))]
    word =" ".join(word)
    print(word)
```

```
human language fill ambiguities make incredibly difficult write software accurately determine intend mean text voice data
homonyms homophones sarcasm idioms metaphors grammar usage exceptions variations sentence structure irregularities human langua
ge take humans years learn programmers must teach natural language drive applications recognize understand accurately start app
lications go useful
several nlp task break human text voice data ways help computer make sense ingest
task include follow speech recognition also call speech text task reliably convert voice data text data
speech recognition require application follow voice command answer speak question
make speech recognition especially challenge way people talk quickly slur word together vary emphasis intonation different acce
nt often use incorrect grammar
part speech tag also call grammatical tag process determine part speech particular word piece text base use context
part speech identify make verb make paper plane noun make car word sense disambiguation selection mean word multiple mean proce
ss semantic analysis determine word make sense give context
example word sense disambiguation help distinguish mean verb make make grade achieve vs make bet place
name entity recognition nem identify word phrase useful entities
nem identify kentucky location fred man name
co reference resolution task identify two word refer entity
common example determine person object certain pronoun refer e g mary also involve identify metaphor idiom text e g instance be
ar animal large hairy person
sentiment analysis attempt extract subjective qualities attitudes emotions sarcasm confusion suspicion text
natural language generation sometimes describe opposite speech recognition speech text task put structure information human lan
guage
```



NLP APPLICATIONS

Hard

Machine translation

Text summarization

Questions answering

Speech recognition

Sentiment analysis

Intelligent document processing

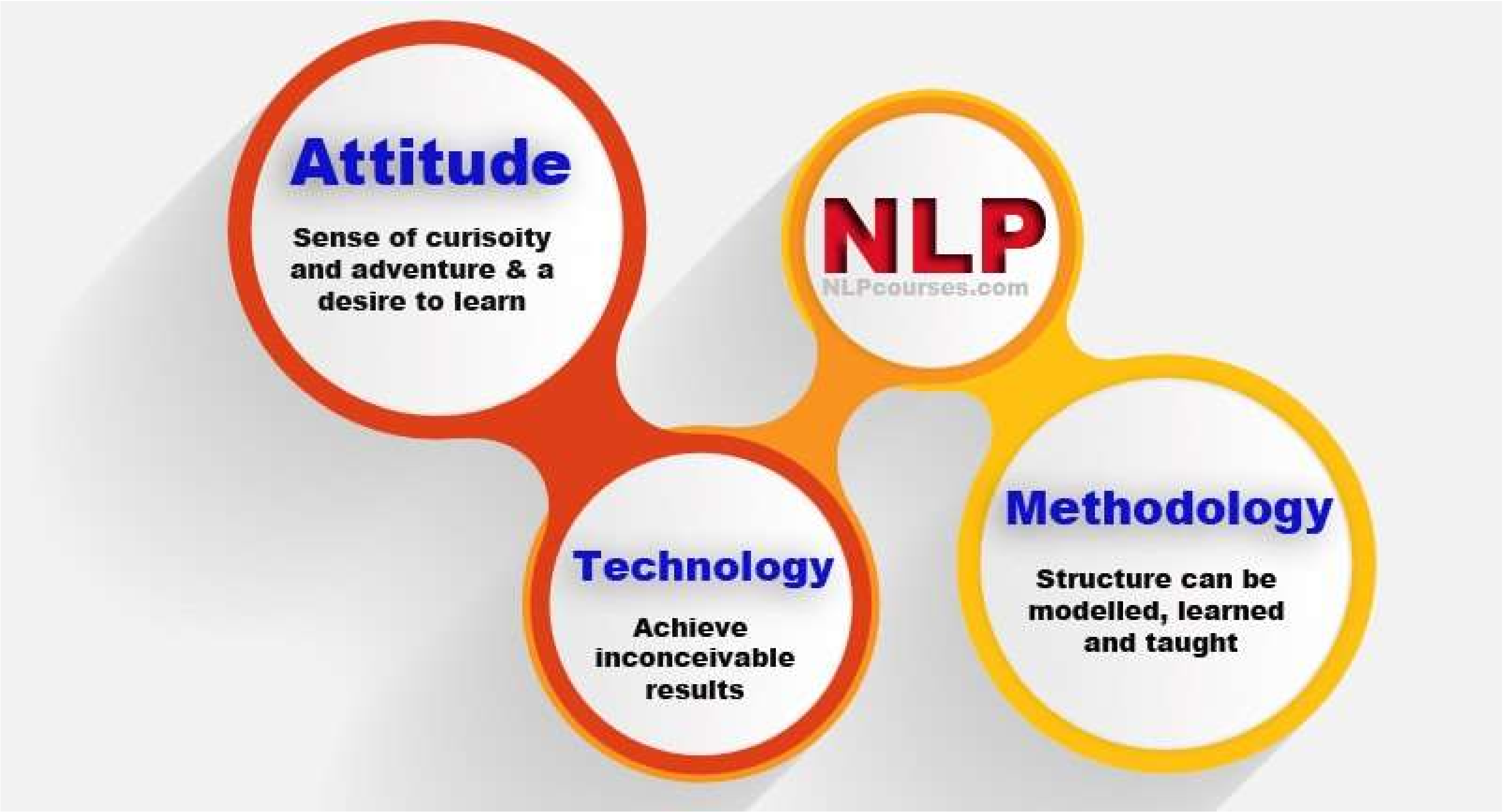
Language detection

Spam detection

Easy

altexsoft
software r&d engineering

### Thanks for watching the Sort Notebook of NLP



**Attitude**

Sense of curisoity and adventure & a desire to learn

**NLP**

NLPcourses.com

**Technology**

Achieve inconceivable results

**Methodology**

Structure can be modelled, learned and taught

In [ ]: