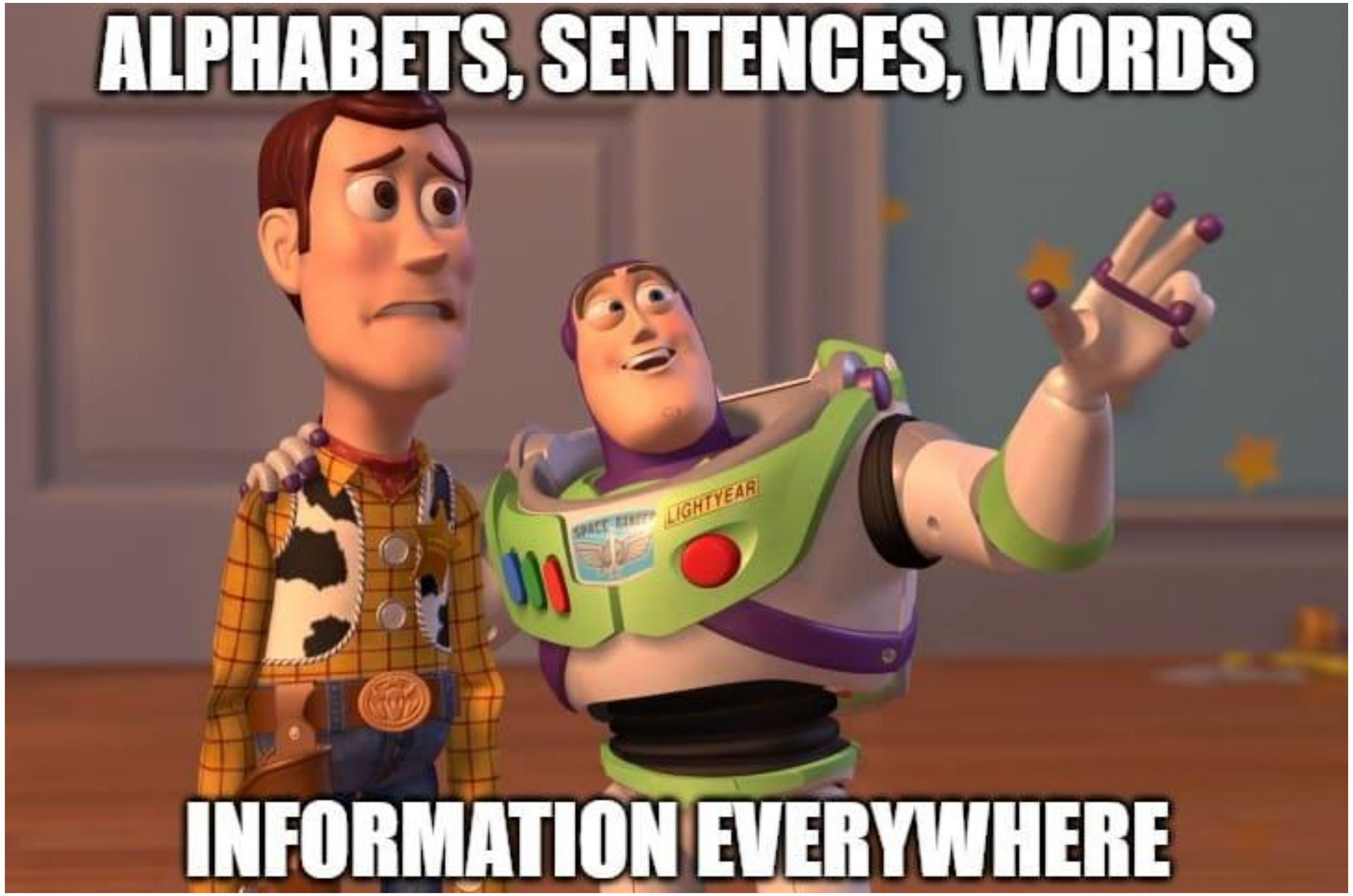
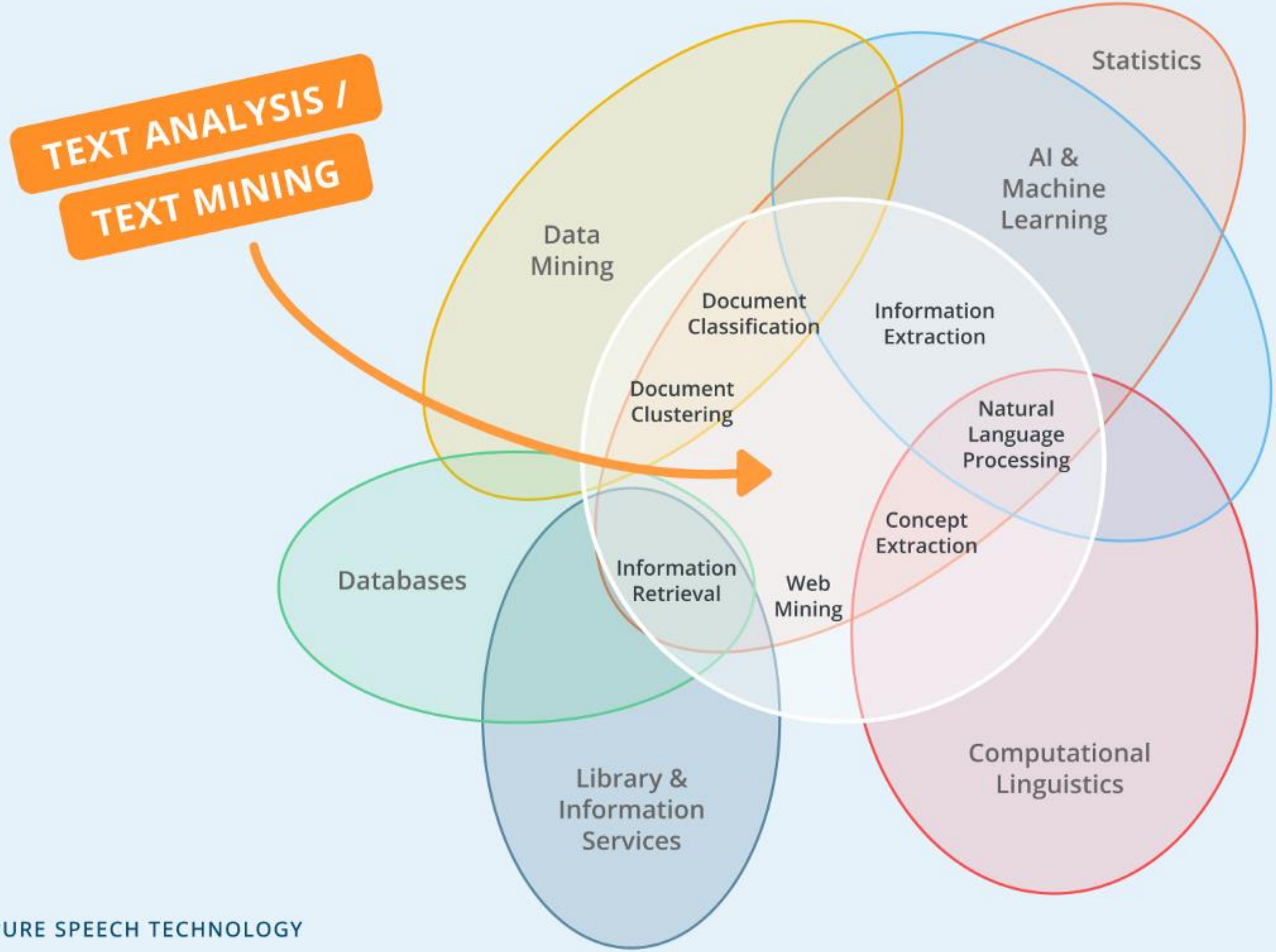
Text Handling

Kunaal Naik

What is text?


ALPHABETS, SENTENCES, WORDS
INFORMATION EVERYWHERE
LIGHTYEAR



PURE SPEECH TECHNOLOGY

Text Analytics Use Cases

Manufacturers

- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products

Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy

Financial Institutions

- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation

Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media

Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications

Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data

Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments

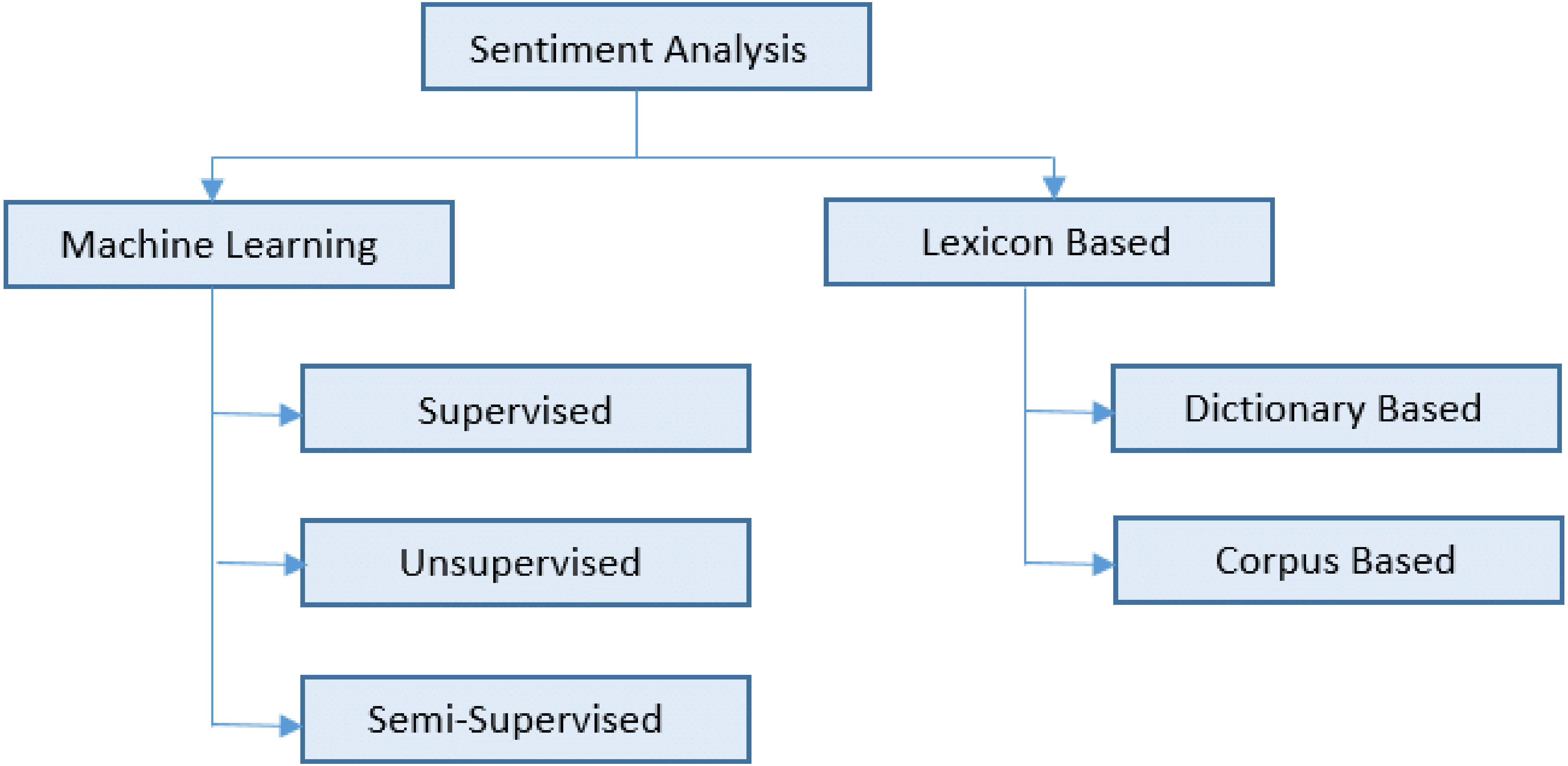
Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials

Insurance

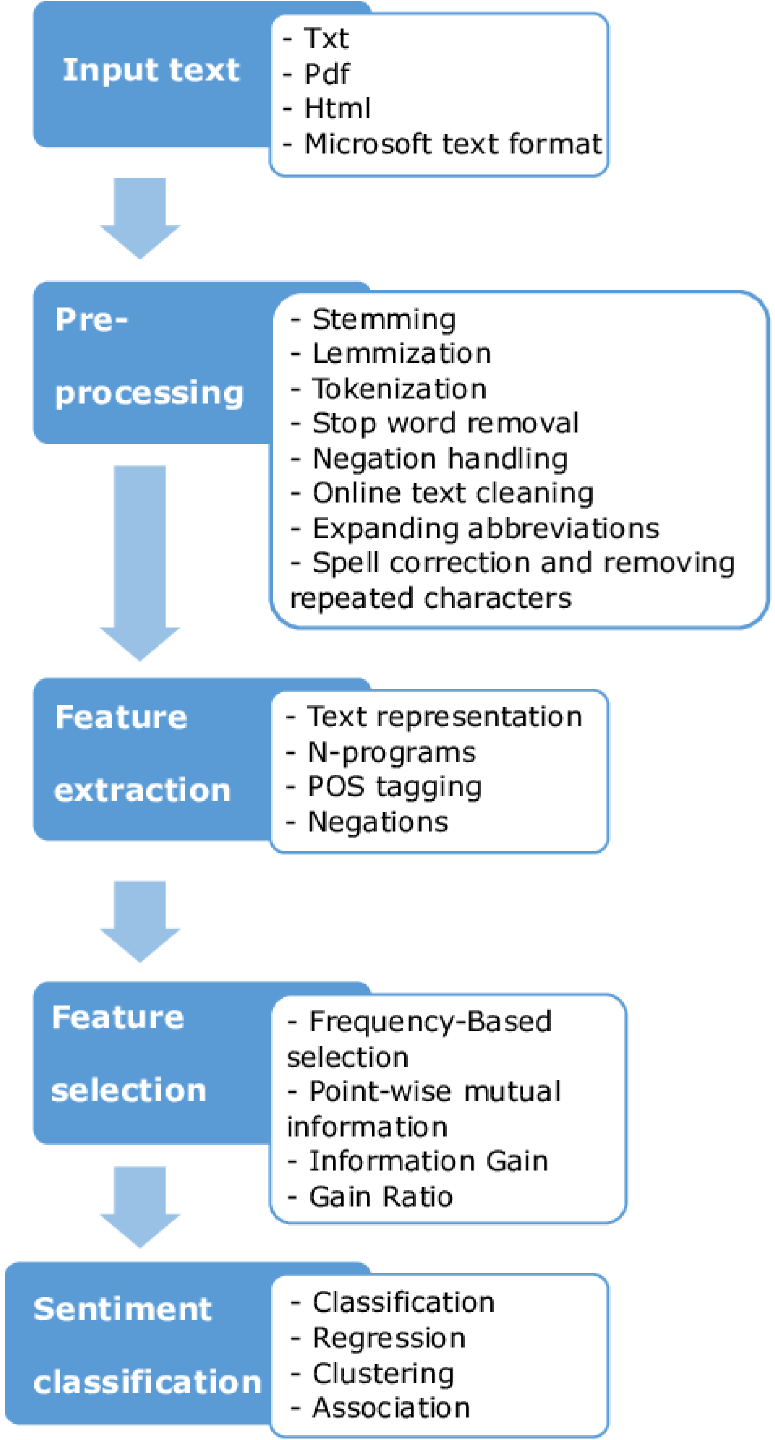
- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

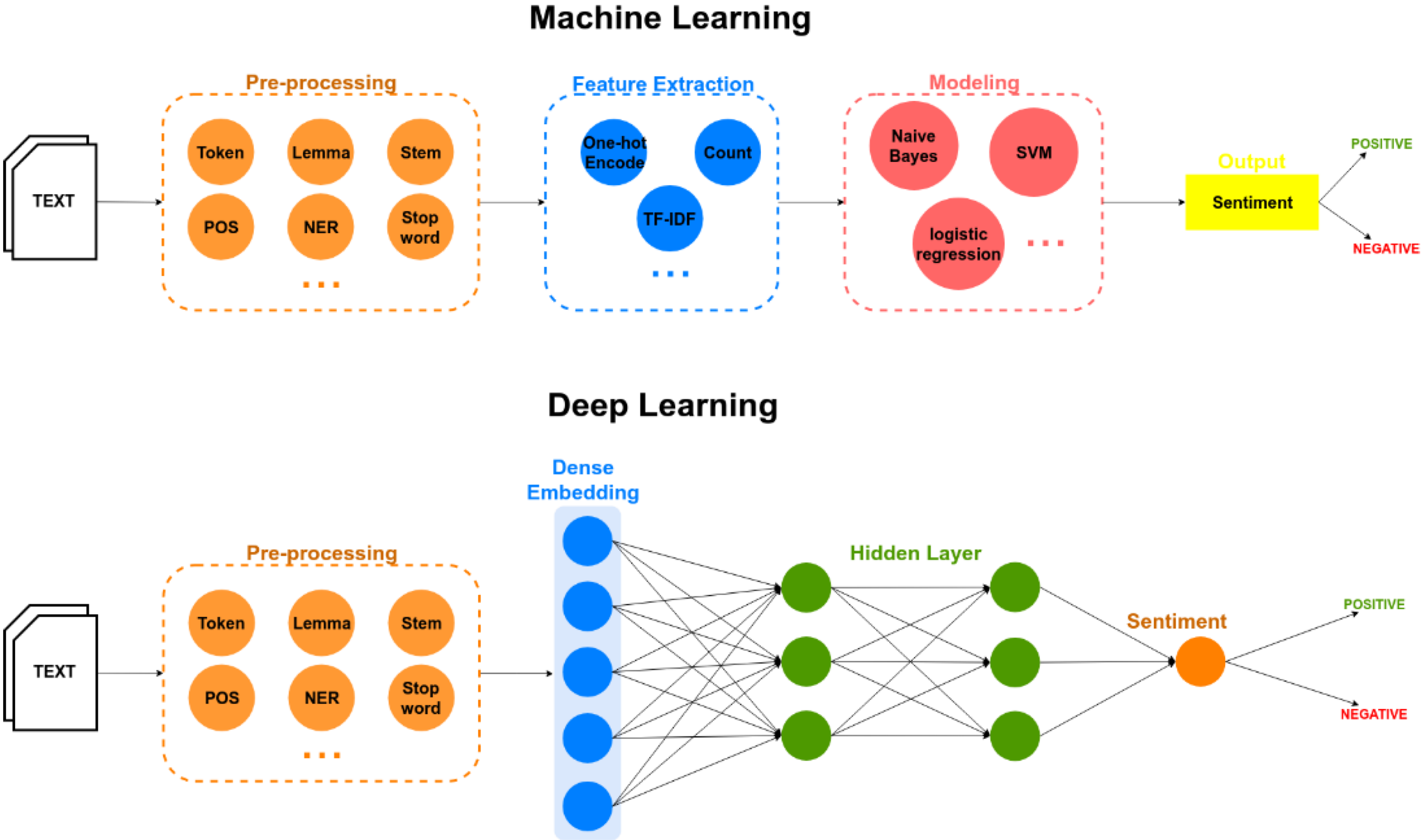
zencos


Sentiment Analysis
Machine Learning
Lexicon Based
Supervised
Unsupervised
Semi-Supervised
Dictionary Based
Corpus Based

How does the process look like?

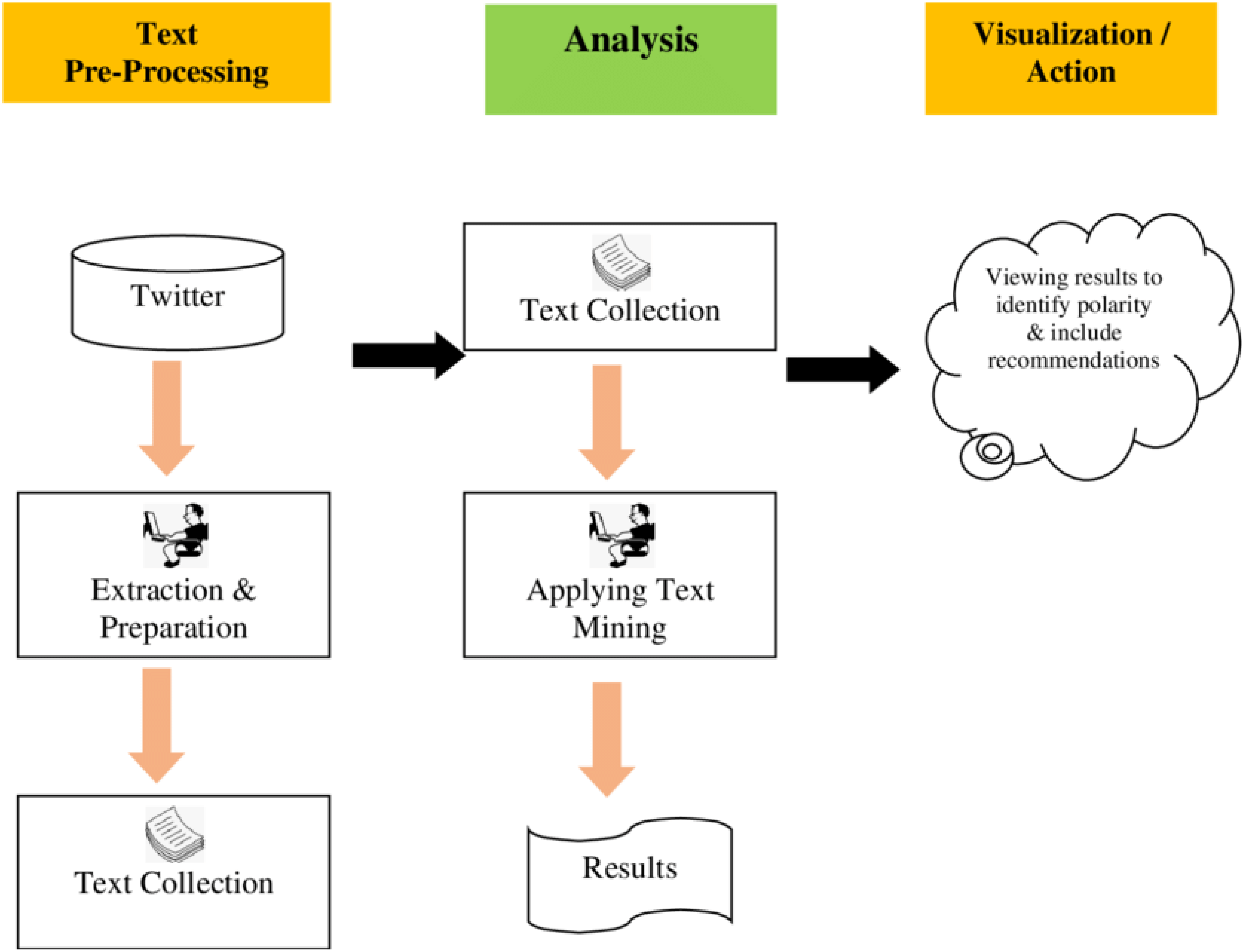
Very complicated!

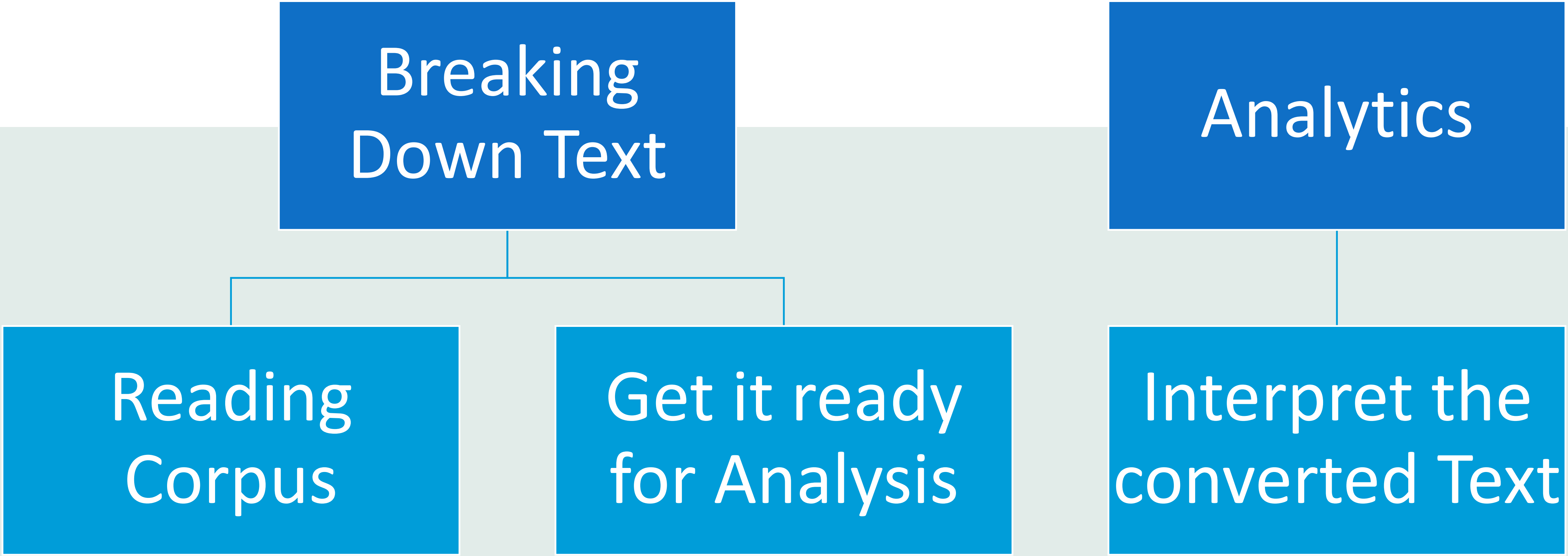

Input text
- Txt
- Pdf
- Html
- Microsoft text format
Pre-processing
- Stemming
- Lemmization
- Tokenization
- Stop word removal
- Negation handling
- Online text cleaning
- Expanding abbreviations
- Spell correction and removing repeated characters
Feature extraction
- Text representation
- N-programs
- POS tagging
- Negations
Feature selection
- Frequency-Based selection
- Point-wise mutual information
- Information Gain
- Gain Ratio
Sentiment classification
- Classification
- Regression
- Clustering
- Association



Machine Learning

Deep Learning


Text Pre-Processing
Analysis
Visualization / Action
Twitter
Extraction & Preparation
Text Collection
Text Collection
Applying Text Mining
Results
Viewing results to identify polarity & include recommendations



Breaking Down Text

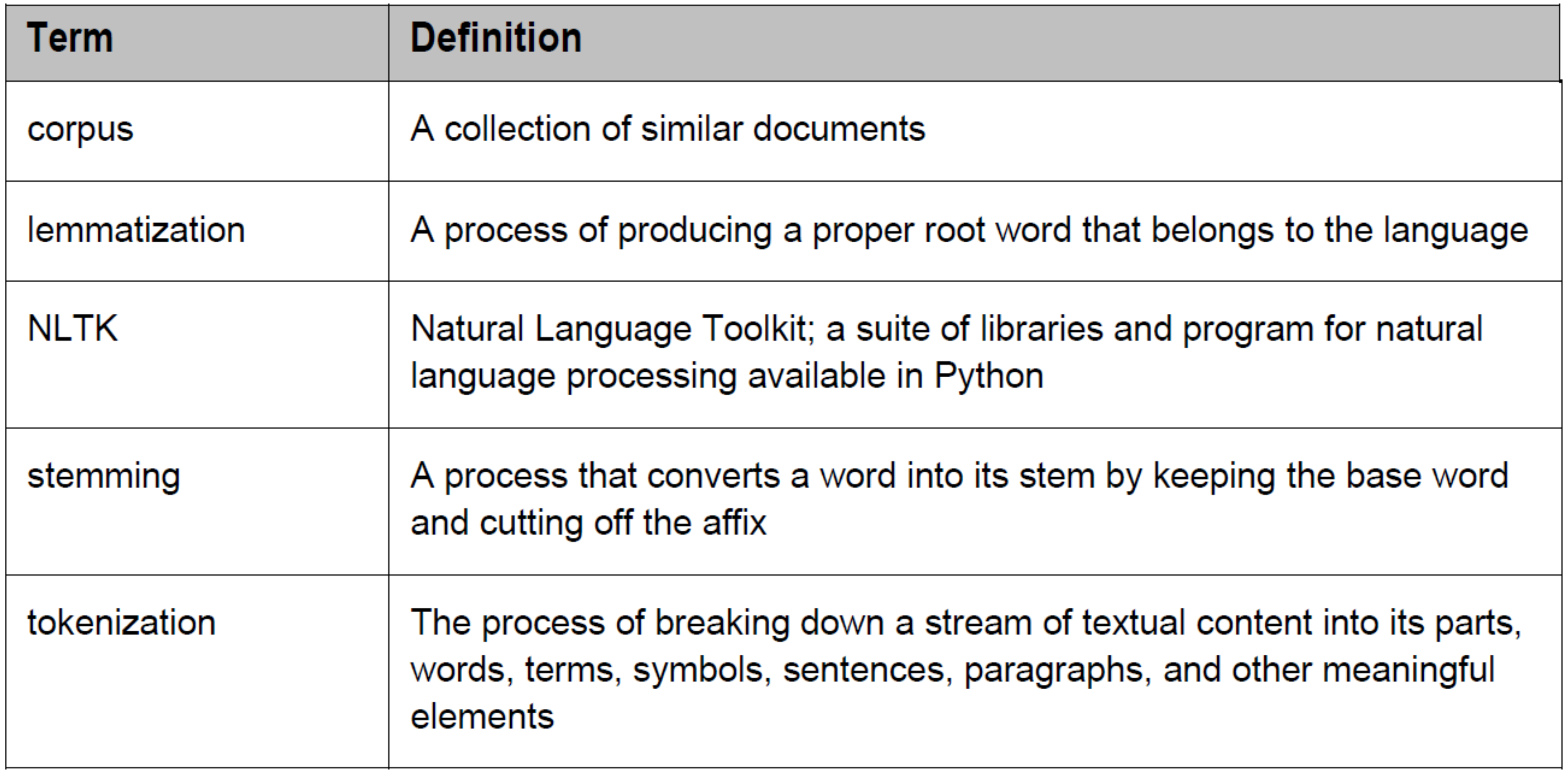
Reading Corpus

Get it ready for Analysis

Analytics

Interpret the converted Text

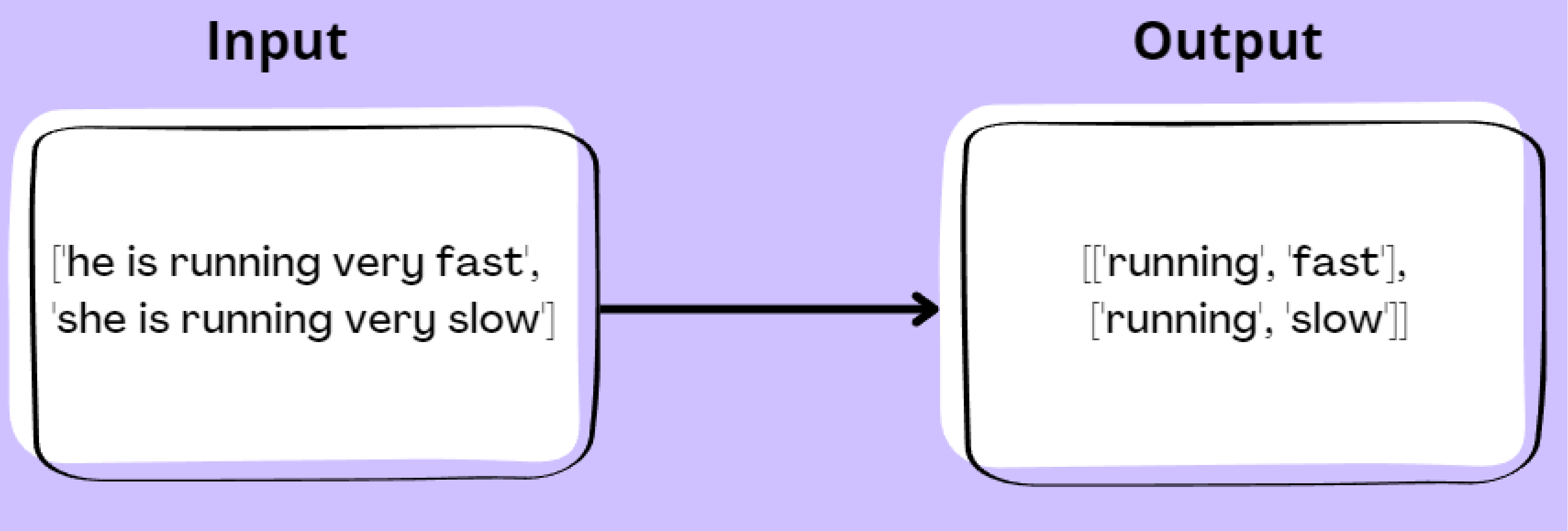
Text Pre-processing (New Terms!)



Term	Definition
corpus	A collection of similar documents
lemmatization	A process of producing a proper root word that belongs to the language
NLTK	Natural Language Toolkit; a suite of libraries and program for natural language processing available in Python
stemming	A process that converts a word into its stem by keeping the base word and cutting off the affix
tokenization	The process of breaking down a stream of textual content into its parts, words, terms, symbols, sentences, paragraphs, and other meaningful elements

- Punctuations
- Stop Words

Stop word Removal



Input

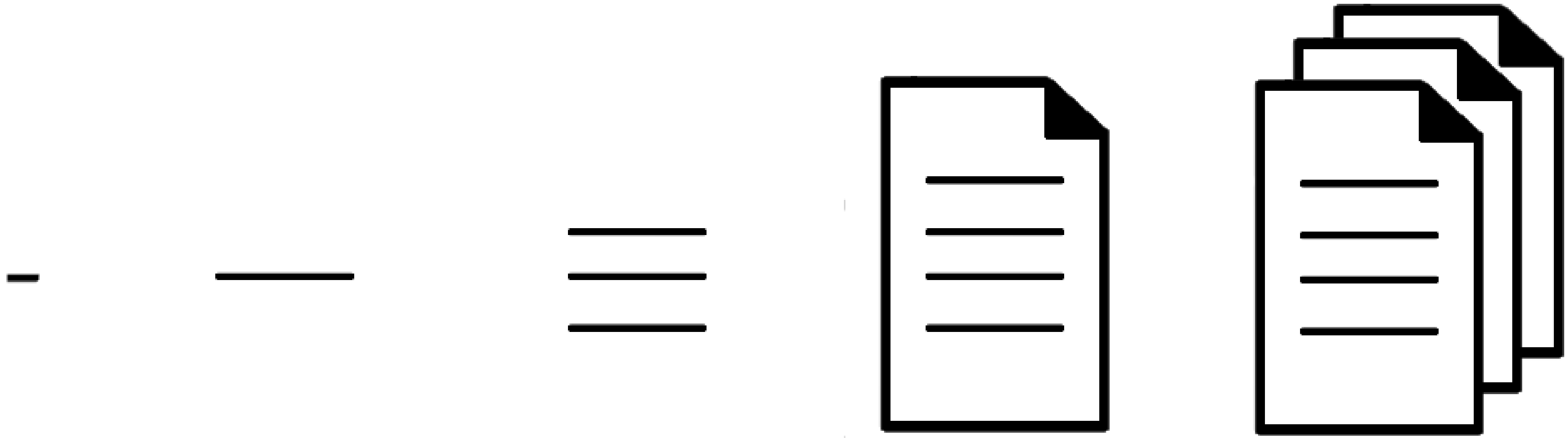
['he is running very fast',
'she is running very slow']

Output

[['running', 'fast'],
['running', 'slow']]

Corpus

A collection of similar documents



Token **Sentence** **Paragraph** **Document** **Corpus**

One Row in a table

Multiple Rows in a table

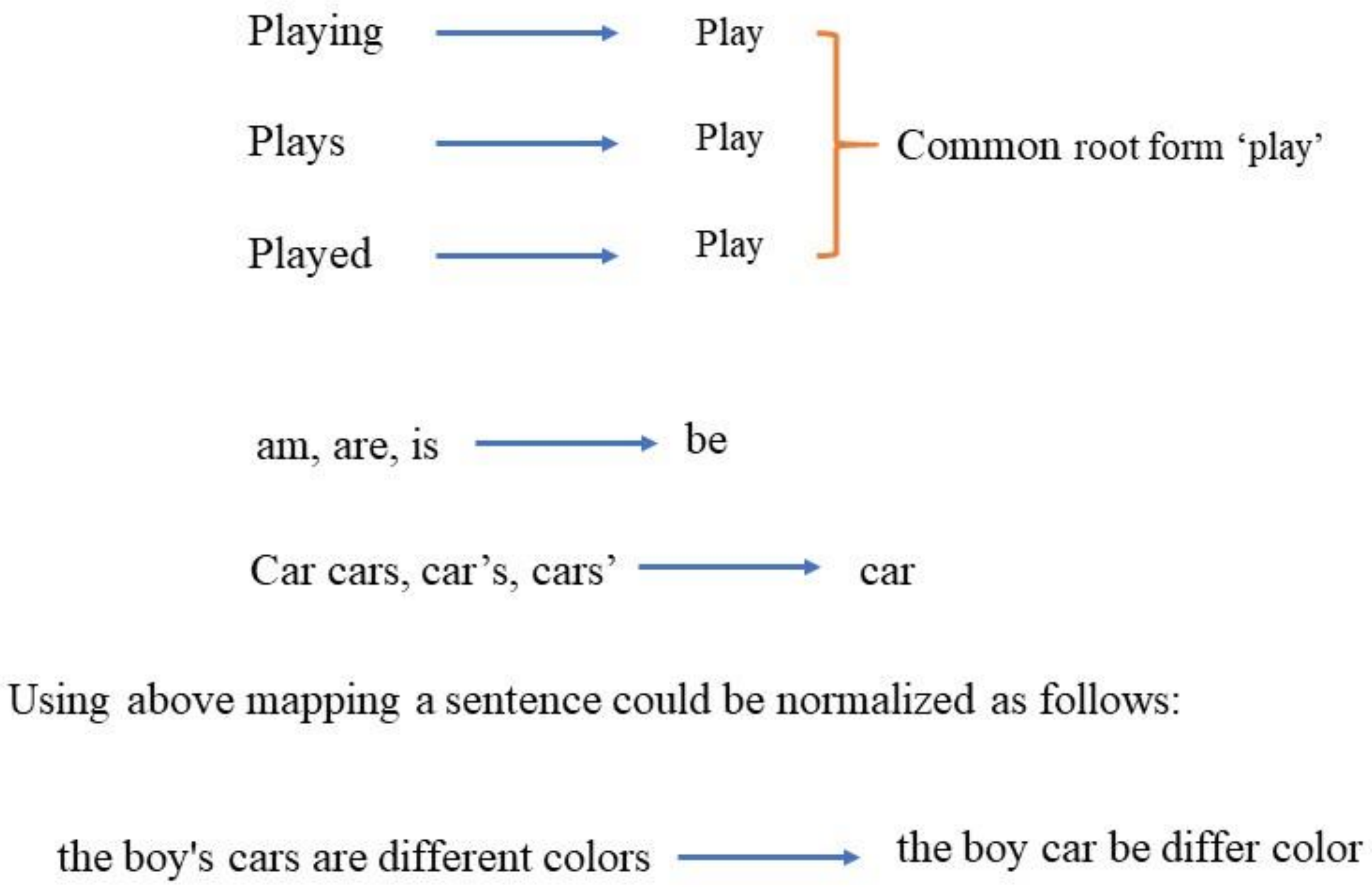
Text vs. Corpus

TEXT	**CORPUS**
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

(Tognini-Bonelli 2001: 3)

Lemmatization

A process of producing a proper root word that belongs to the language



Playing → Play

Plays → Play

Played → Play

Common root form ‘play’

am, are, is → be

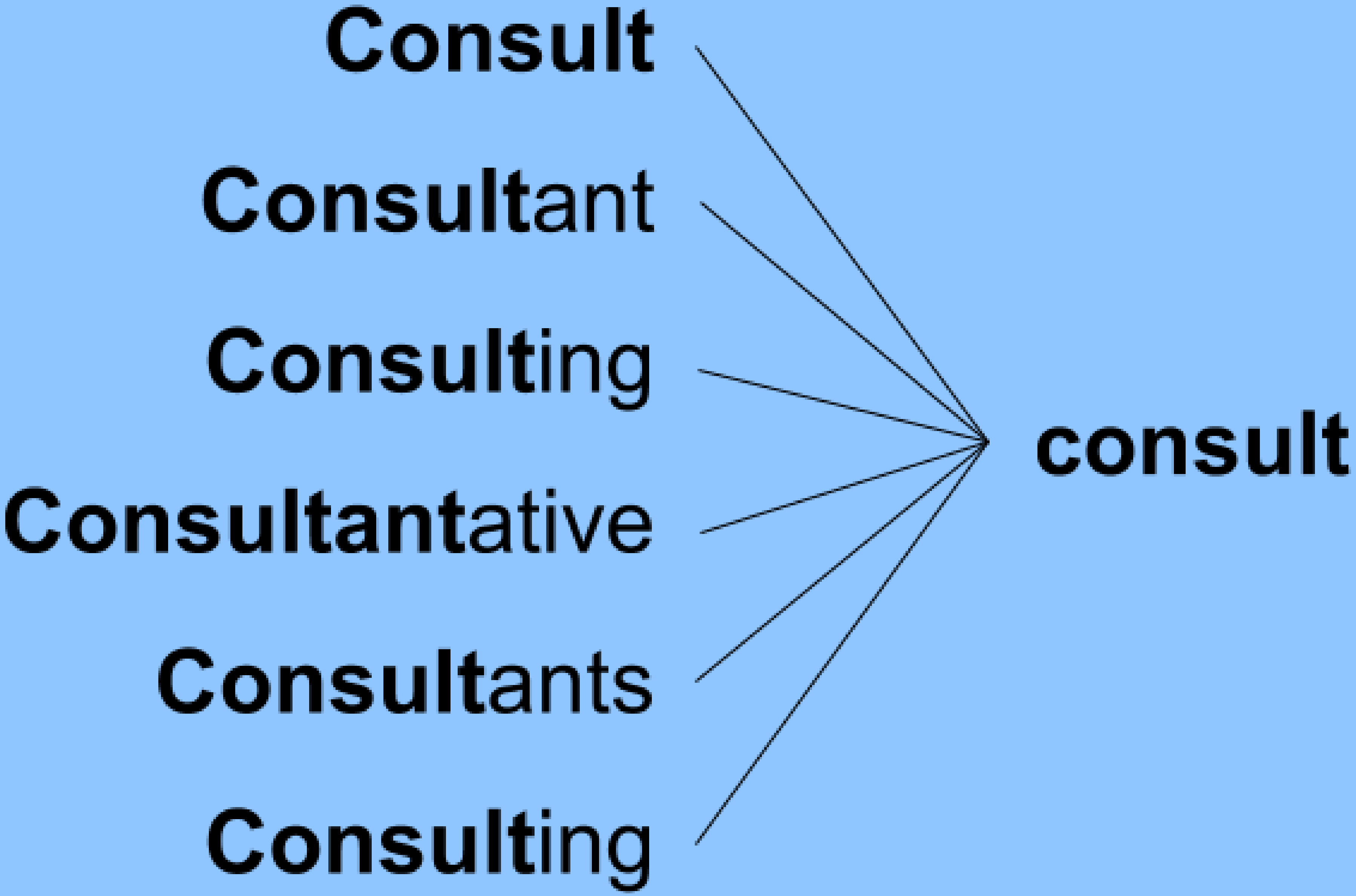
Car cars, car’s, cars’ → car

Using above mapping a sentence could be normalized as follows:

the boy's cars are different colors → the boy car be differ color

Stemming

A process that converts a word into its stem by keeping the base word and cutting off the affix



Consult

Consultant

Consulting

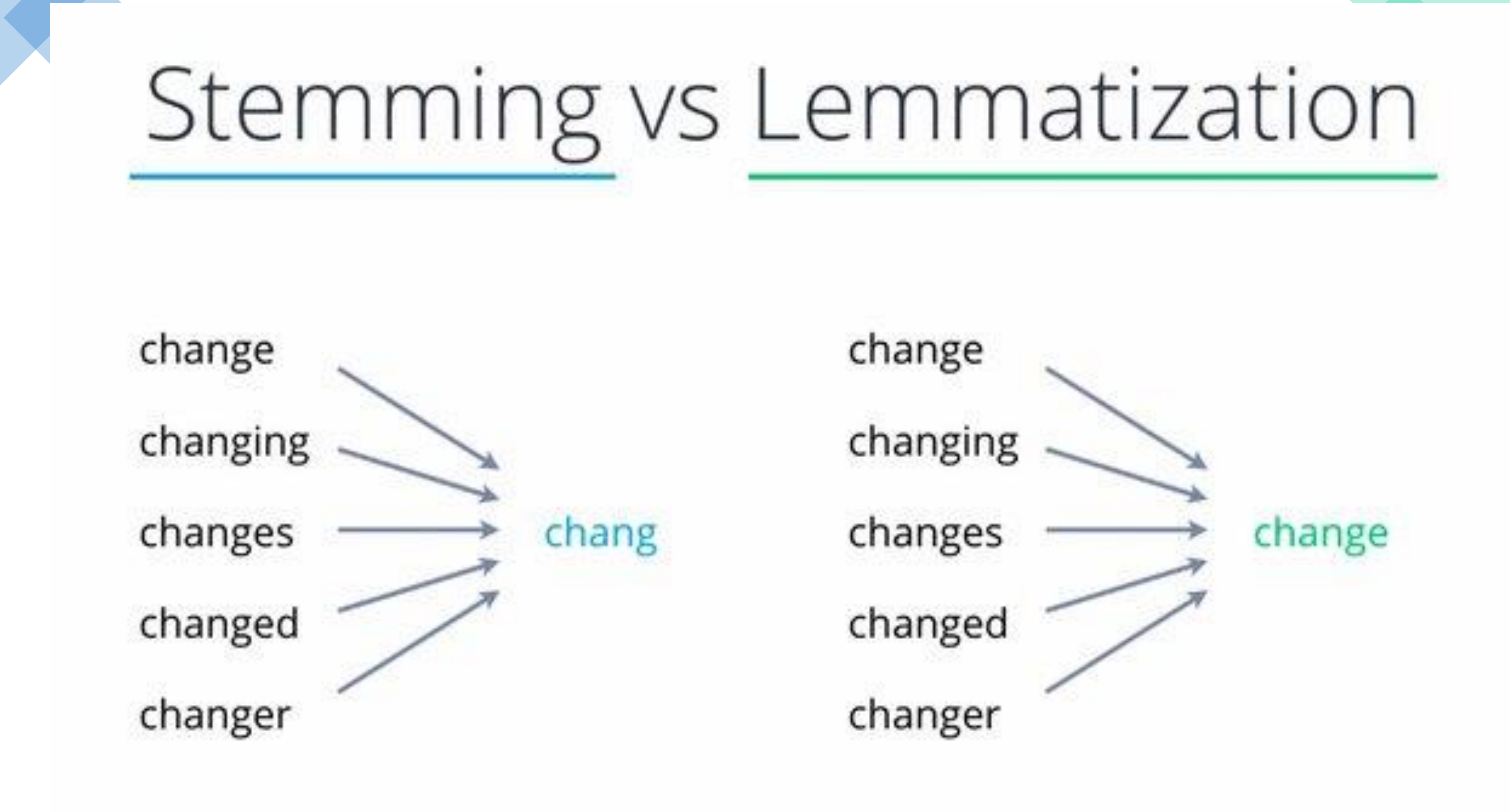
Consultantative

Consultants

Consulting

consult

Stemming vs Lemmatization



change
changing
changes
changed
changer

chang

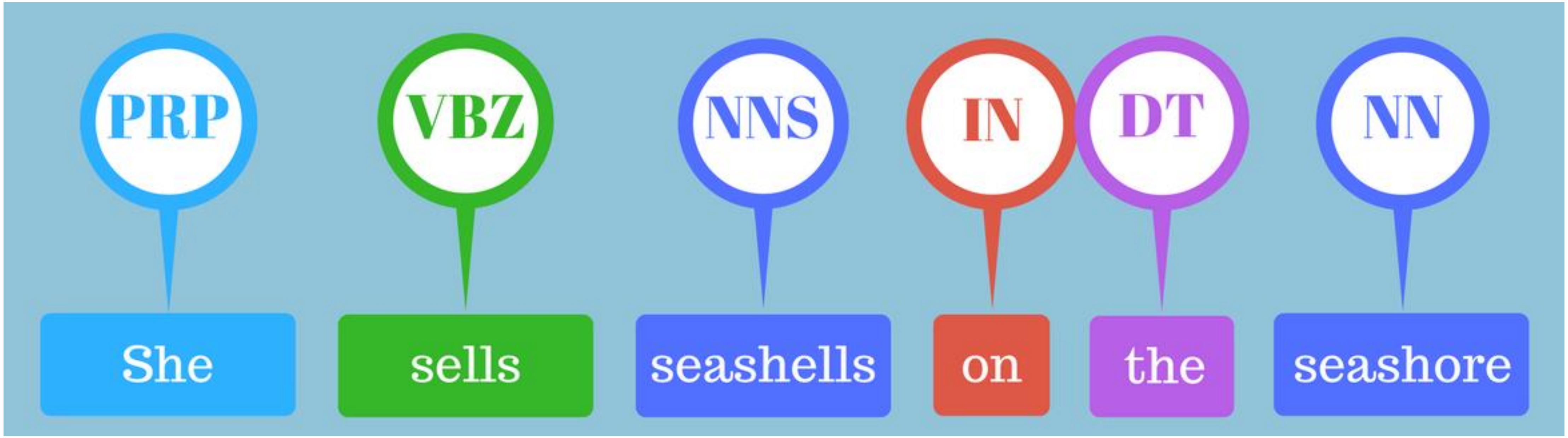
change
changing
changes
changed
changer

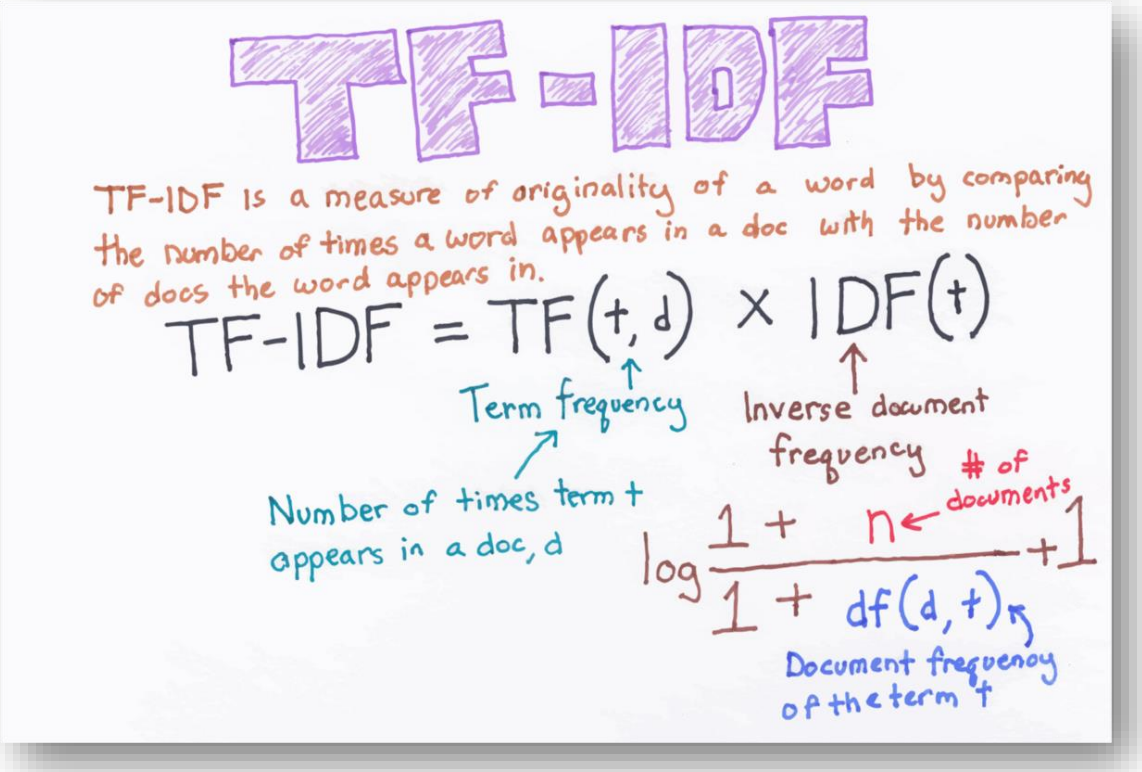
change

N-GRAM

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

POS Tagging





TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

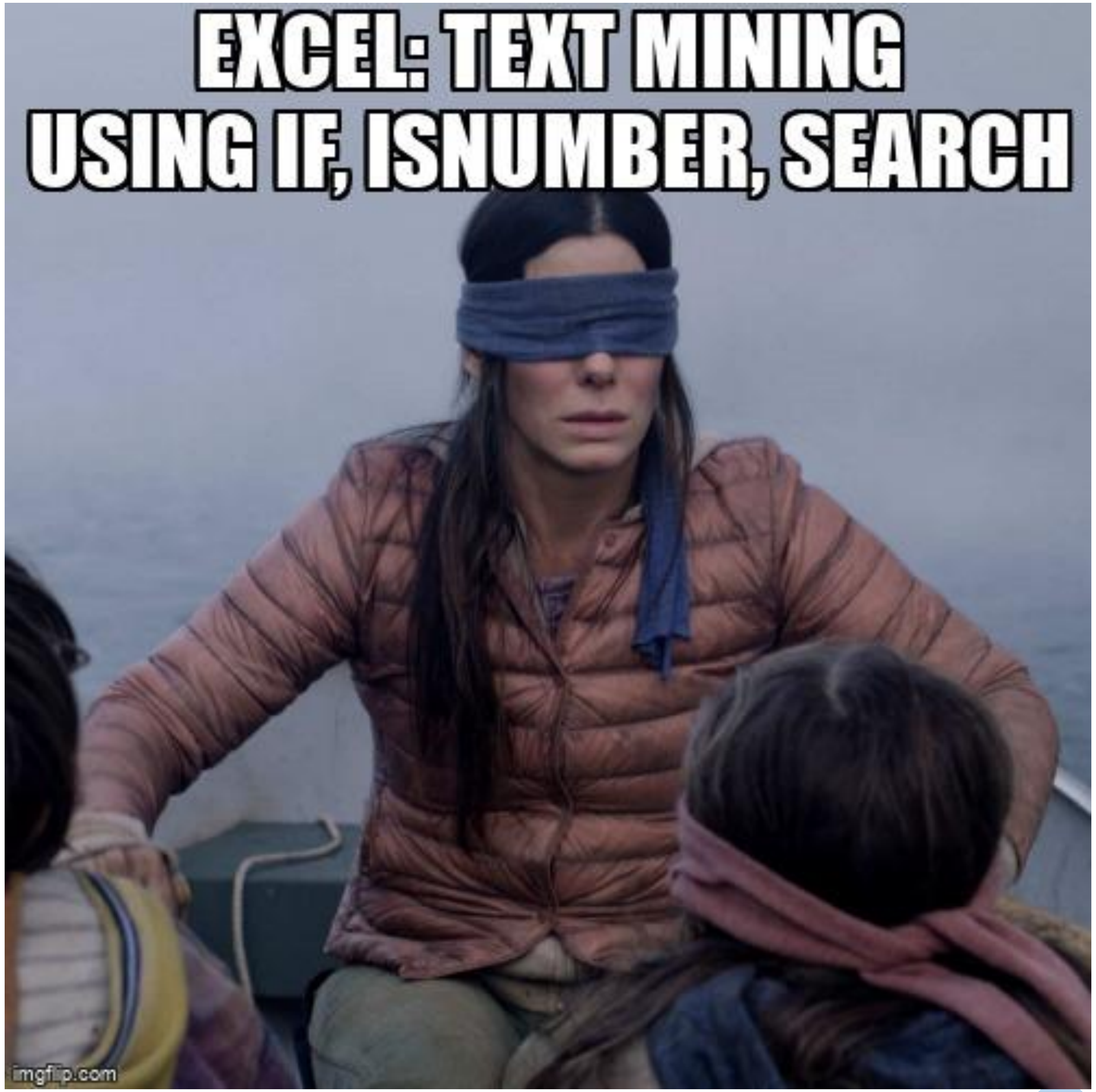
$$\log \frac{1 + n}{1 + df(d, t)} + 1$$

# of documents

Document frequency of the term t

Document	She	Loves	Food	With	Cheese	Her	Favourite	is	Italian	Lives	in	State
Doc 1	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0	0	0
Doc 2	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0
Doc 3	$\frac{1}{5}$	0	0	0	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

So how to you do all those fancy text mining stuff?


EXCEL: TEXT MINING
USING IF, ISNUMBER, SEARCH
imgflip.com

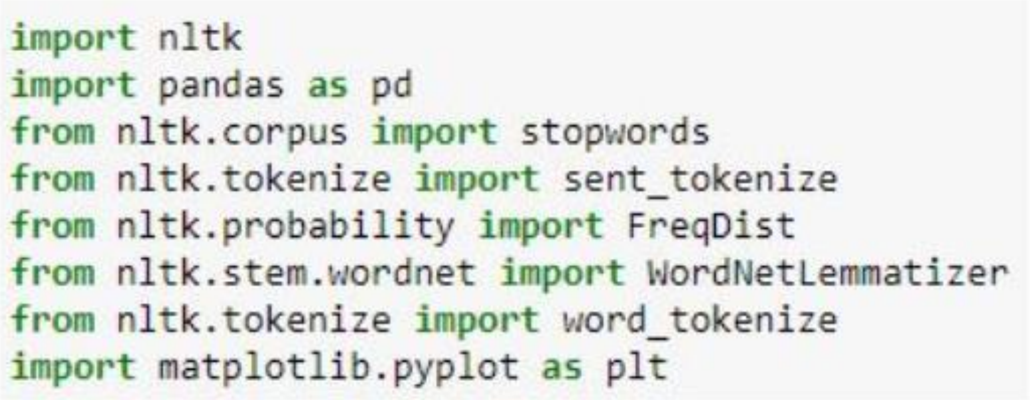
WHAT IF I TOLD YOU
THERE IS A BETTER WAY?

NATURAL LANGUAGE PROCESSING

USING

NLTK

PYTHON



```
import nltk
import pandas as pd
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize
from nltk.probability import FreqDist
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import matplotlib.pyplot as plt
```