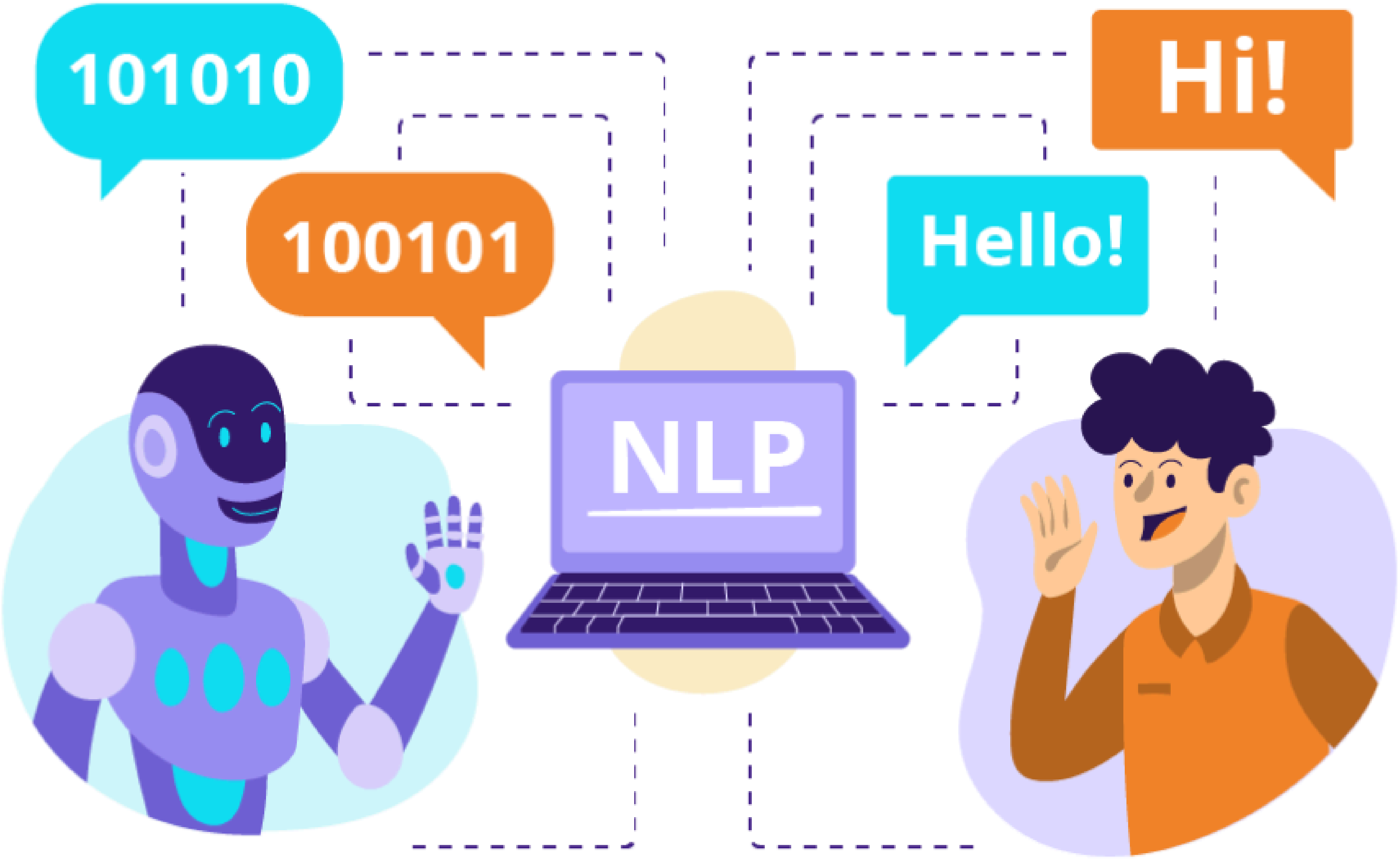
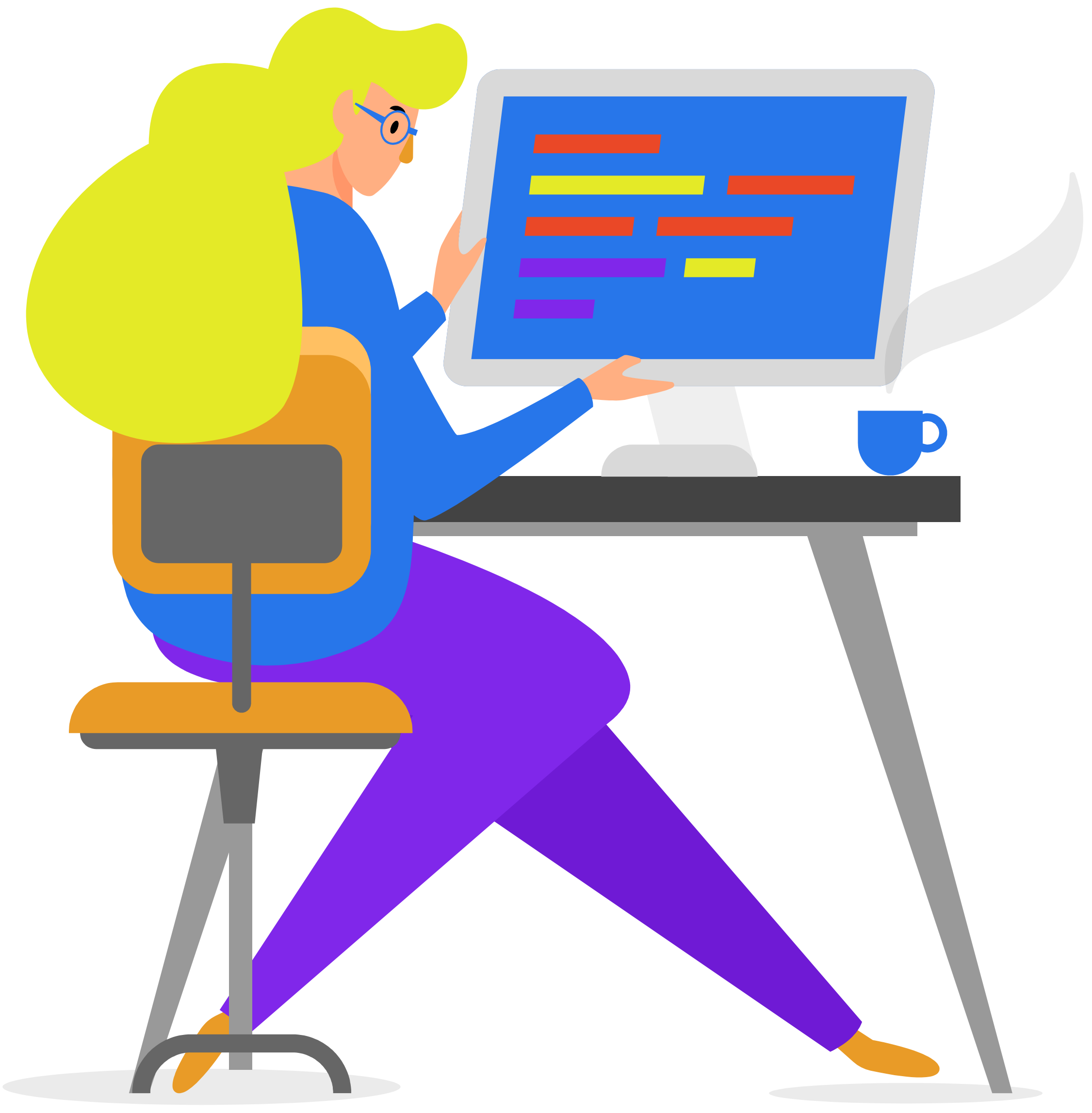


Intro to NLP

Summer Project:
Frames to Fables

What is NLP?



01 Utilizing Natural Language to Facilitate Human-Computer Interaction

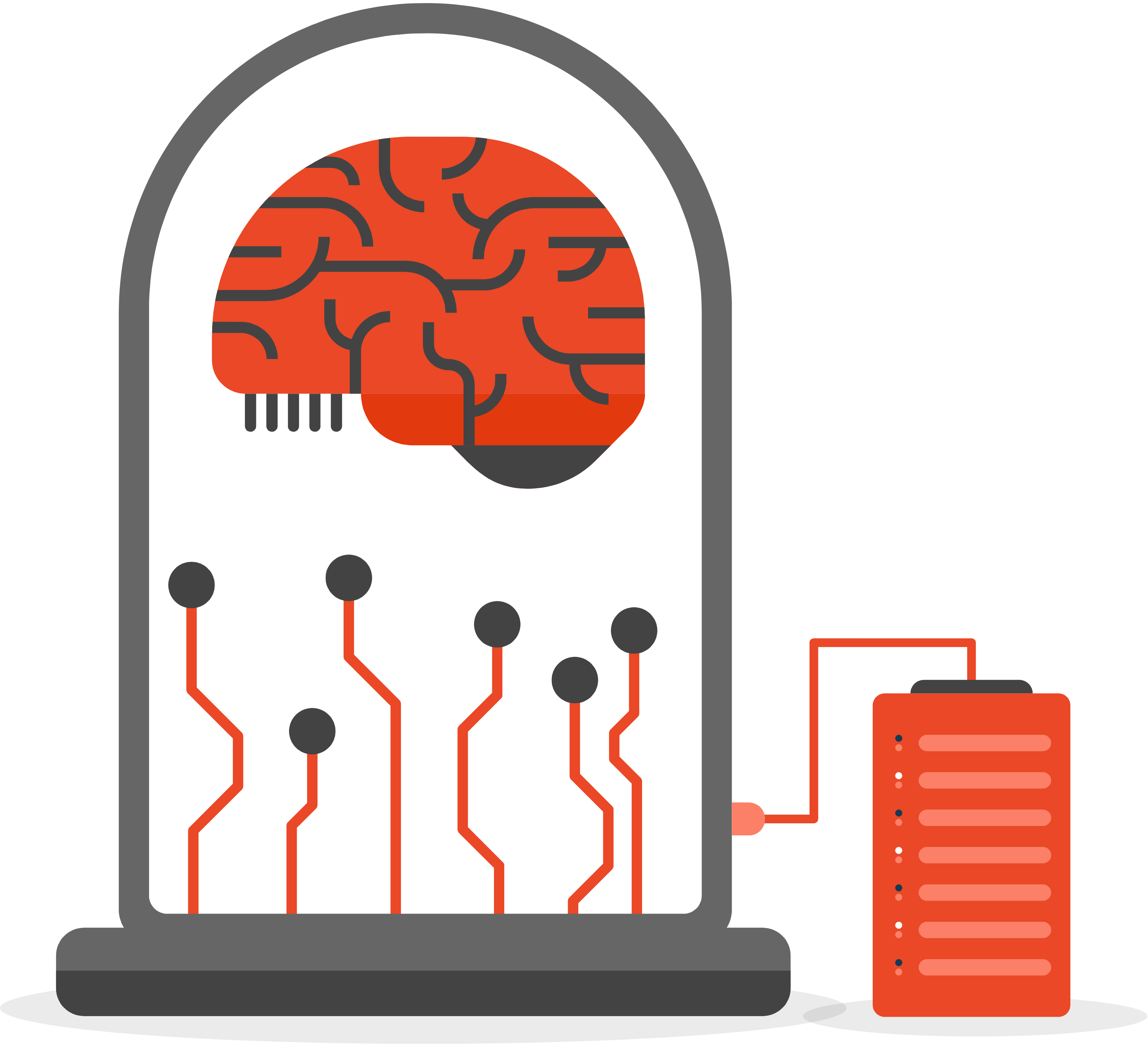
02 Make machines able to interpret, analyse and generate human like text

03 Text analysis by statistical methods, machine learning, deep learning etc

04 Language translation, speech recognition, text translation etc

Pre-Processing

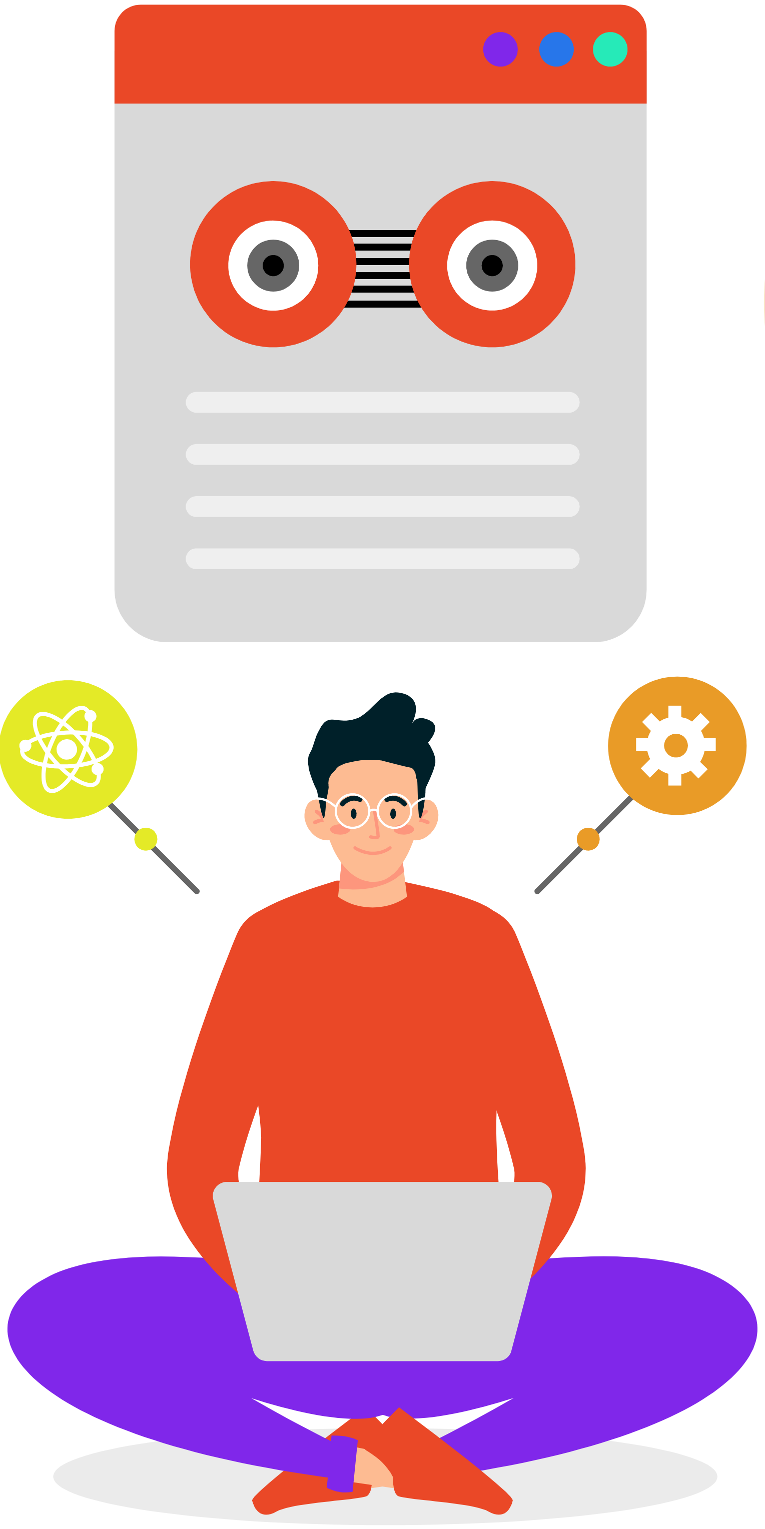
01	**What is pre-processing?**
	Cleaning and transforming raw data
02	**Why pre-processing?**
	Models can extract meaningful patterns and insights from data
03	**Some techniques**
	• Tokenisation • Stemming • Lemmatization • Stopwords Removal



Tokenisation

Tokenisation

Breaking text into small tokens. Mostly these are words or subwords



Your Turn!

Let’s see how well you learn!

Example

Sentence: I love IIT Kanpur
Output: [’I’, ‘love’, ‘IIT’, ‘Kanpur’]

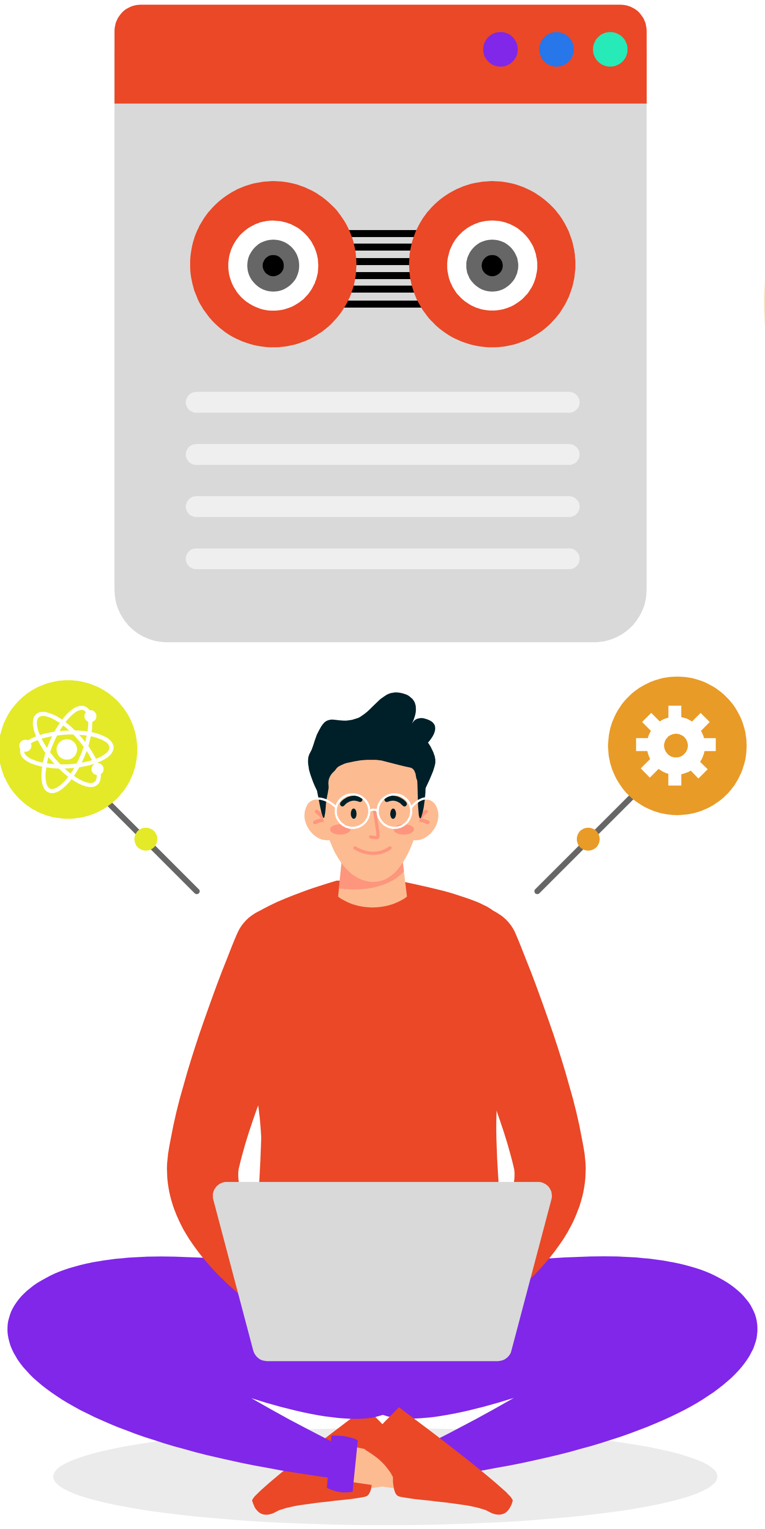
Example

Sentence: I can guess this right
Output: [’I’, ‘can’, ‘guess’, ‘this’, ‘right’]

Stopwords

Stopwords

Words carrying little to no semantic meaning
Example: 'The', 'And', 'Is'



Your Turn!

Ace it this time guys!!!

Example

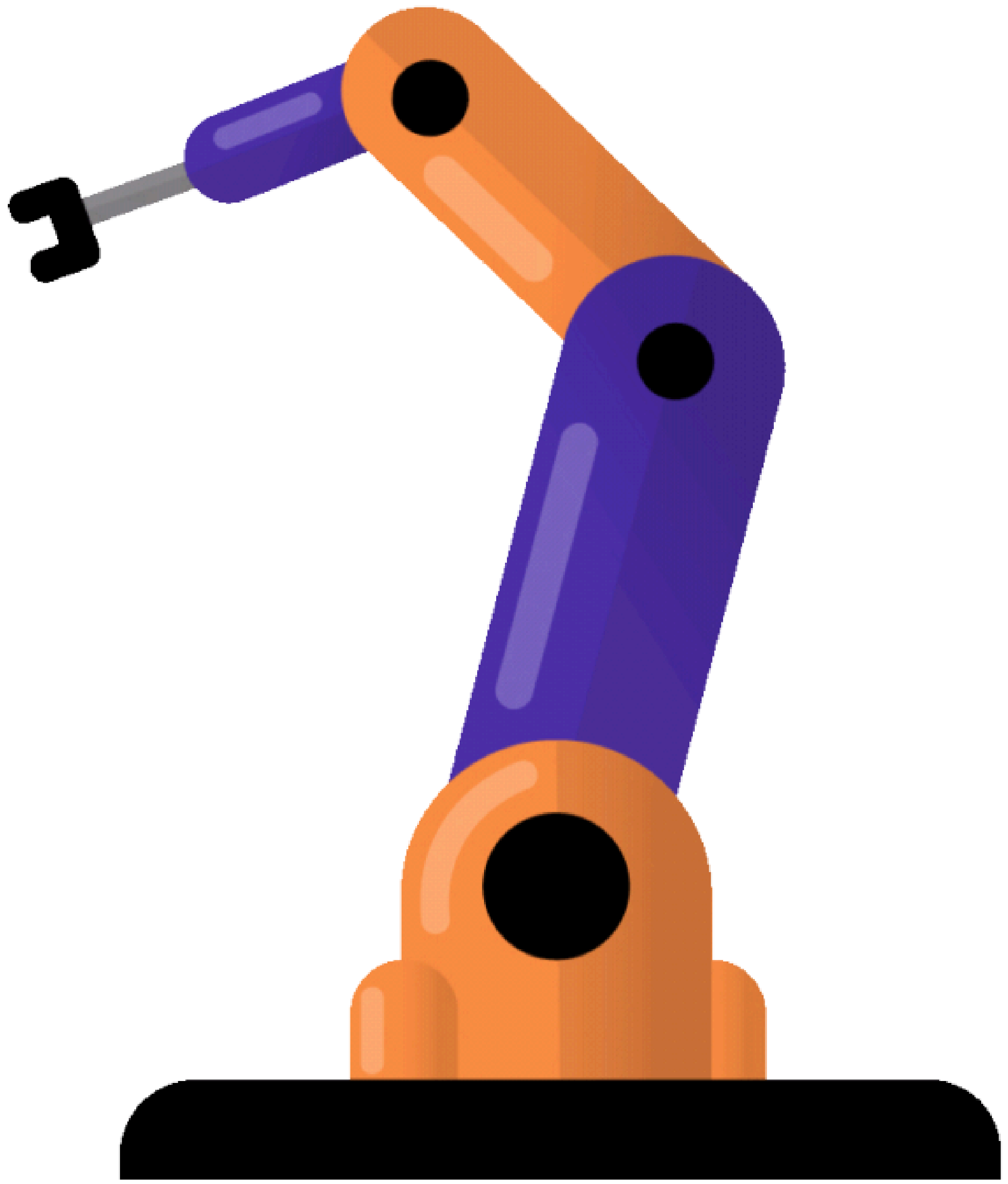
Sentence: The Sun is bright
Output: Sun bright

Example

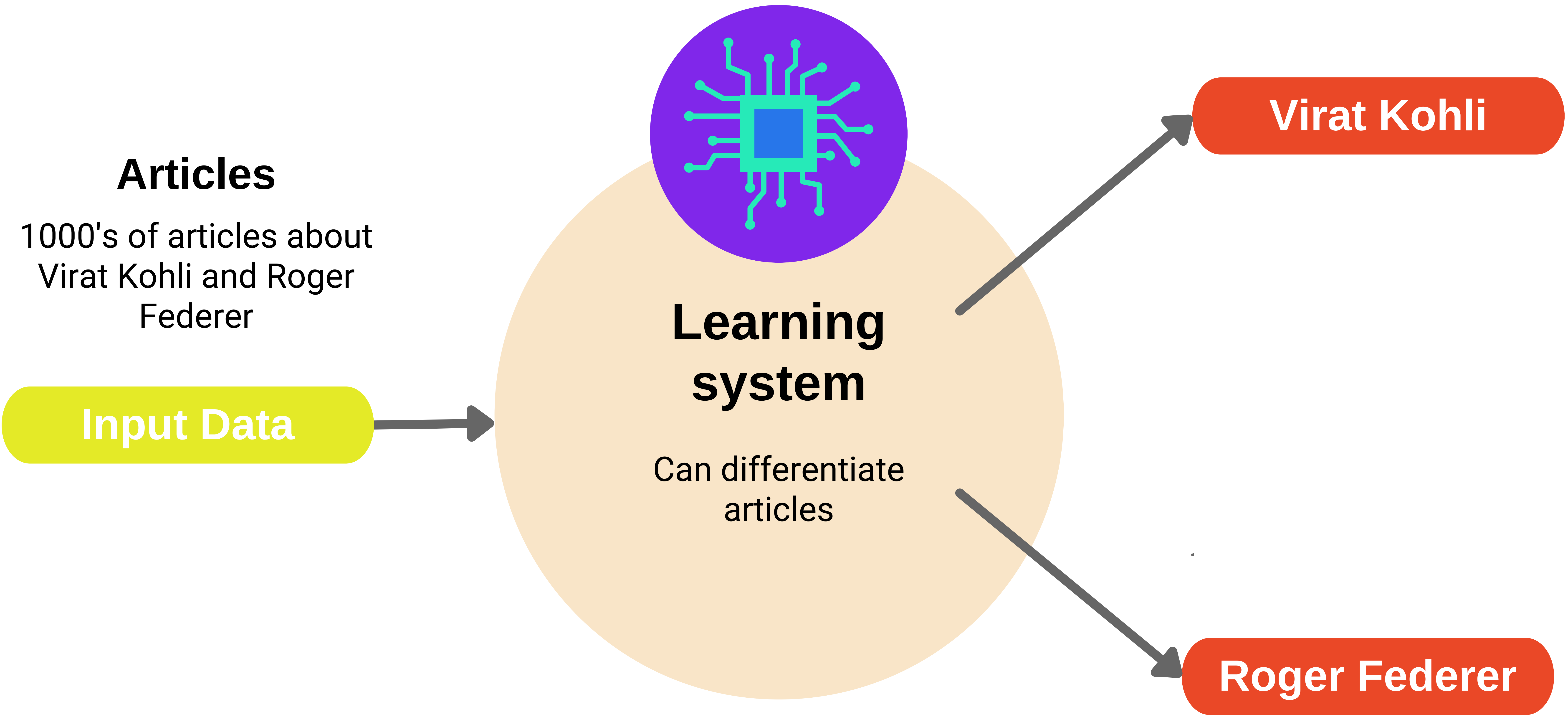
Sentence: It is easy to get 9+ CPI
Output: easy get 9 + CPI

Feature Engineering

01	**What is feature engineering?**
	Selecting, transforming or creating relevant features (attributes or variables) from raw data
02	**Use in NLP?**
	Representing text data in a way that captures its essential characteristics
03	**Some techniques**
	• BoW (Bag of Words): A numerical representation of text data using frequency • N-grams: Sequence of ‘n’ words • TF-IDF: A smarter version of BoW



Another Classifier Model



Articles

1000's of articles about Virat Kohli and Roger Federer

Input Data

Learning system

Can differentiate articles

Virat Kohli

Roger Federer

Bag of Words

01

Vocabulary

List down all the unique words

02

Frequency

Frequency of each word

03

Representation

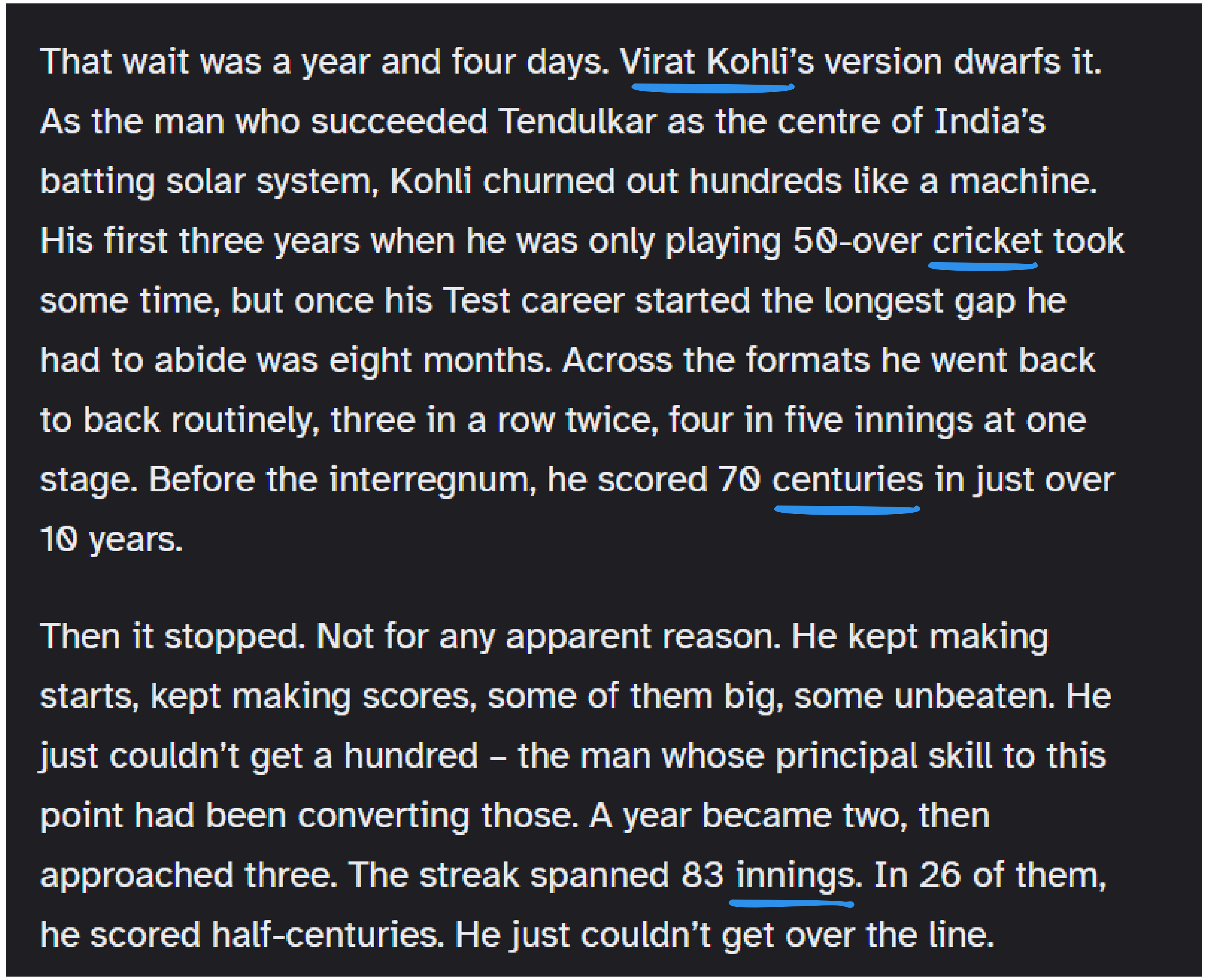
Store in a table format for insights

04

Feed the model

Feed this data to the model

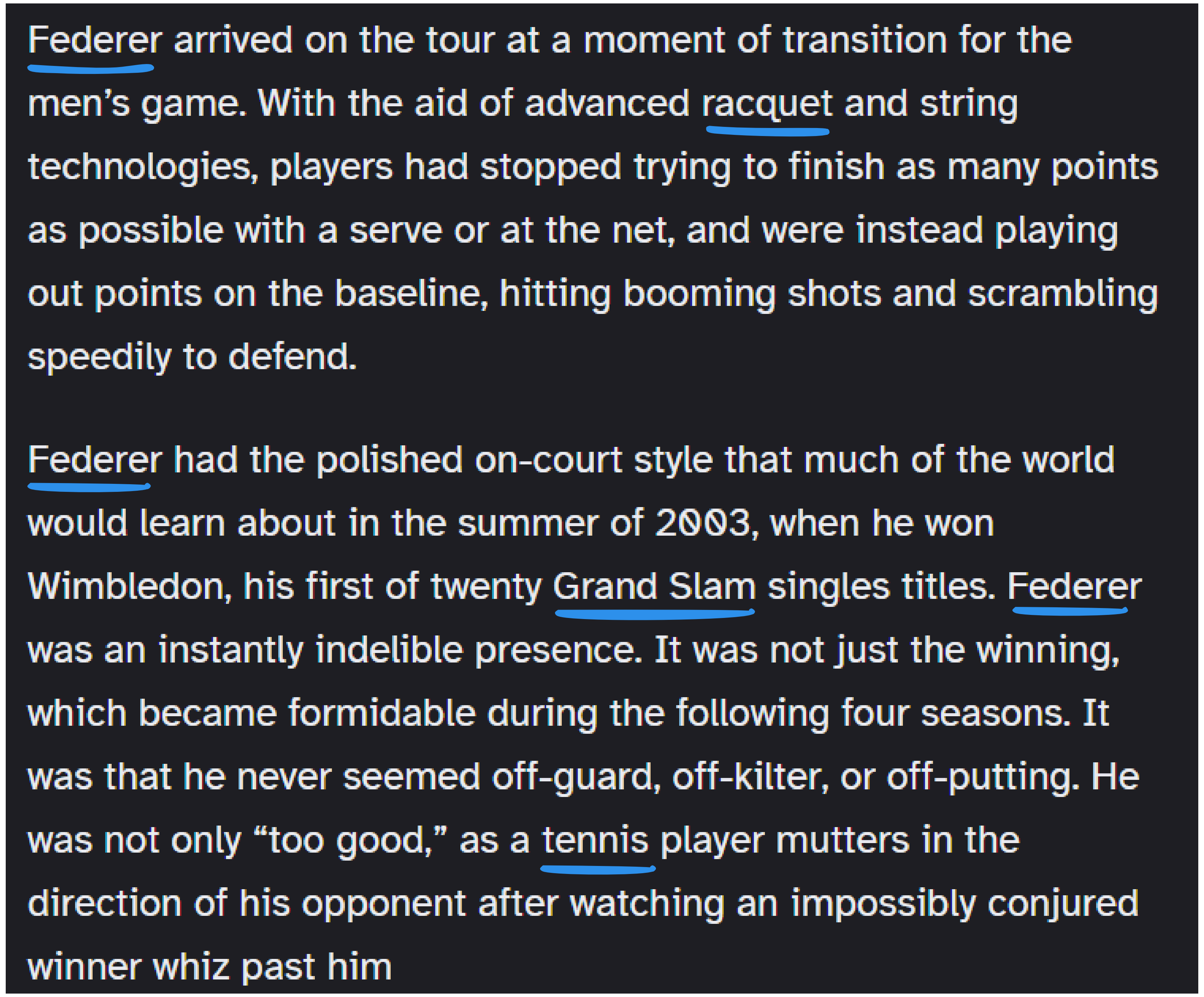
Virat Kohli



That wait was a year and four days. Virat Kohli's version dwarfs it. As the man who succeeded Tendulkar as the centre of India's batting solar system, Kohli churned out hundreds like a machine. His first three years when he was only playing 50-over cricket took some time, but once his Test career started the longest gap he had to abide was eight months. Across the formats he went back to back routinely, three in a row twice, four in five innings at one stage. Before the interregnum, he scored 70 centuries in just over 10 years.

Then it stopped. Not for any apparent reason. He kept making starts, kept making scores, some of them big, some unbeaten. He just couldn't get a hundred – the man whose principal skill to this point had been converting those. A year became two, then approached three. The streak spanned 83 innings. In 26 of them, he scored half-centuries. He just couldn't get over the line.

Roger Federer



Federer arrived on the tour at a moment of transition for the men's game. With the aid of advanced racquet and string technologies, players had stopped trying to finish as many points as possible with a serve or at the net, and were instead playing out points on the baseline, hitting booming shots and scrambling speedily to defend.

Federer had the polished on-court style that much of the world would learn about in the summer of 2003, when he won Wimbledon, his first of twenty Grand Slam singles titles. Federer was an instantly indelible presence. It was not just the winning, which became formidable during the following four seasons. It was that he never seemed off-guard, off-kilter, or off-putting. He was not only "too good," as a tennis player mutters in the direction of his opponent after watching an impossibly conjured winner whiz past him

	Cricket	Tennis	Virat	Kohli	Innings	Federer	Racquet	Century	Grandslam	........}
Article 1	1	0	1	1	1	0	0	1	0	.....
Article 2	0	1	0	0	0	3	1	0	1	....

Now it’s very easy for the model

Virat kohli article : [1 0 1 1 1 0 0 1 0 ...]

Roger Federer article : [0 1 0 0 0 3 1 0 1 ...]

	Cricket	Tennis	Virat	Kohli	Innings	Federer	Racquet	Century	Grandslam	........}
Article 1	1	0	1	1	1	0	0	1	0	.....
Article 2	0	1	0	0	0	3	1	0	1	....

Now it’s very easy for the model

Virat kohli article : [1 0 1 1 1 0 0 1 0 ...]

Roger Federer article : [0 1 0 0 0 3 1 0 1 ...]

N-Grams

01

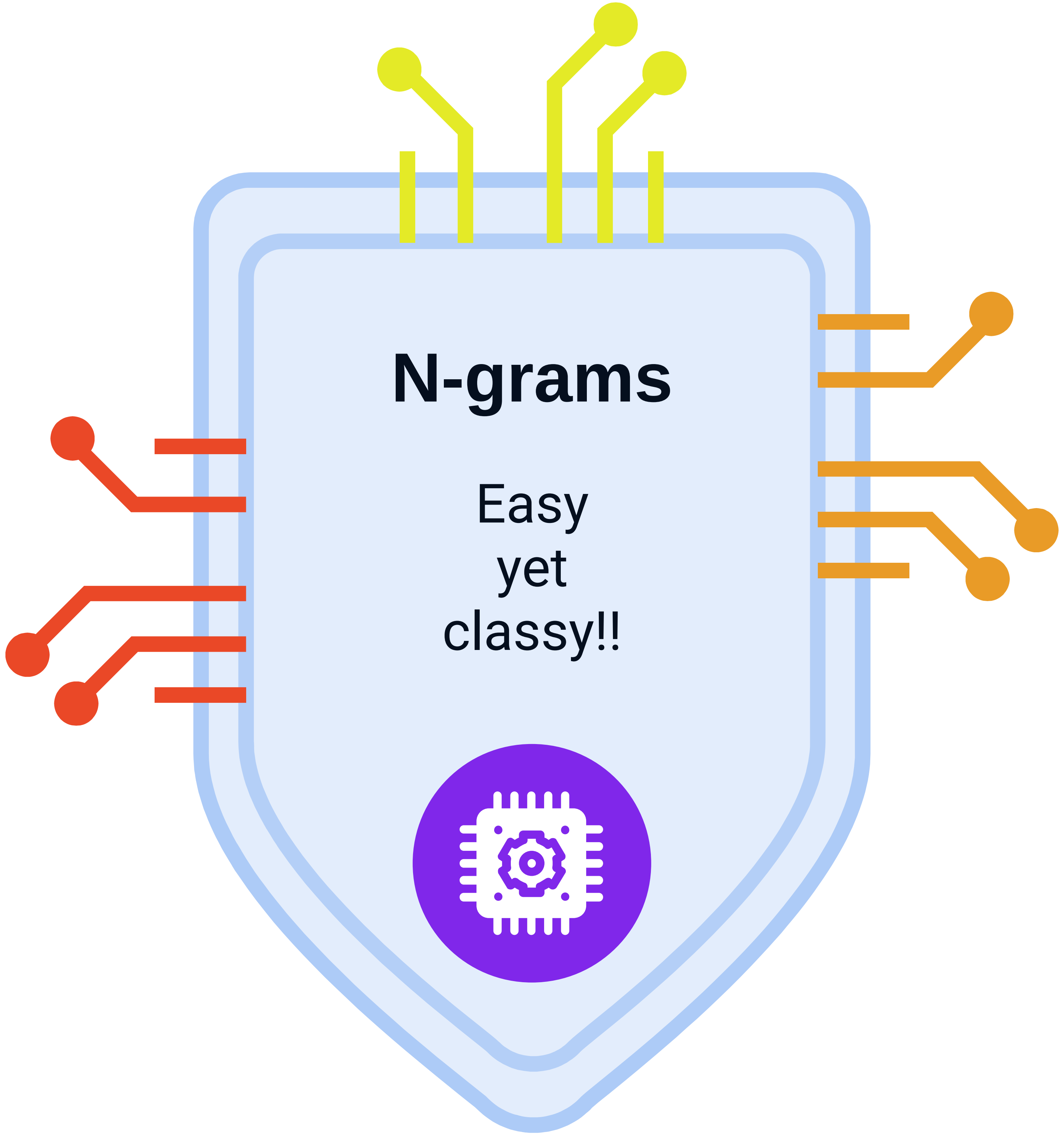
Sequence

A sequence of ‘n’ words

03

Order of words

In language, order of words is important



N-grams

Easy yet classy!!

02

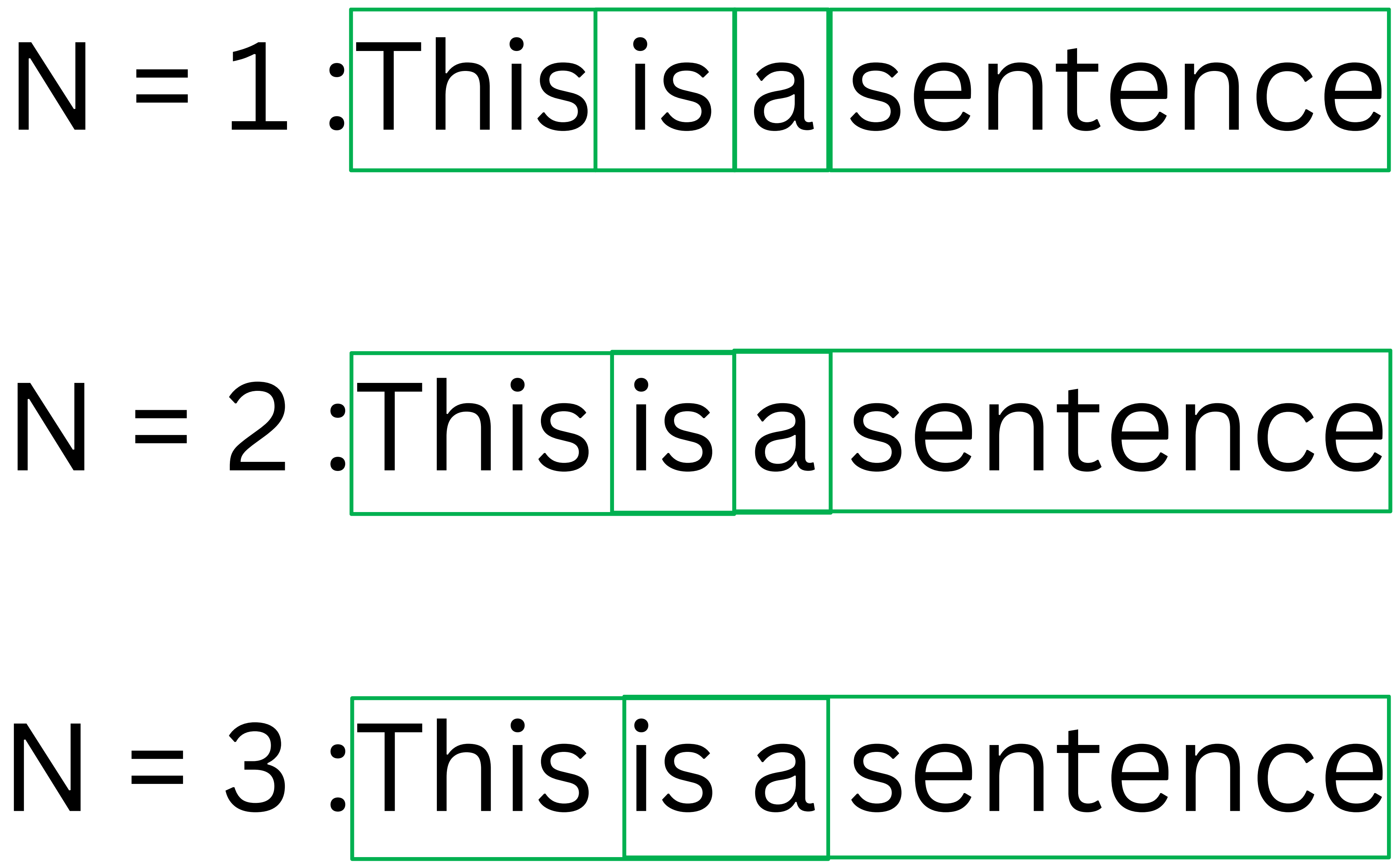
Local context

Relationship between adjacent words

04

Examples

Unigram/BoW: [’is’, ‘it’]
Bigram: [’Is it’, ‘it good’]



N = 1 : This is a sentence

Unigram

this,
is,
a,
sentence

N = 2 : This is a sentence

Bigram

this is,
is a,
a sentence

N = 3 : This is a sentence

Trigram

this is a,
is a sentence

$$\text{TF}(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$\text{IDF}(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

The TF-IDF score is the product of TF and IDF:

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

For Document 1:

$$\textbf{TF-IDF}\ (\text{cat, Document 1, D}) = 0.167 \times 0.176 \approx 0.029$$

For Document 2:

$$\textbf{TF-IDF}(\text{cat, Document 2, D}) = 0 \times 0.176 = 0$$