NATURAL LANGUAGE PROCESSING

By :- SHOBHIT TYAGI

You will Learn…

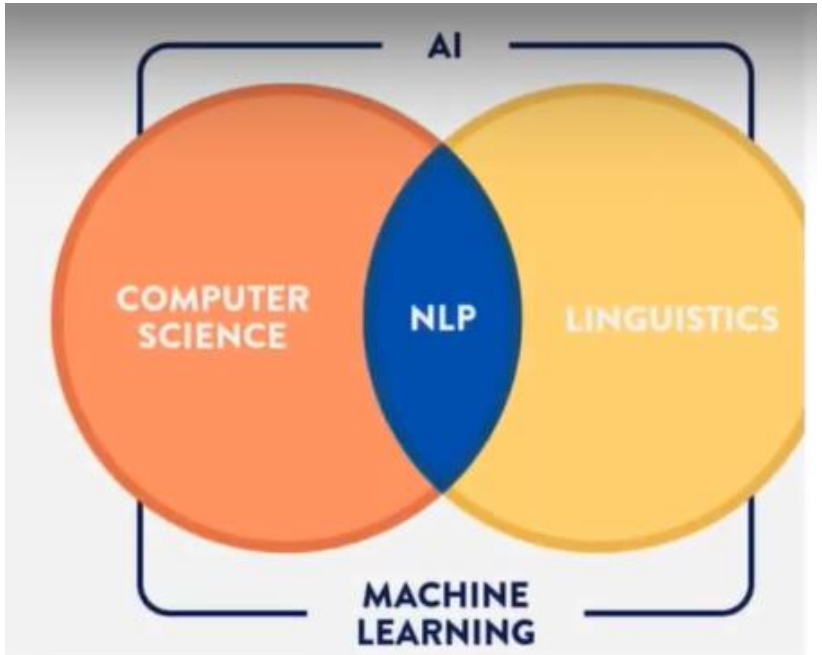
- What is NLP?
- Why NLP is important?
- Applications of NLP

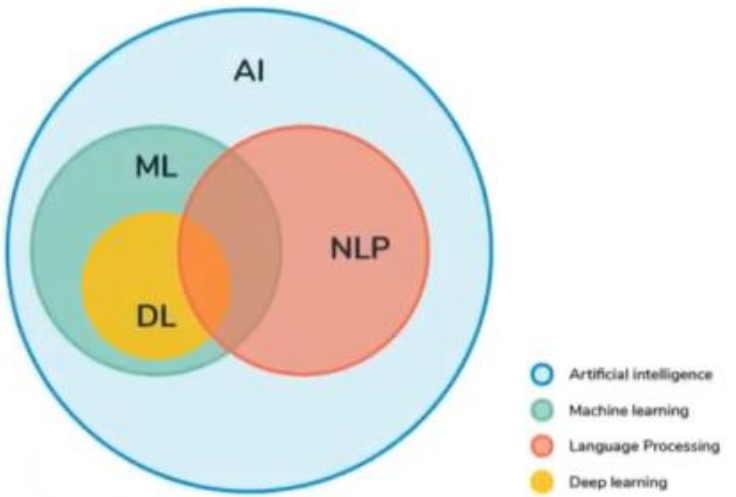
NATURAL LANGUAGE PROCESSING - NLP

Interactions between computers and human natural languages

NATURAL LANGUAGE PROCESSING - NLP

Natural language processing is a form of artificial intelligence (AI) that gives computers the ability to read, understand and interpret human language.





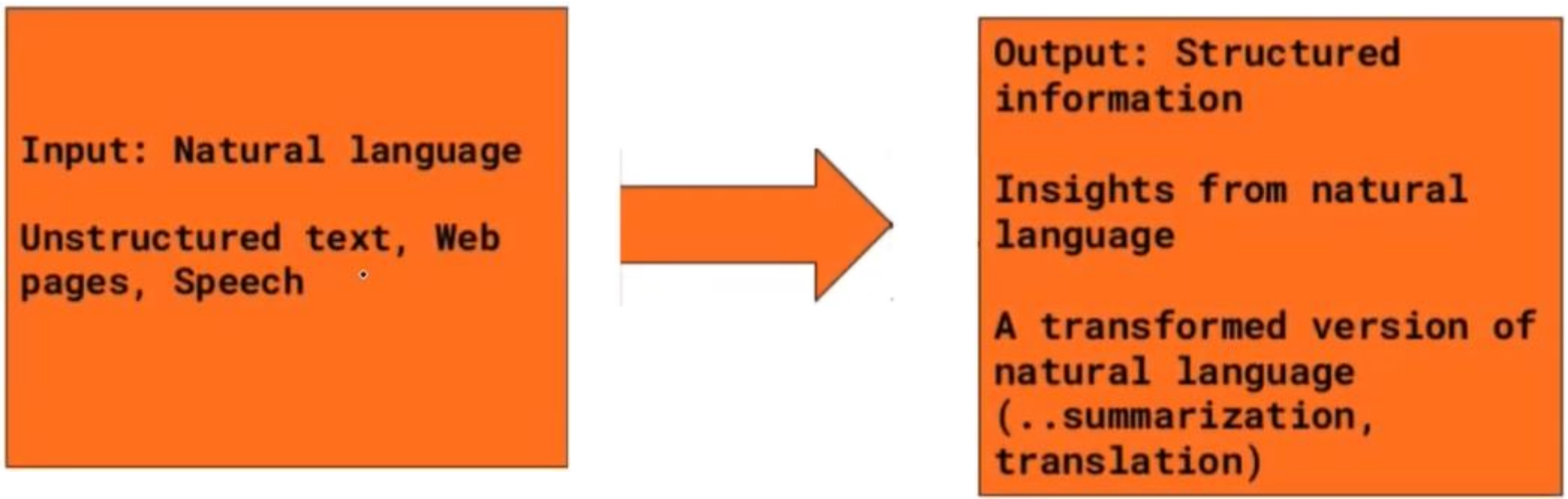
Why NLP is Important?

NLP used to handle Human Text/Language data generated from various Data Sources.

Handling Large volumes of textual data

Why NLP is Important?

Structuring the highly unstructured data



Why NLP is Important?

News:

AN EARTHQUAKE struck Indonesia today - a strapping 7.7 magnitude earthquake that struck early today off the northern coast of the island of Sumatra. It caused minor damage and there are no reports of any deaths, although electricity was interrupted in several places.

Location : Indonesia
Magnitude: 7.7
Region: Sumatra (Northern Cost)
Deaths: Nil
Damage: Minor

Tweet

@nokia announces release of new PDA phones see is.gd/iuTuY

Who: Nokia

What: Product announcement

APPLICATIONS OF NLP

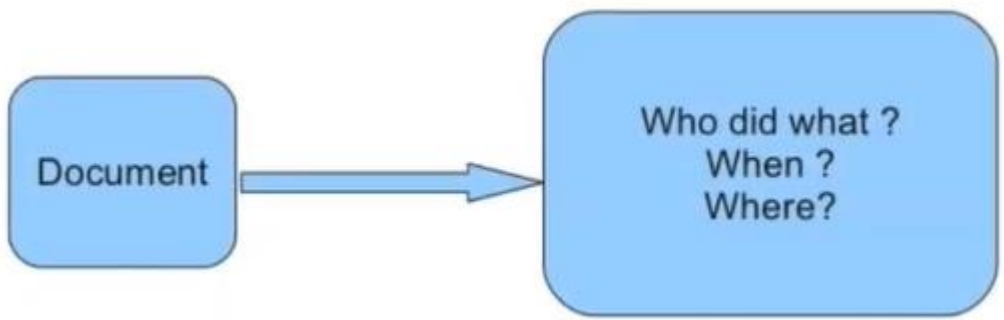
You will Learn…

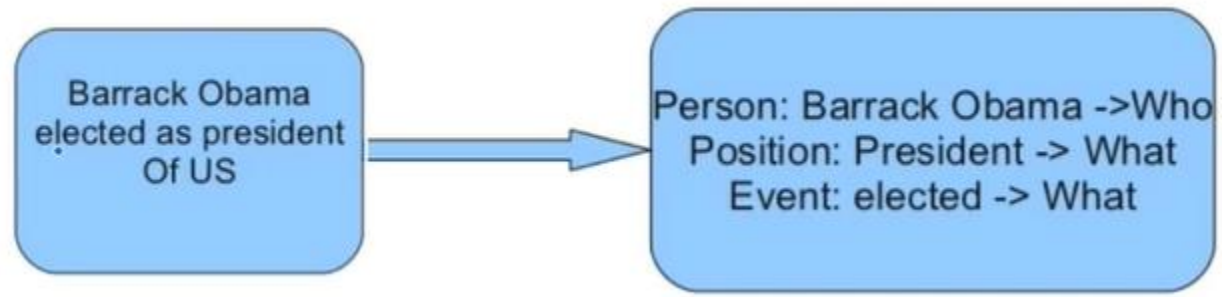
- **Applications of NLP**
- **Examples**

Application of NLP

Information Extraction

Extraction of Meaningful Information from text.





Application of NLP

Machine Translation
Automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language.

Application of NLP

Question answering (QA)

Systems that automatically answer questions posed by humans in a natural language.

When was Tesla born?

Application of NLP

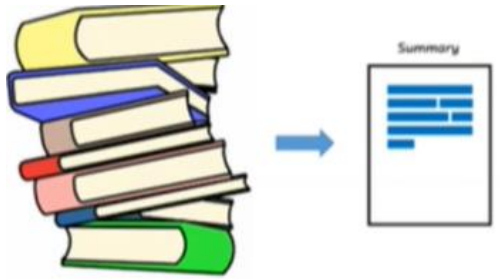
SENTIMENT

INPUT TEXT

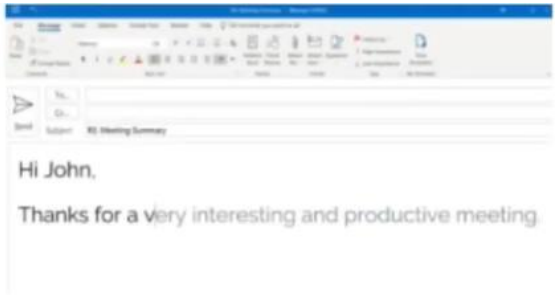
RESULT

CLEAR

SUBMIT



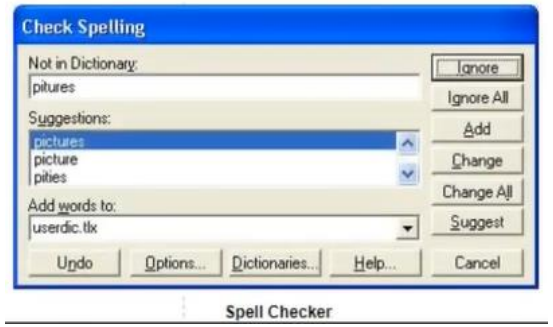
Text Summarization



Predictive Next Word...

Text Classification

Chatbots



Spell Checker

More Applications of NLP

TOOLS IN NLP

You will Learn…

- **NLP Tools**
- **Framework each tool belong**
- **Task perform in each tool**

Tools in NLP

Research And Development in NLP

WordNet, Stop words, Tokenization, Sentiment Analysis

Tools in NLP

Library for advanced
. NLP

Syntactic Parser, Named
NER, Tokenization, Fast
Processing, Visualization

Tools in NLP

NLP Research library, developing deep learning models

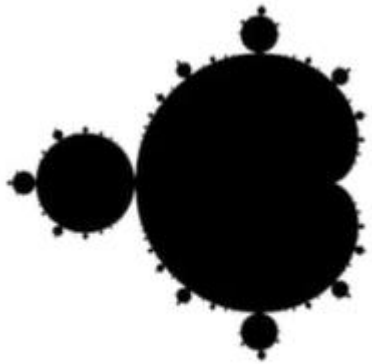
Question and Answering, Semantic Role Labeling, Within Document Co-reference, Textual Entailment, Text to SQL

Tools in NLP

Provide scalability for processing large amounts of data and performing complex operations.

Data Scrapping
Social Media Analysis
Conversational Chatbots

Tools in NLP



TextBlob

flair

Genism

NLP PIPELINE

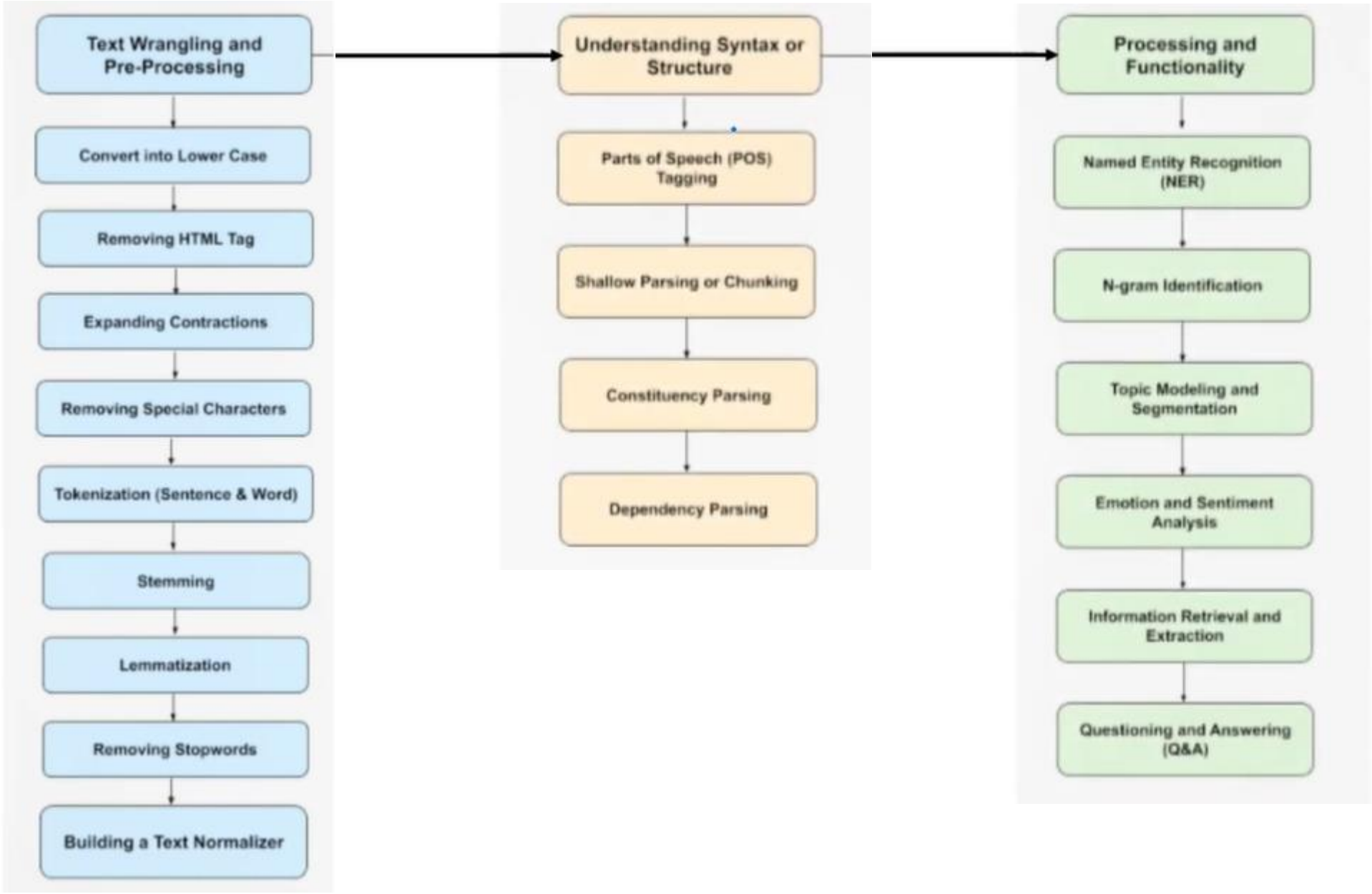
You will Learn…

- **NLP Pipeline**
- **Role of Pipeline in NLP**
- **Pipeline Components**
- **Building Pipeline**

How can we make use of NLP tools?

What are the steps in NLP to transform the text?

NLP PIPELINES



Building NLP Pipeline

Input Text

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

(Source: Wikipedia article "London")

Final Output Text

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium. London's ancient core, the City of London, largely retains its 1.12-square-mile (2.9 km2) medieval boundaries.

Building NLP Pipeline

Input Text

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

(Source: Wikipedia article "London")

Sentence Segmentation

1. "London is the capital and most populous city of England and the United Kingdom."
2. "Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia."
3. "It was founded by the Romans, who named it Londinium."

Building NLP Pipeline

Input Text

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

(Source: Wikipedia article "London")

Word Tokenization

"London is the capital and most populous city of England and the United Kingdom."

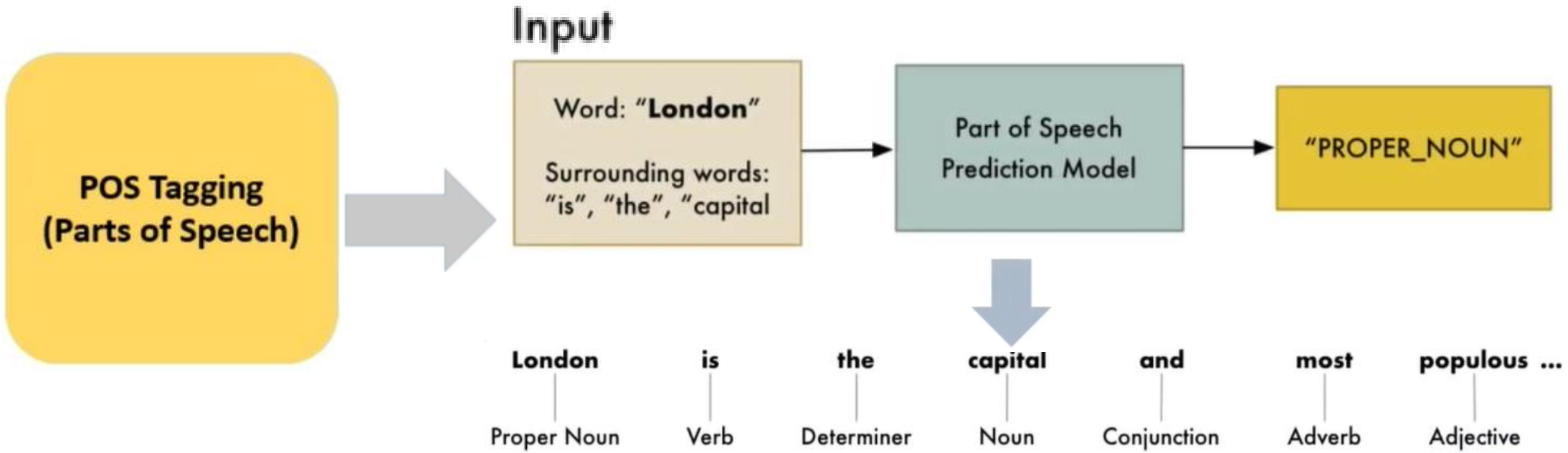
"London", "is", " the", "capital", "and", "most", "populous", "city", "of", "England", "and", "the", "United", "Kingdom", "."

Building NLP Pipeline

Input Text

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

(Source: Wikipedia article "London")

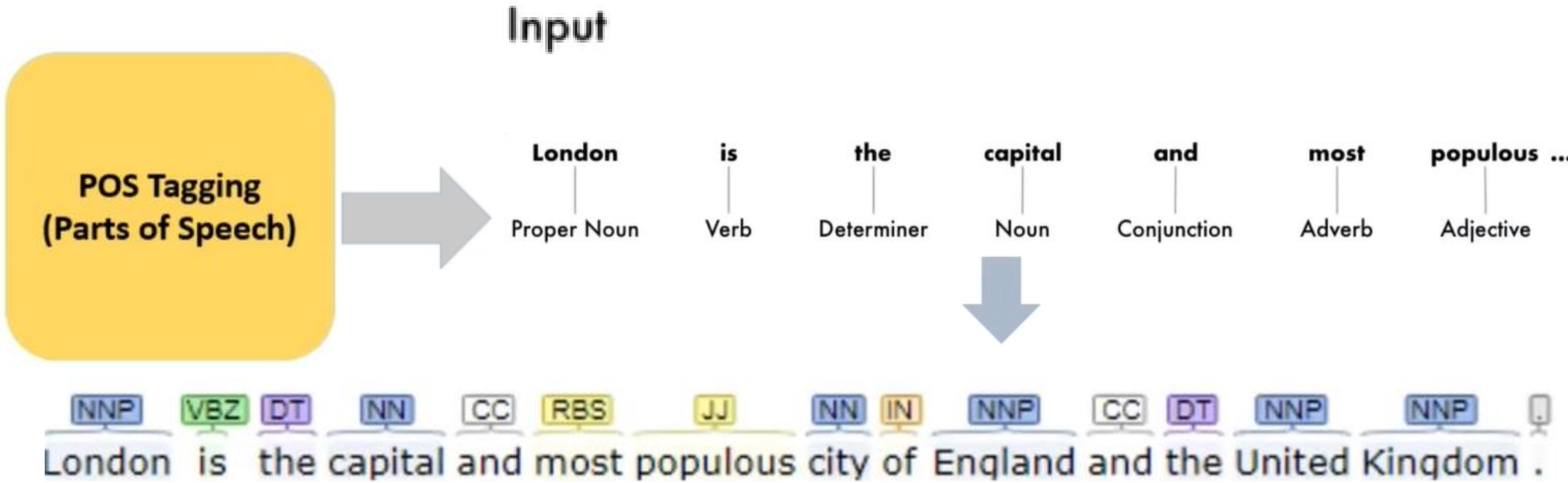


Building NLP Pipeline

Input Text

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

(Source: Wikipedia article "London")



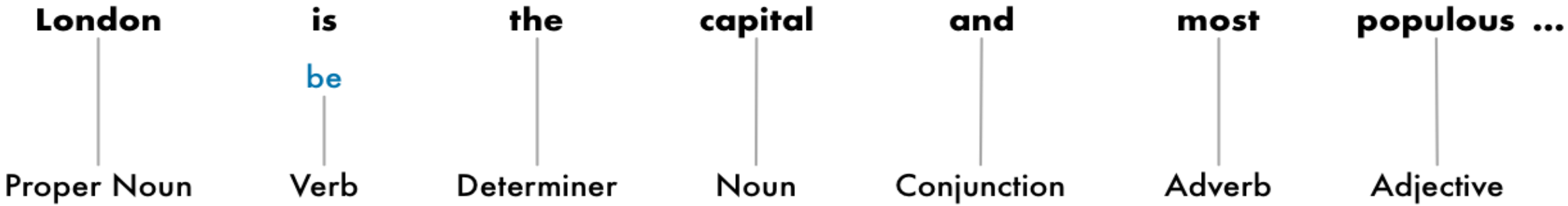
NLP: Example

- *London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.*

Text Lemmatization / Stemming

- Stemming and Lemmatization are **Text Normalization** techniques in NLP
- It is the process of reducing inflection in words to their root forms
- **For example:**
  - I had a **pony**.
  - I had two **ponies**.

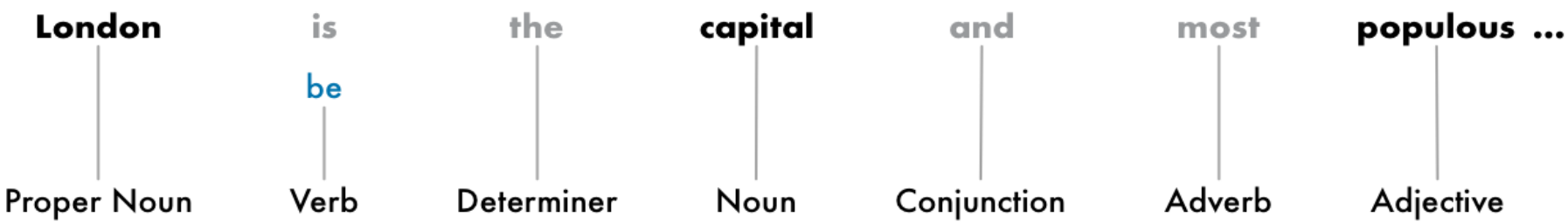
} Pony



Identifying Stop Words

- A **stop word** is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore.

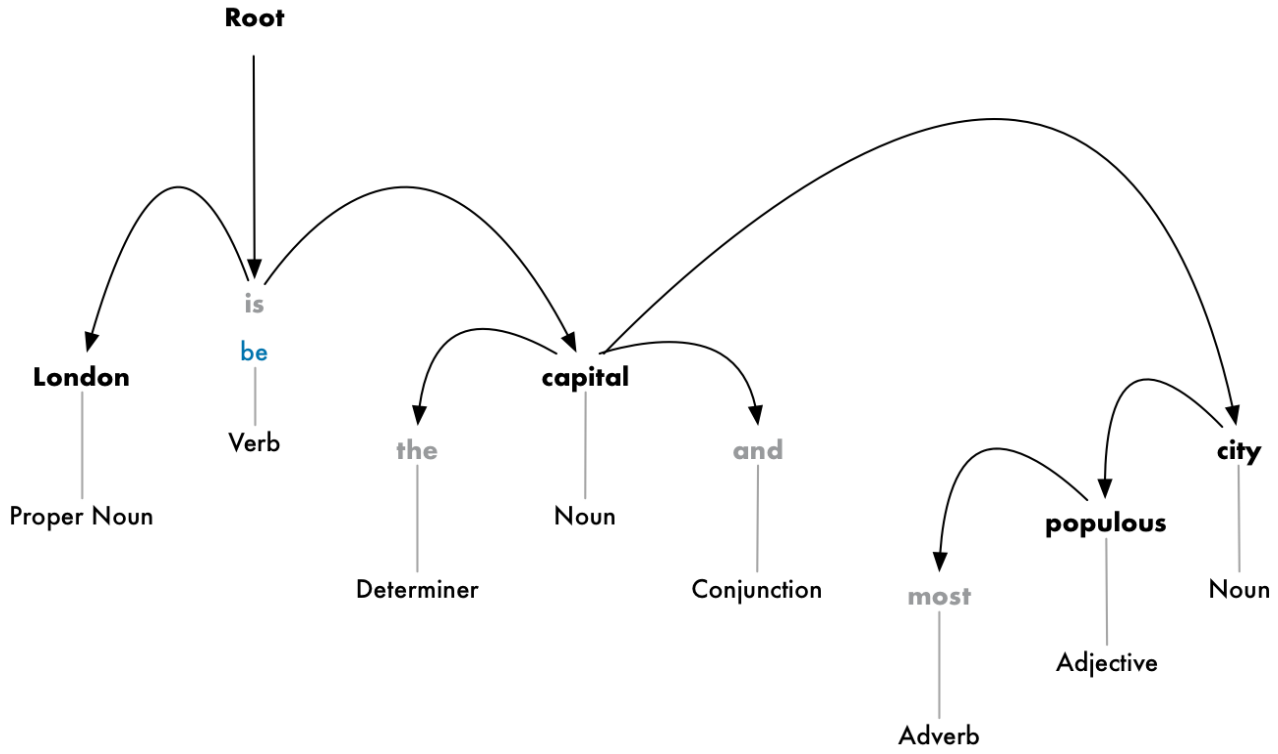
- **Example:**



Dependency Parsing

- A **dependency parser** analyzes the grammatical structure of a sentence, establishing relationships between "head" words and words which modify those heads.

- The goal is to build a tree that assigns a single **parent** word to each word in the sentence.



NER EXTRACTION

- extracting 'named-entities' from text. Named-entities denotes to words in a sentence representing real-world objects with proper names like:
  - Person's name (Ramu, Raja, Seeta, etc.),
  - Countries (India, Sri Lanka, etc),
  - Organization (Google, Facebook, etc.)
  - or anything that has been given a specific name.

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

NLP
Word Embedding

You will Learn…

- **word embedding**
- **BOW**
- **Term-Frequency**
- **TF-IDF**
- **Examples**

Word Embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques in language modeling where words or phrases from the vocabulary are mapped to vectors of real numbers.

- Every word has a unique word embedding (or "vector"), which is just a list of numbers for each word.
- The word embeddings are multidimensional; typically for a good model, embeddings are between 50 and 500 in length.
- For each word, the embedding captures the "meaning" of the word.
- Similar words end up with similar embedding values.

The man and women live happily.

The king and Queen live happily.

vocabulary { “The”, “ man”, “king”, “and”, “women”, “queen”, “live”, “happily”}

```
The     - [1,0,0,0,0,0,0,0]
man     - [0,1,0,0,0,0,0,0]
king    - [0,0,1,0,0,0,0,0]
and     - [0,0,0,1,0,0,0,0]
women   - [0,0,0,0,1,0,0,0]
queen   - [0,0,0,0,0,1,0,0]
live    - [0,0,0,0,0,0,1,0]
happily- [0,0,0,0,0,0,0,1]
```

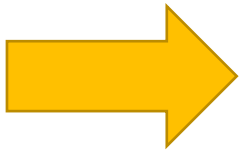
- Word2vec
- Glove

Bag-of-Words (BoW)

Bag of Words (BoW) model is the simplest form of text representation in numbers.

Review 1: This movie is very scary and long
Review 2: This movie is not scary and is slow
Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6



Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]

Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]

Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]

Term Frequency-Inverse Document Frequency (TF-IDF)

"Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus."

Term Frequency (TF) It is a measure of how frequently a term, t, appears in a document, d:

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

- numerator, n is the number of times the term "t" appears in the document "d".
- each document and term would have its own TF value.

Review 2: This movie is not scary and is slow

Vocabulary: 'This', 'movie', 'is', 'very', 'scary', 'and', 'long', 'not', 'slow', 'spooky', 'good'

Number of words in Review 2 = 8

TF('movie') = 1/8
TF('is') = 2/8 = 1/4
TF('very') = 0/8 = 0
TF('scary') = 1/8
TF('and') = 1/8
TF('long') = 0/8 = 0
TF('not') = 1/8
TF('slow') = 1/8
TF('spooky') = 0/8 = 0
TF('good') = 0/8 = 0

Term Frequency (TF)

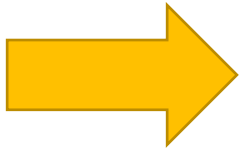
Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

Inverse Document Frequency (IDF)

IDF is a measure of how important a term is.
We need the IDF value because computing just the TF alone is not sufficient to understand the importance of words:

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

IDF('movie',) = log(3/3) = 0
IDF('is') = log(3/3) = 0
IDF('not') = log(3/1) = log(3) = 0.48
IDF('scary') = log(3/2) = 0.18
IDF('and') = log(3/3) = 0
IDF('slow') = log(3/1) = 0.48



Term	Review 1	Review 2	Review 3	IDF
This	1	1	1	0.00
movie	1	1	1	0.00
is	1	2	1	0.00
very	1	0	0	0.48
scary	1	1	0	0.18
and	1	1	1	0.00
long	1	0	0	0.48
not	0	1	0	0.48
slow	0	1	0	0.48
spooky	0	0	1	0.48
good	0	0	1	0.48

TF-IDF

$$(tf\_idf)_{t,d} = tf_{t,d} * idf_t$$

TF-IDF('movie', Review 2) = 1/8 * 0 = 0
TF-IDF('is', Review 2) = 1/4 * 0 = 0
TF-IDF('not', Review 2) = 1/8 * 0.48 = 0.06
TF-IDF('scary', Review 2) = 1/8 * 0.18 = 0.023
TF-IDF('and', Review 2) = 1/8 * 0 = 0
TF-IDF('slow', Review 2) = 1/8 * 0.48 = 0.06

Term	Review 1	Review 2	Review 3	IDF	TF-IDF (Review 1)	TF-IDF (Review 2)	TF-IDF (Review 3)
This	1	1	1	0.00	0.000	0.000	0.000
movie	1	1	1	0.00	0.000	0.000	0.000
is	1	2	1	0.00	0.000	0.000	0.000
very	1	0	0	0.48	0.068	0.000	0.000
scary	1	1	0	0.18	0.025	0.022	0.000
and	1	1	1	0.00	0.000	0.000	0.000
long	1	0	0	0.48	0.068	0.000	0.000
not	0	1	0	0.48	0.000	0.060	0.000
slow	0	1	0	0.48	0.000	0.060	0.000
spooky	0	0	1	0.48	0.000	0.000	0.080
good	0	0	1	0.48	0.000	0.000	0.080

Thank you.

Questions?