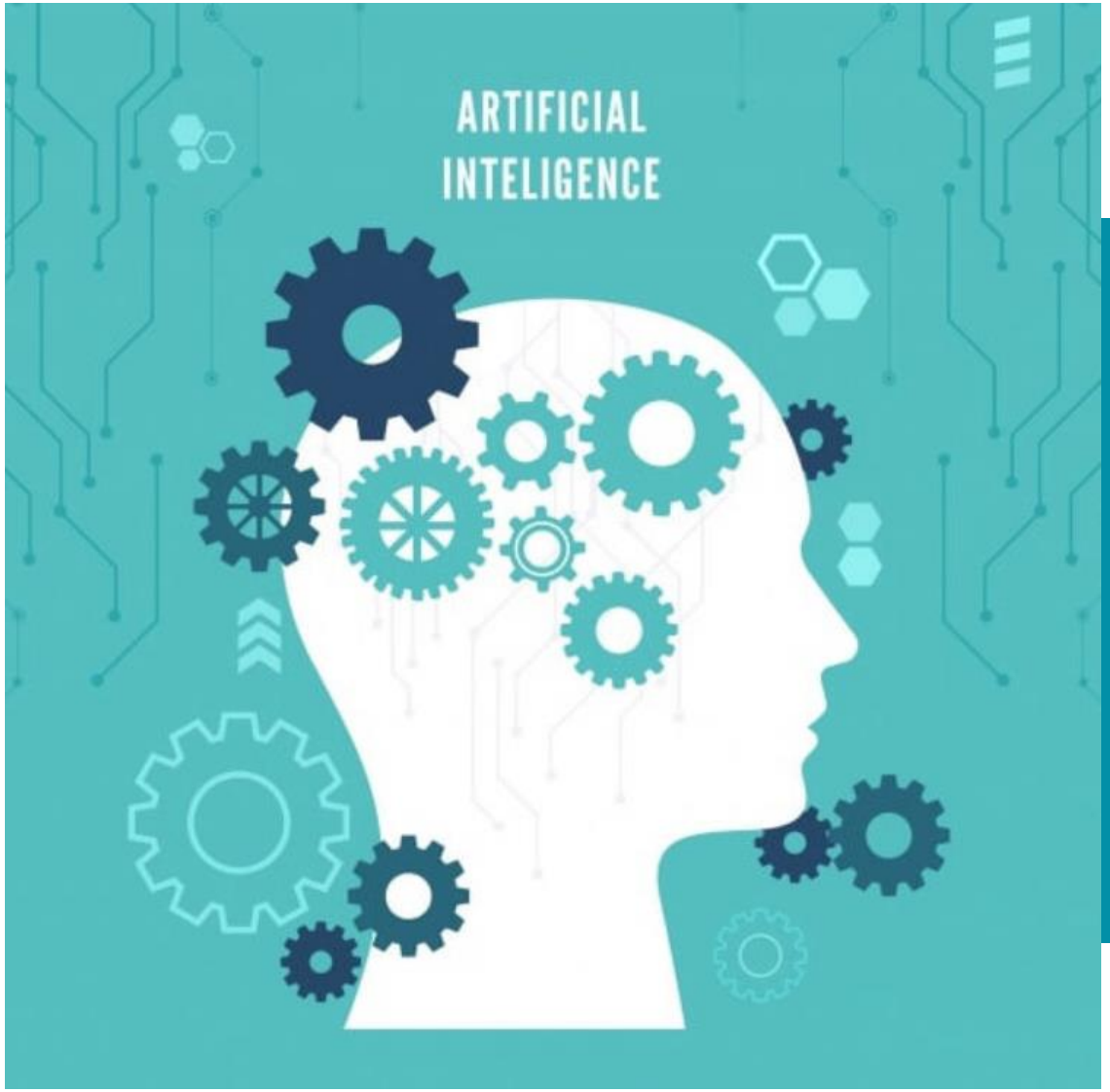
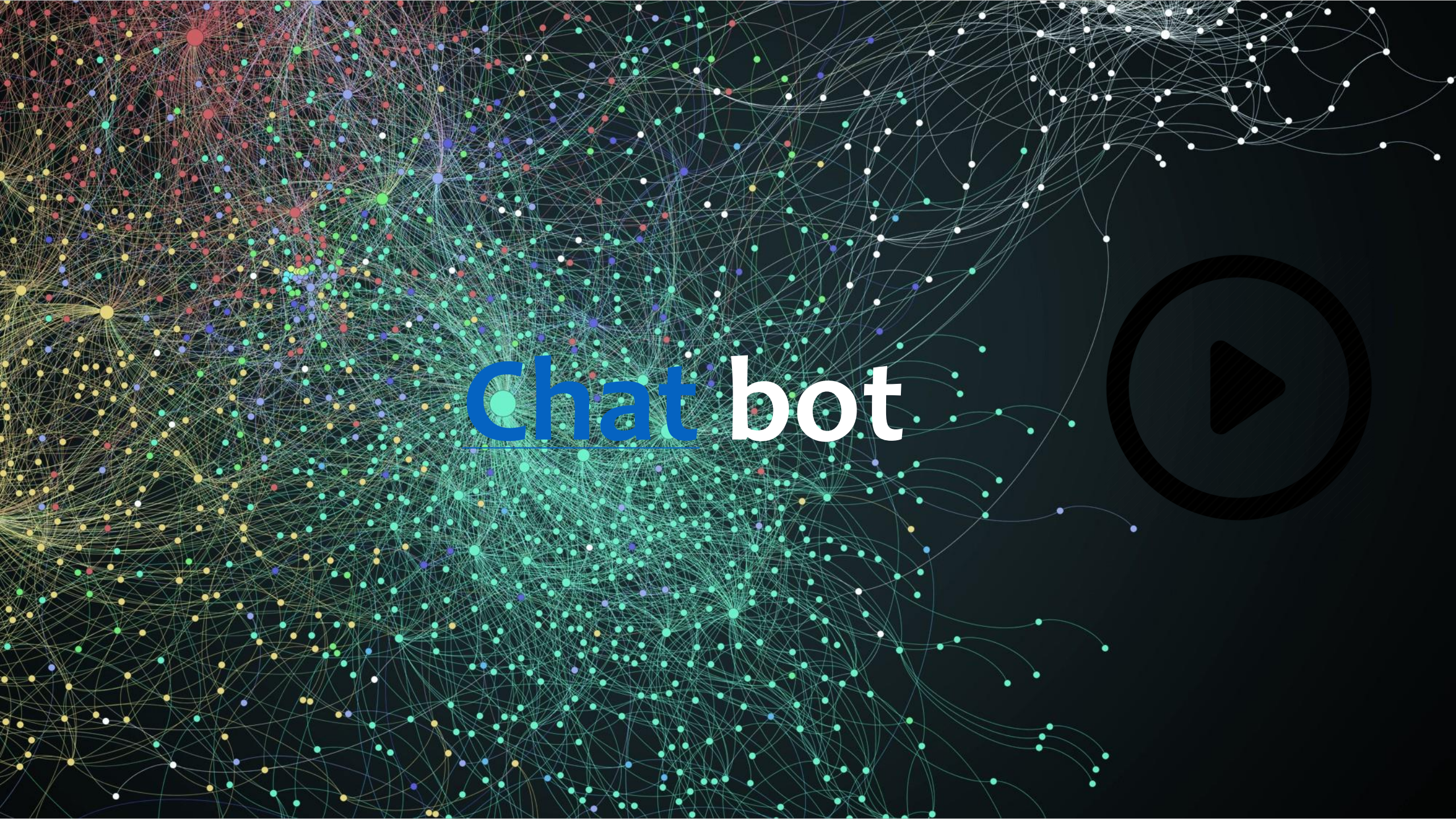


MEDEVA

An Introduction to Text Mining


Chat bot

80% of typical customer questions can be handled by AI chatbots

$8 billion per year can potentially be saved by investing in AI chatbots

90% success rate with chatbot interactions expected by 2022

Working with Text

- Dr Venugopala Rao Manneni
- www.statsvenu.com

Content

- Why Natural Language Processing
- What is Natural Language Processing
- Tasks Of Natural Language Processing
- Machine / Deep Learning and NLP
- How To Learn Natural Language Processing
- Hands on

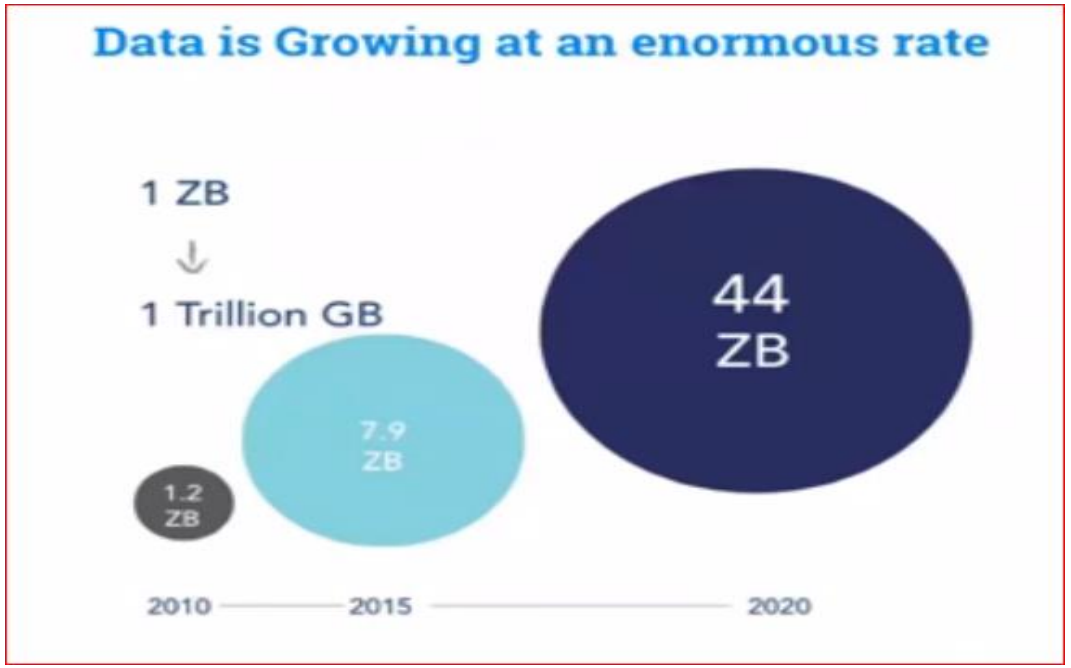
Why NLP

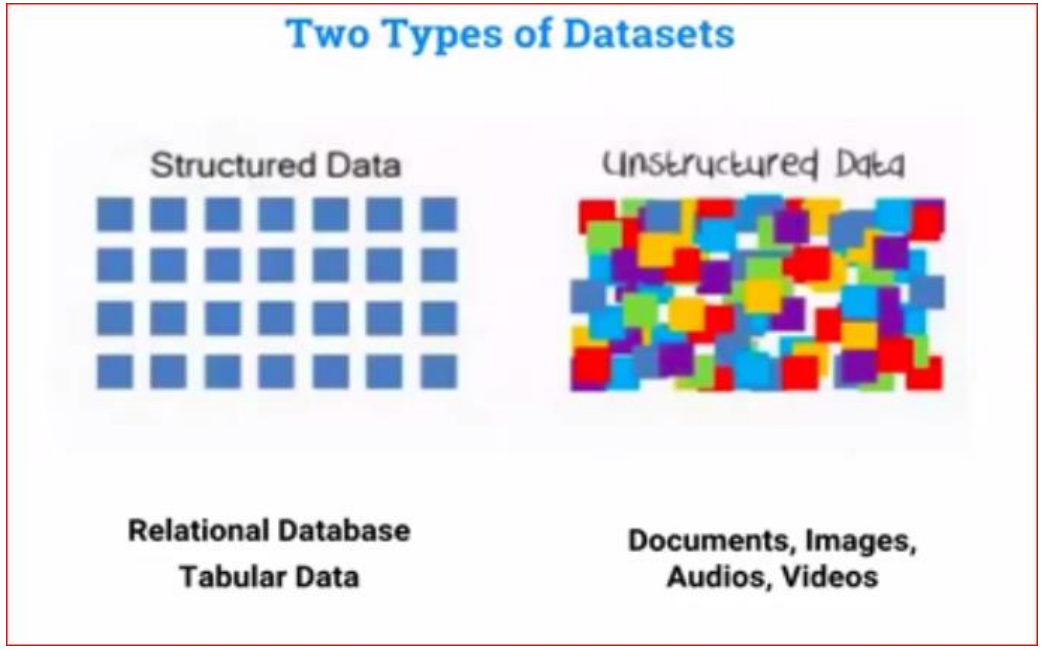
1. Tremendous growth in data in last few years and natural language processing is required to deal with such data

2. Number of Business Use-Cases which can only be solved using natural language processing

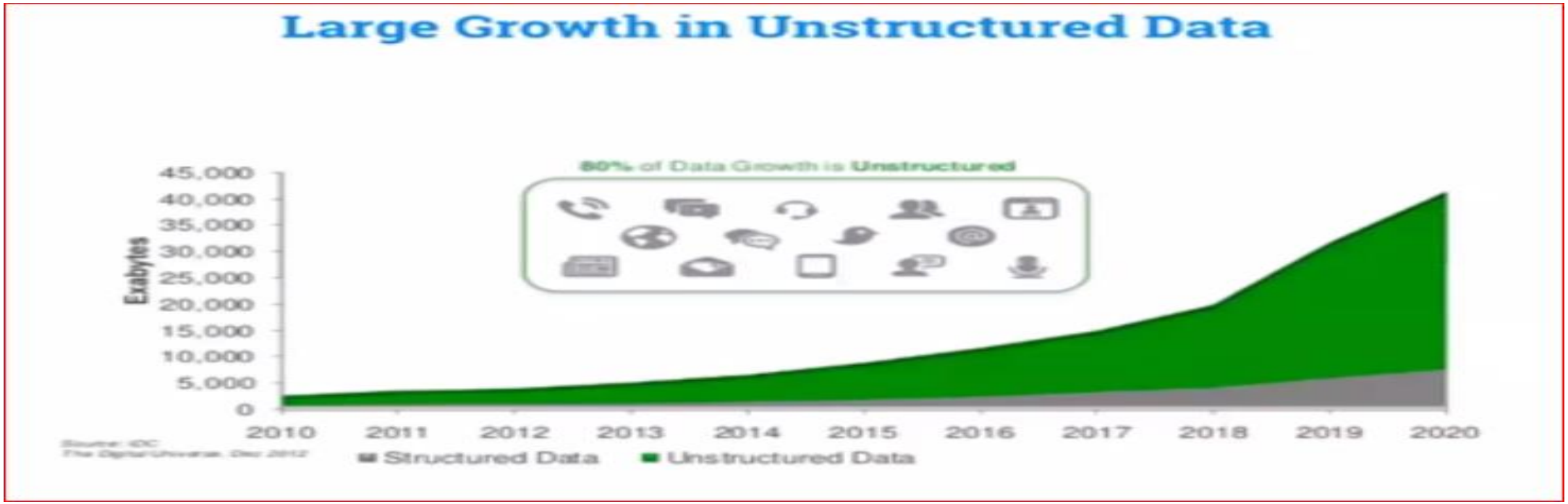
3. Number of Applications in which the use of natural language processing has improved the user experiences

Data Growth

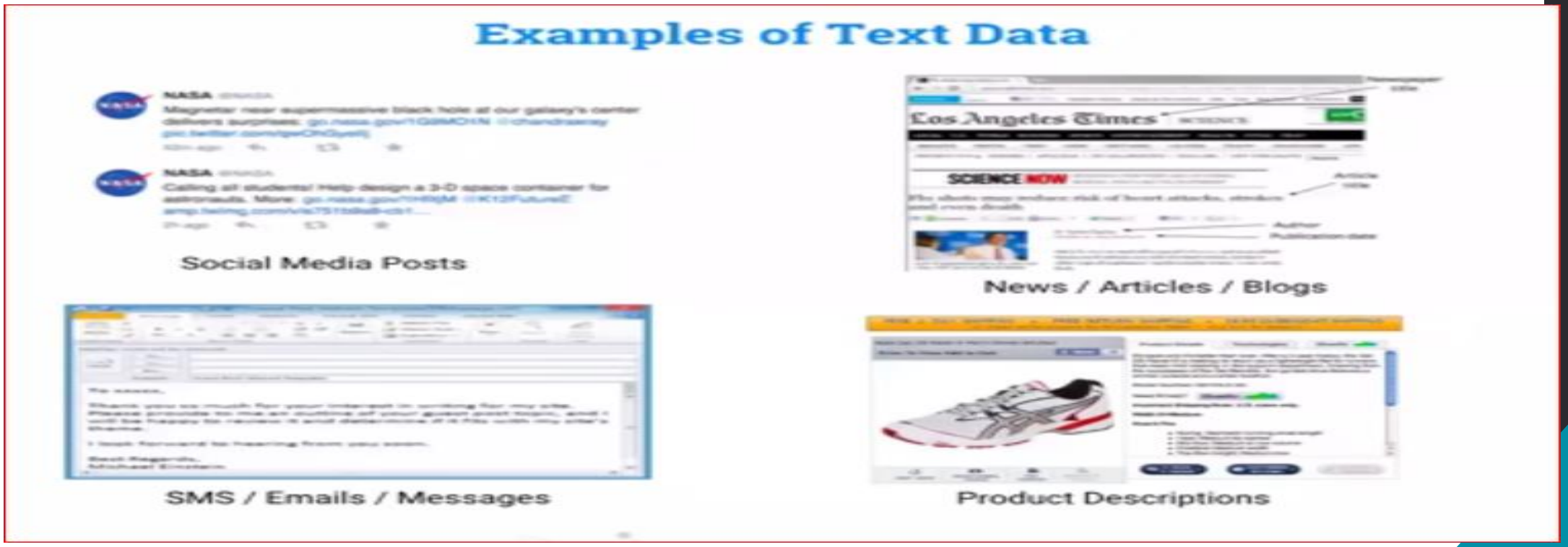

Data is Growing at an enormous rate
1 ZB
1 Trillion GB
44 ZB
7.9 ZB
1.2 ZB
2010
2015
2020


Two Types of Datasets
Structured Data
Unstructured Data
Relational Database
Tabular Data
Documents, Images,
Audios, Videos

Out of 100 percept 80% growth is unstructured data

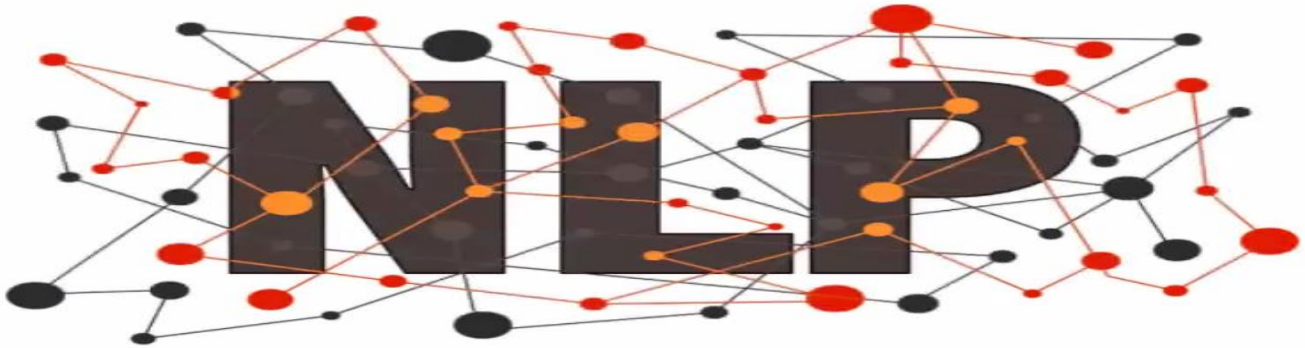


Text Data


Examples of Text Data
Social Media Posts
News / Articles / Blogs
SMS / Emails / Messages
Product Descriptions

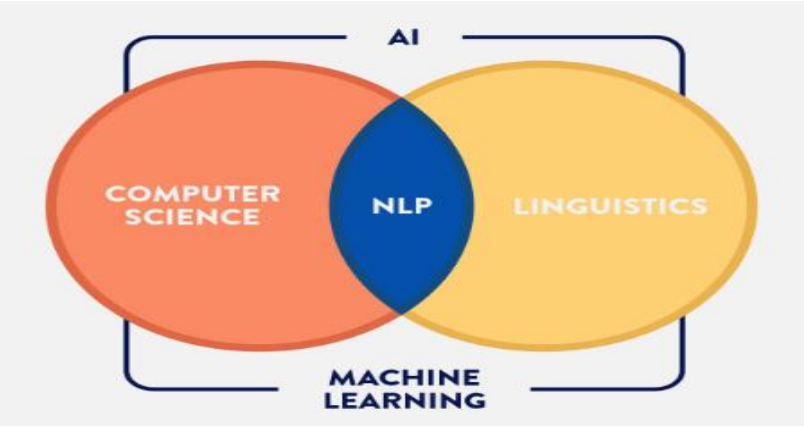
The Solution is

How to deal with such data ?



NATURAL LANGUAGE PROCESSING

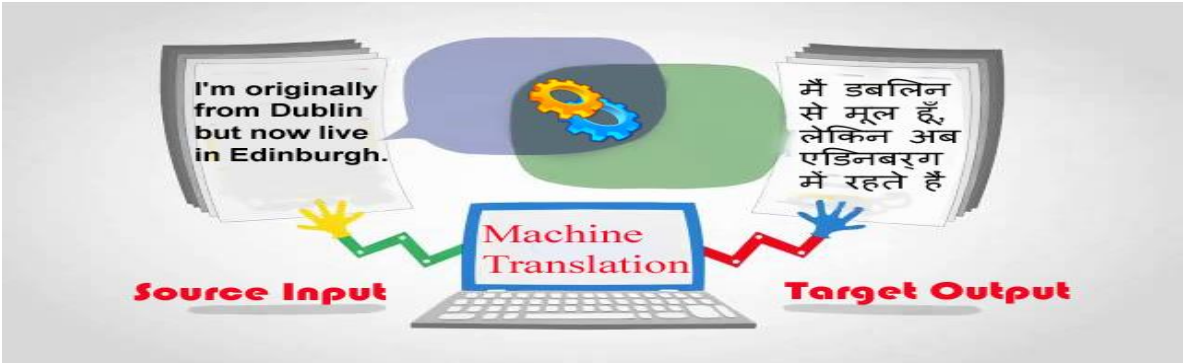
- NLP is an intersection of computer science , linguistics(scientific study of language), and machine learning that is connected with communication between computers and humans in natural language



Number of NLP Applications

APPLICATIONS OF NLP:

- **MACHINE TRANSLATION:** it is process by which computer software is used to translate a text from one natural language to another.



- **SENTIMENT ANALYS** : sentiment analysis is also known as opnion mining is a field within NLP that builds systems that try to identify and extract opnion within text.

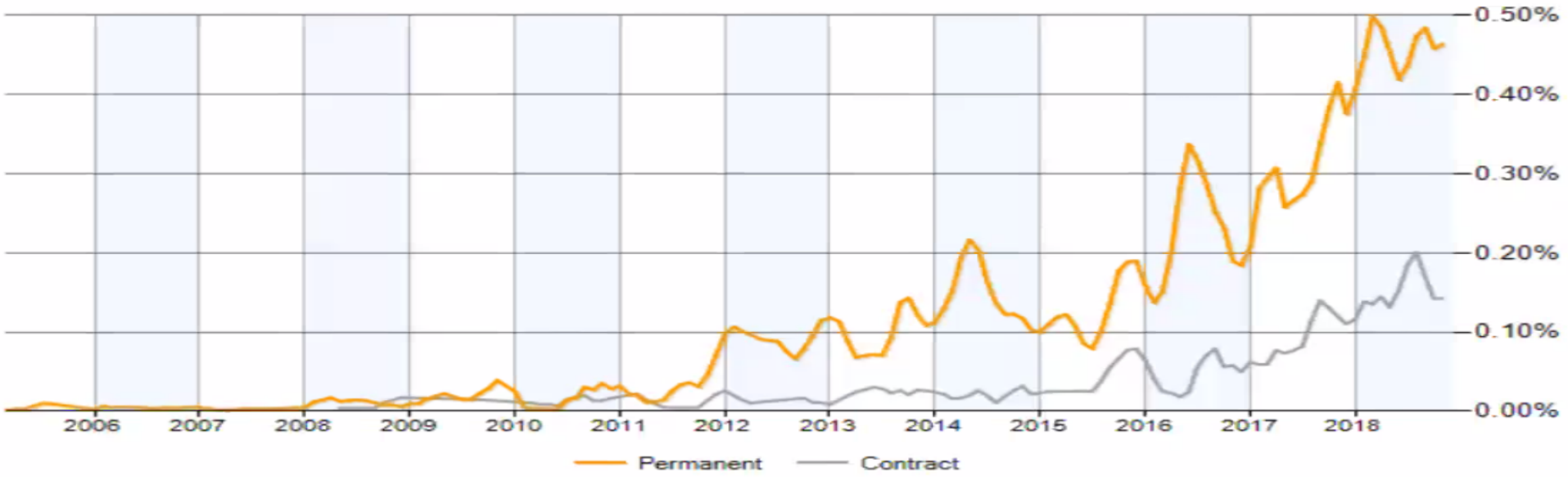
MEDEVA

- **IMAGE CAPTIONING:** this is the process of generating textual description of an image .it uses both NLP and computer vision to generate the captions.

- **AUTOMATIC SUMMARIZATION:** it is the process of shortening a text document with software,in order to create a summary with the major points of the original documents.

MEDEVA

Industry Trends : NLP - The hottest skill



What is Natural Language Processing

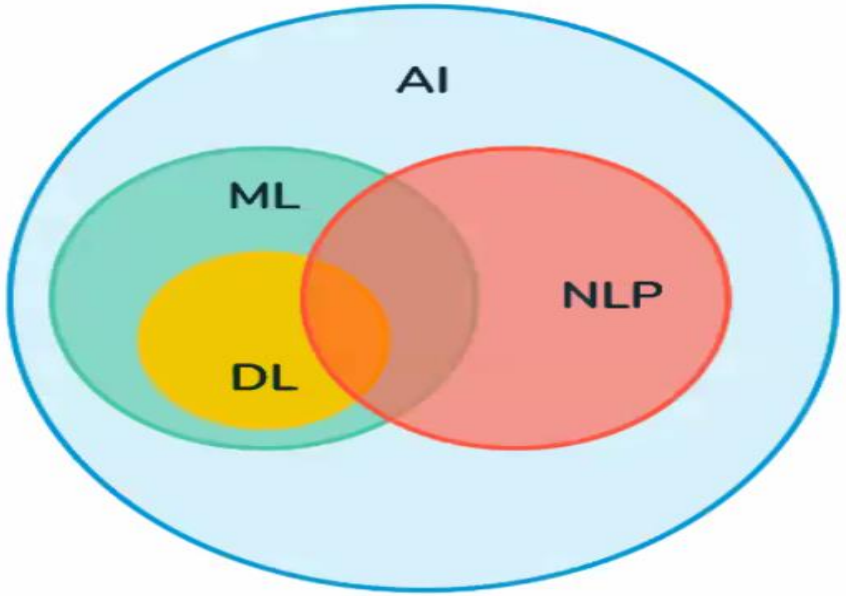
Branch of Data Science, Computer Science and Linguistics that deals with techniques to derive information from the text data

Structured Information can be used as below:

Natural Language Processing Outcomes

1. Used to generate actionable Insights that matter for businesses
2. Used as features to machine learning or deep learning problems
3. Used to create knowledge databases that powers many other applications (chatbots, searchengines, recommendation engines)
4. Used to automate many business process

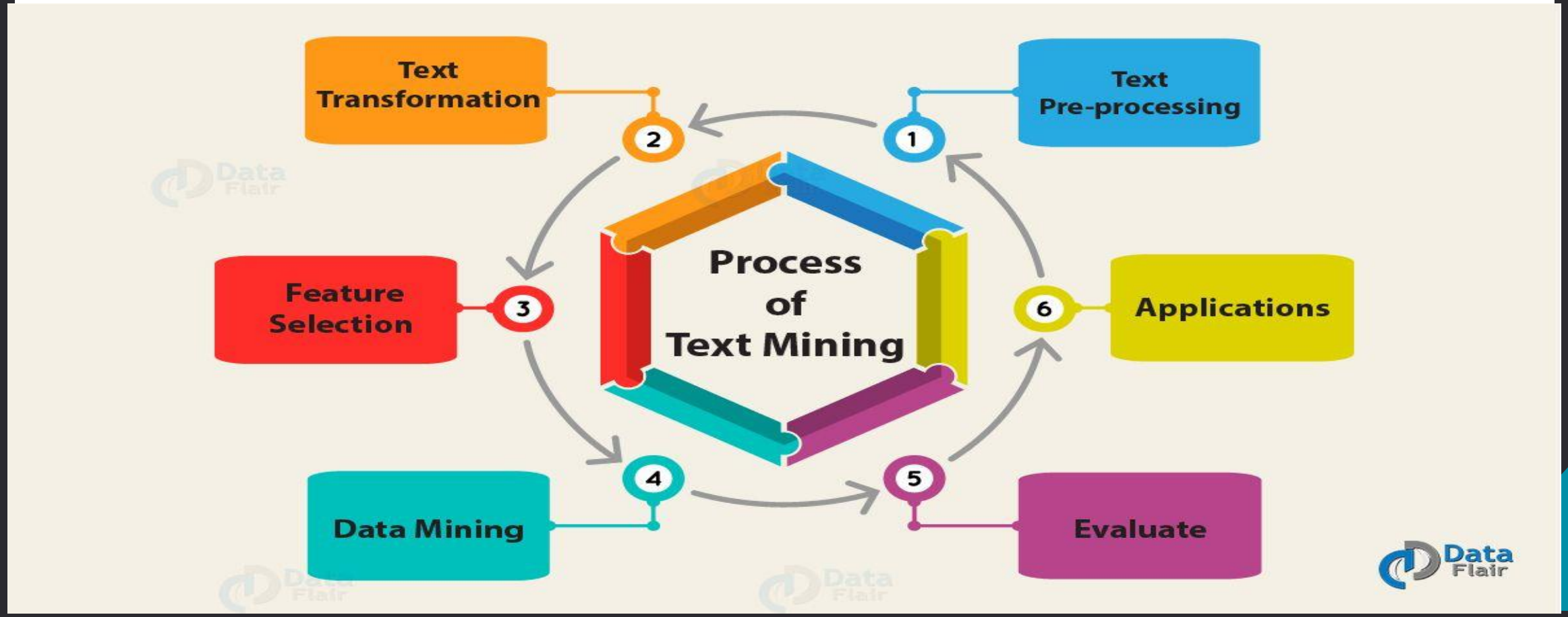
What is Natural Language Processing

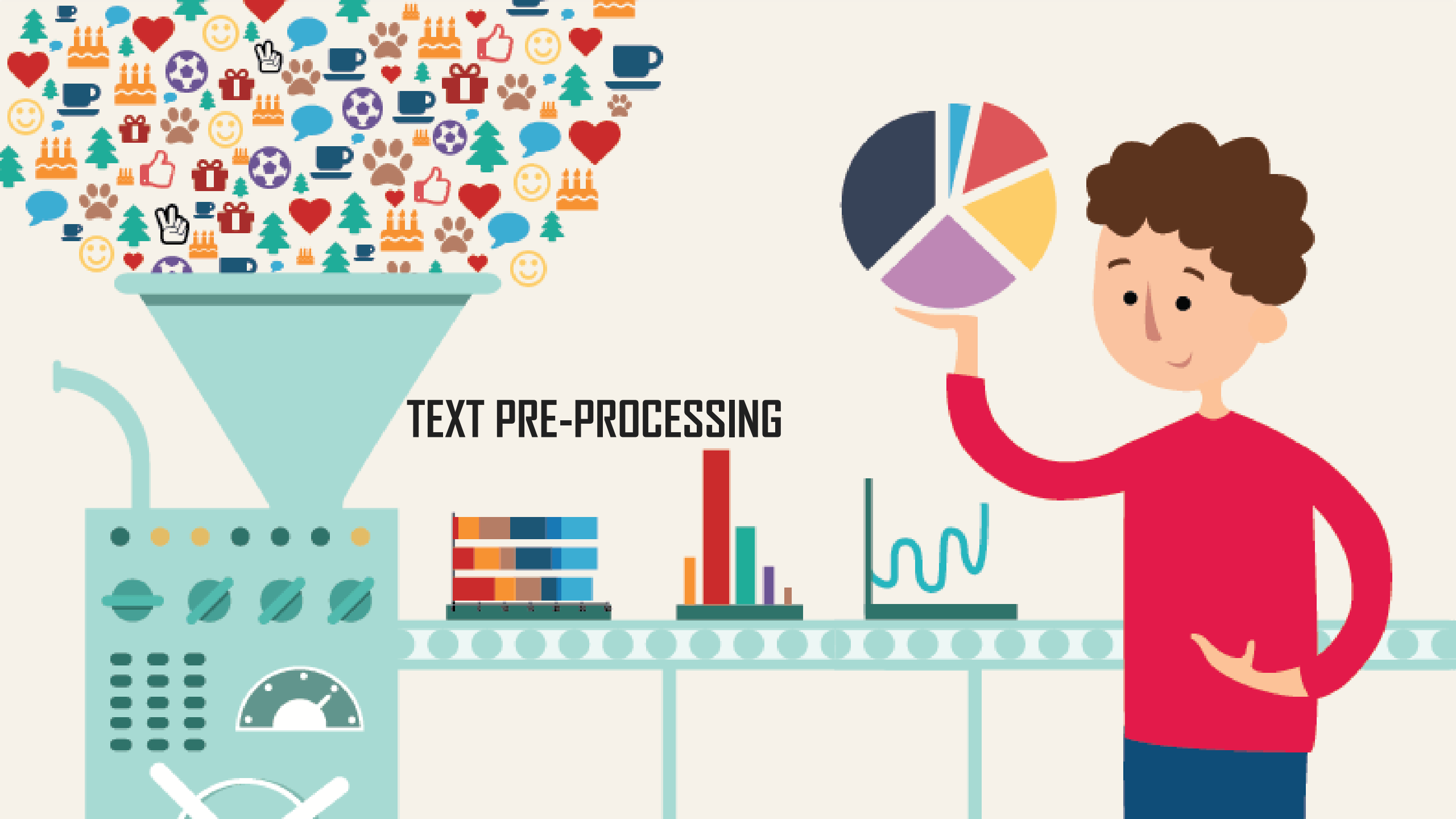


NLP comprises of set of different tasks related to text data.

Some are overlapped with Machine Learning and Some with Deep Learning

STANDARD NLP WORKFLOW




TEXT PRE-PROCESSING

What is text preprocessing?

To preprocess your text simply means to bring your text into a form that is **predictable** and **analyzable** for your task

Types of text pre-processing techniques

There are different ways to preprocess your text. Here are some of the approaches that you should know about

Lowercasing:
Lowercasing ALL your text data

Noise Removal:
This includes punctuation removal, spell checking, special character removal

Removing HTML tags:

Expanding Contractions:
Contractions are shortened version of words or syllables.
Ex: do not to ***don't*** and ***I would*** to ***I'd***

Removing accented characters :
Converting into ASCII
Ex: converting **é** to **e**.

Removing Special Characters:
Special characters and symbols are usually non-alphanumeric

Dataset PreProcessing : Text Cleaning

Process of removal of noise present in the text data

Different Steps of a Text Cleaning Pipeline

1. Removal of stopwords, punctuations, html entities
2. Keyword Normalization
3. Spelling Correction, Grammar Mistakes
4. Standardization of words, removal of slangs

Natural Language Processing Tasks

Tasks

1. Tokenization
2. Normalization
3. Part of Speech Tagging
4. Dependency Parsing
5. Entity Extraction

Problems

1. Topic Modelling
2. Text Classification
3. Text Generation
4. Machine Translation
5. Text Summarization

Text Pre Processing - Tokenization

NLP Tasks : Tokenization

Process of splitting a text object into smaller units (tokens)

Smaller Units : words, numbers, symbols, ngrams, characters

White space tokenizer / Unigram tokenizer

Sentence : "I went to New-York to play football"
Tokens : "I", "went", "to", "New-York", "to", "play", "football"

Normalization

Normalization : Process of converting a token into its base form
Two Types : Stemming and Lemmatization

Original	Stemmed	Lemmatized
visibilities	visibl	visibility
adhere	adher	adhere
adhesion	adhes	adhesion
appendicitis	append	appendicitis
oxen	oxen	ox
indices	indic	index
swum	swum	swim

FEATURE ENGINEERING

Feature Extractor

Process of Quantifying text data into features (numbers)

- Meta Features
  - Word Counts / Character Counts
- NLP Features
  - Part of Speech / Topics Present
- TF IDF Features
  - Term Frequency and Inverse Document Frequency
  - Word / Character level
- Word Vector Notations
  - Word Embeddings

BAG OF N-GRAMS MODEL

The Bag of Words model doesn't consider order of words. But what if we also wanted to take into account phrases or collection of words which occur in a sequence?
N-grams help us achieve that

- An N-gram is basically a collection of word tokens from a text document such that these tokens are contiguous and occur in a sequence.
- Bi-grams indicate n-grams of order 2 (two words), Tri-grams indicate n-grams of order 3 (three words), and so on.

	bacon eggs	beautiful sky	beautiful today	blue beautiful	blue dog	blue sky	breakfast sausages	brown fox	dog lazy	eggs ham	...	lazy dog	love blue	love green	quick blue	quick brown	sausages bacon	sausages ham	sky beautiful
0	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	1	0	1	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	...	1	0	0	0	1	0	0	0
3	1	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	1	...	0	0	1	0	0	1	0	0
5	0	0	0	0	1	0	0	1	1	0	...	0	0	0	1	0	0	0	0
6	0	0	1	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	1	1	0	...	0	0	0	0	0	0	0	0

8 rows × 29 columns

Bi-gram based feature vectors using the Bag of N-Grams Model

MEDEVA

TF-IDF MODEL

TF-IDF stands for Term Frequency-Inverse Document Frequency.

Term Frequency: This summarizes how often a given word appears within a document.

Inverse Document Frequency: This downscales words that appear a lot across documents.

	bacon	beans	beautiful	blue	breakfast	brown	dog	eggs	fox	green	ham	jumps	kings	lazy	love	quick	sausages	sky	toast	today
0	0.00	0.00	0.60	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.0
1	0.00	0.00	0.49	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00	0.49	0.00	0.0
2	0.00	0.00	0.00	0.00	0.00	0.38	0.38	0.00	0.38	0.00	0.00	0.53	0.00	0.38	0.00	0.38	0.00	0.00	0.00	0.0
3	0.32	0.38	0.00	0.00	0.38	0.00	0.00	0.32	0.00	0.00	0.32	0.00	0.38	0.00	0.00	0.00	0.32	0.00	0.38	0.0
4	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00	0.47	0.39	0.00	0.00	0.00	0.39	0.00	0.39	0.00	0.00	0.0
5	0.00	0.00	0.00	0.37	0.00	0.42	0.42	0.00	0.42	0.00	0.00	0.00	0.00	0.42	0.00	0.42	0.00	0.00	0.00	0.0
6	0.00	0.00	0.36	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.00	0.5
7	0.00	0.00	0.00	0.00	0.00	0.45	0.45	0.00	0.45	0.00	0.00	0.00	0.00	0.45	0.00	0.45	0.00	0.00	0.00	0.0

Our TF-IDF model based document feature vectors

The TF-IDF based feature vectors for each of our text documents show scaled and normalized values as compared to the raw Bag of Words model values.

MEDEVA

WORD EMBEDDING

Word embedding is one of the most popular representations of document vocabulary.It is capable of capturing context of a word in a document, semantic and syntactic similarity,relation with other words, etc

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

Example for Word Embedding

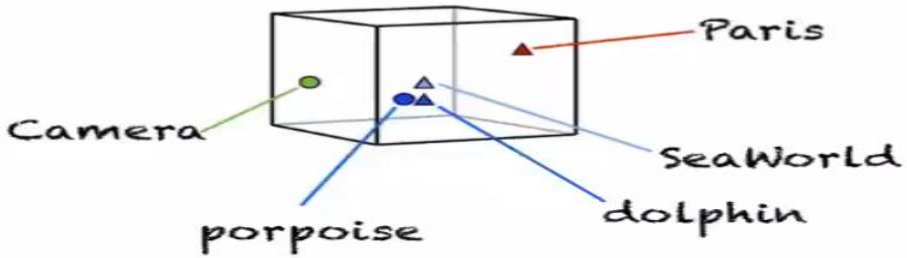
WORD2VEC :is one of the most popular technique to learn word embeddings
This model was created by Google in 2013 and is a predictive deep learning based model

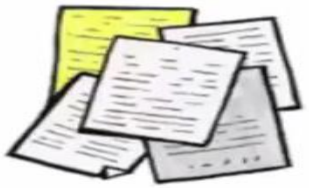
GLOVE MODEL : The GloVe model stands for Global Vectors which is an unsupervised learning model which can be used to obtain dense word vectors similar to Word2Vec

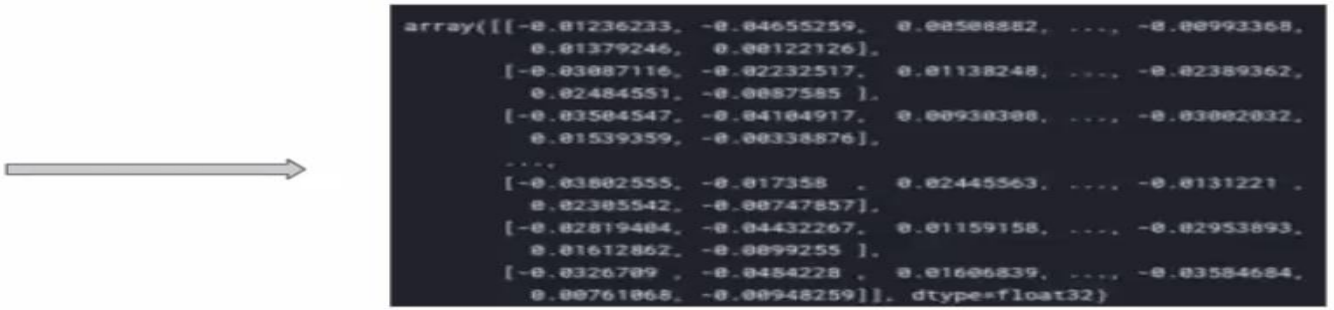
THE FAST TEXT MODEL : The model was first introduced by Facebook in 2016 as an extension and supposedly improvement of the vanilla Word2Vec mode

Word Embeddings : Vector Notations

Representation of words in an n-dimensional vector space





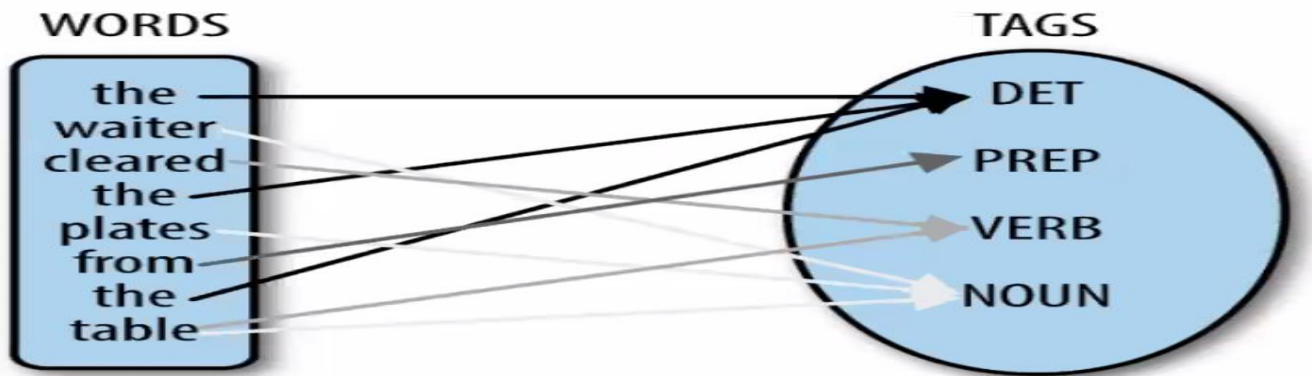


```
array([[-0.01236233, -0.04655259,  0.08508882, ..., -0.00993368,
         0.01379246,  0.00122126],
       [-0.03087116, -0.02232517,  0.01138248, ..., -0.02389362,
         0.02484551, -0.0087585 ],
       [-0.03504547, -0.04104917,  0.00930308, ..., -0.03002032,
         0.01539359, -0.00338876],
       ...,
       [-0.03802555, -0.017358  ,  0.02445563, ..., -0.0131221 ,
         0.02305542, -0.00747857],
       [-0.02819404, -0.04432267,  0.01159158, ..., -0.02953893,
         0.01612862, -0.0099255 ],
       [-0.0326709 , -0.0454228 ,  0.01606839, ..., -0.03584684,
         0.00761068, -0.00948259]], dtype=float32)
```

Role / Function of given word

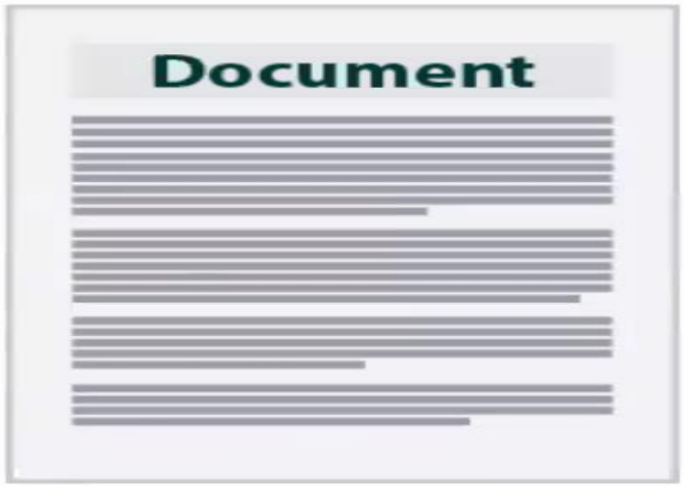
NLP Tasks : Part of Speech Tagging

Defining the Part of Speech tag of every word in the sentence which defines the role of words in the sentence



NLP Tasks : Entity Extraction

Process of identifying key entities present in the text data



People	
Companies	
Places	
Money	$ € £ ¥
Links	foo.com, acme.org, bar.com
Phone #s	1-800-12345, +353-44-34-254
	...

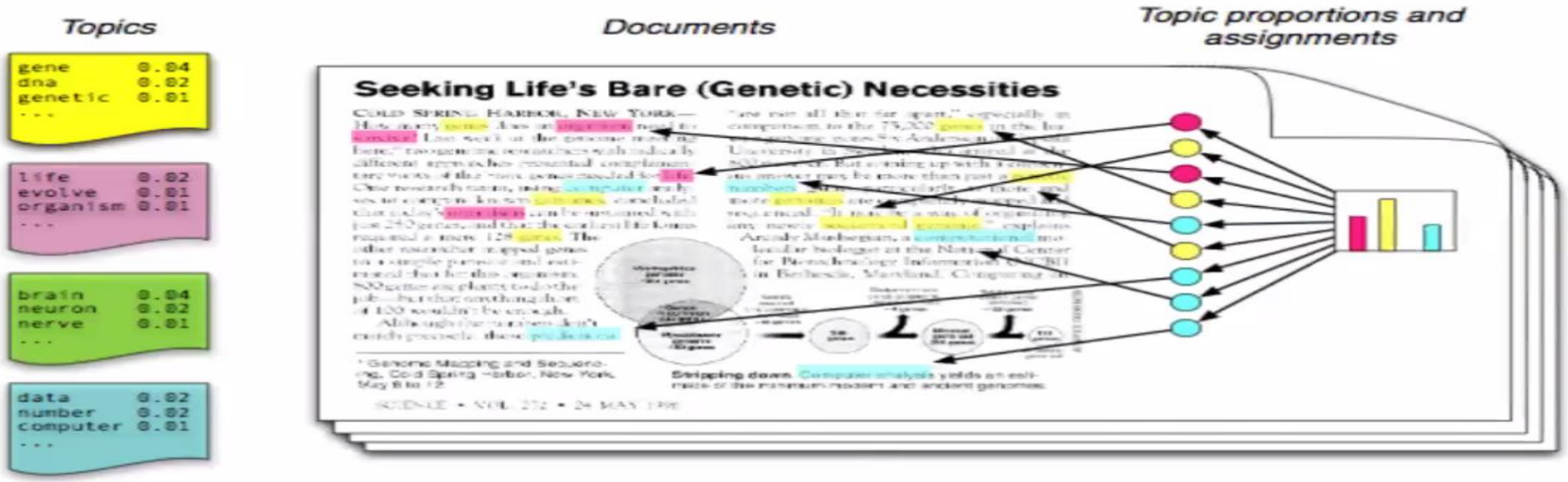
NLP Tasks : Entity Extraction

Identification of Person / Location / Company / Dates etc

Albert Einstein PER Albert Einstein was born in Ulm LOC in Germany LOC on March 14, 1879. Six weeks later the family moved to Munich LOC , where he later on began his schooling at the Luitpold Gymnasium ORG . In 1896 he entered the Swiss Federal Polytechnic School ORG in Zurich LOC to be trained as a teacher in physics and mathematics.

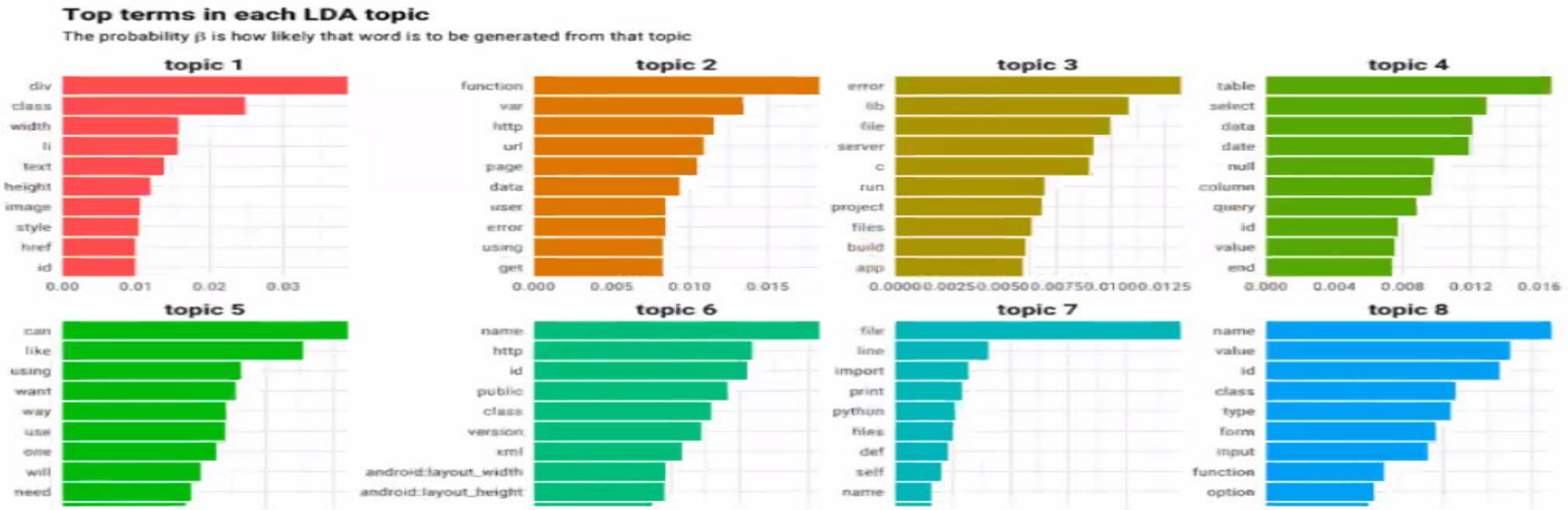
NLP Tasks : Topic Modelling

Process of automatically identifying topics from the documents



NLP Tasks : Topic Modelling

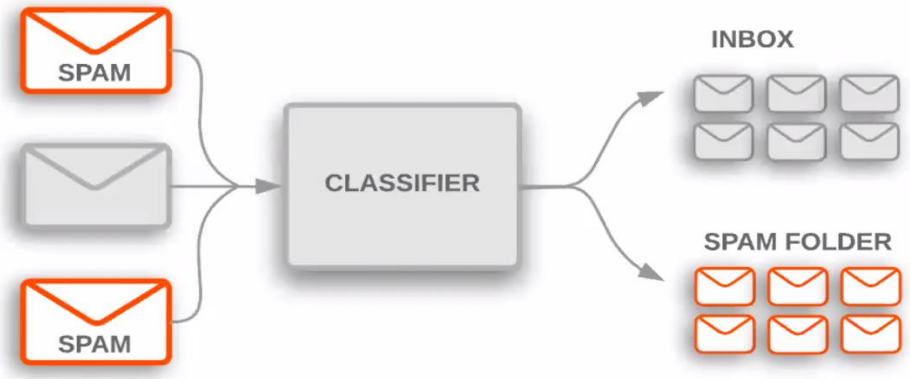
Topic Modelling Outputs



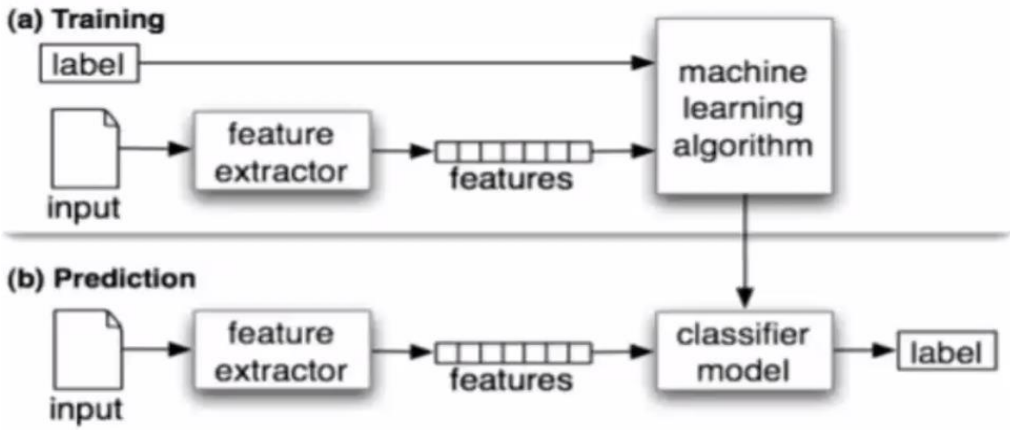
Machine Learning and NLP

Text Classification Problem

Classify the given documents / text objects into a fixed set of classes



Text Classification Tasks



Text Classification Models

Rule Based
- Hand Crafted Rules

Probability Based
- Naïve Bayes

Learning Based
- Logistic Regression
- State Vector Machines
- Ensemble Models
- Neural Networks

Deep Learning and NLP

Convolutional Neural Networks (CNNs)

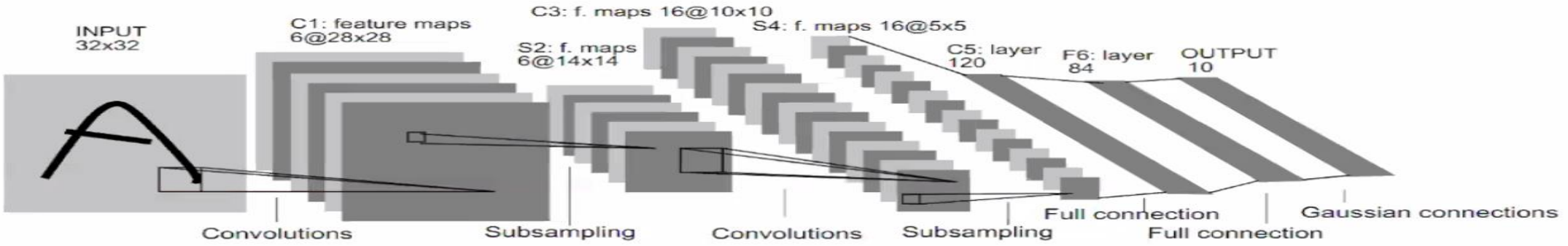
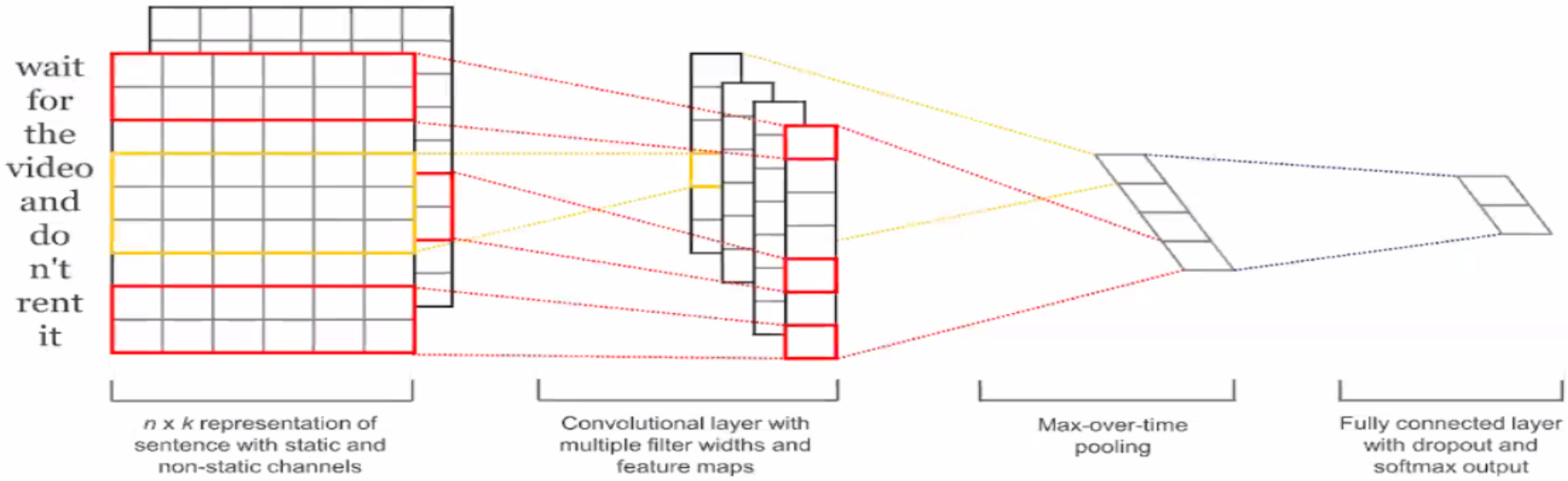
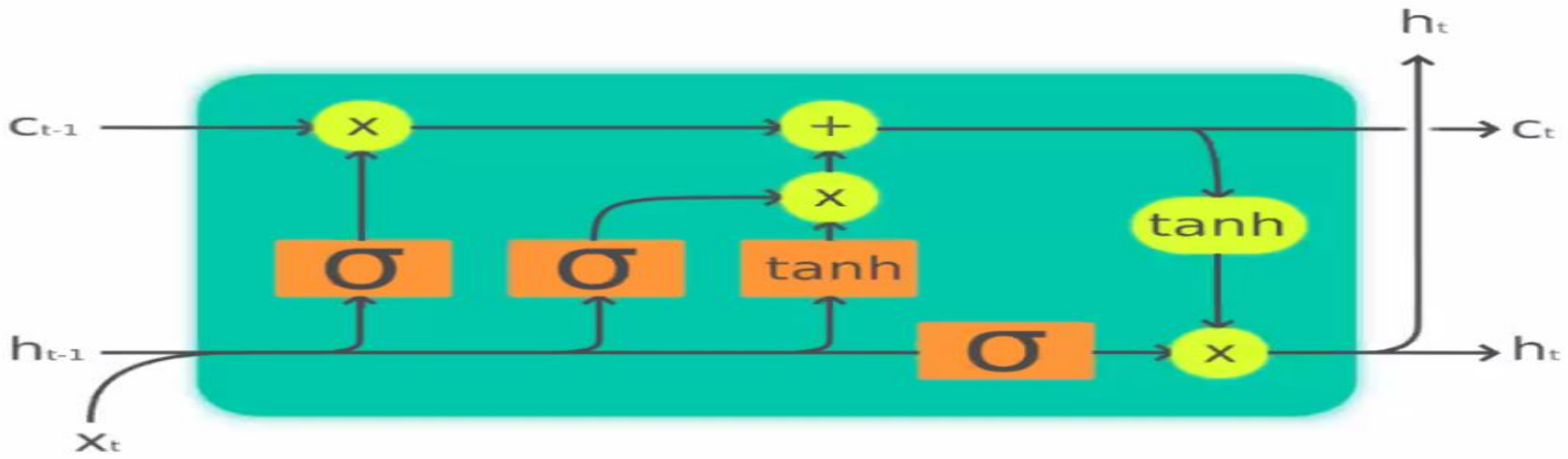


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

CNNs for Text Classification



Long Short Term Memory (LSTM) Models



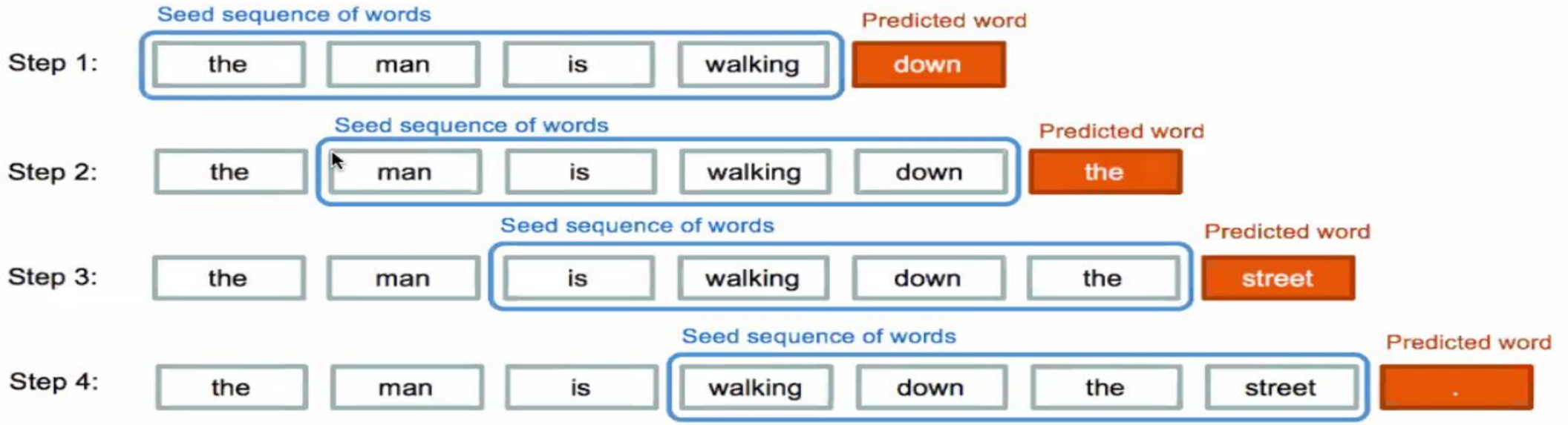
Legend:

Layer

Pointwize op

Copy

LSTMs for Text Generation



6 Steps to Learn NLP

Step 1

Practice the Basics

Step 2

NLP Tasks and Libraries

Step 3

Entity Extraction, Topic Modelling

Step 4

Information Retrieval using NLP

Step 5

Word Embeddings, Feature Engineering

Step 6

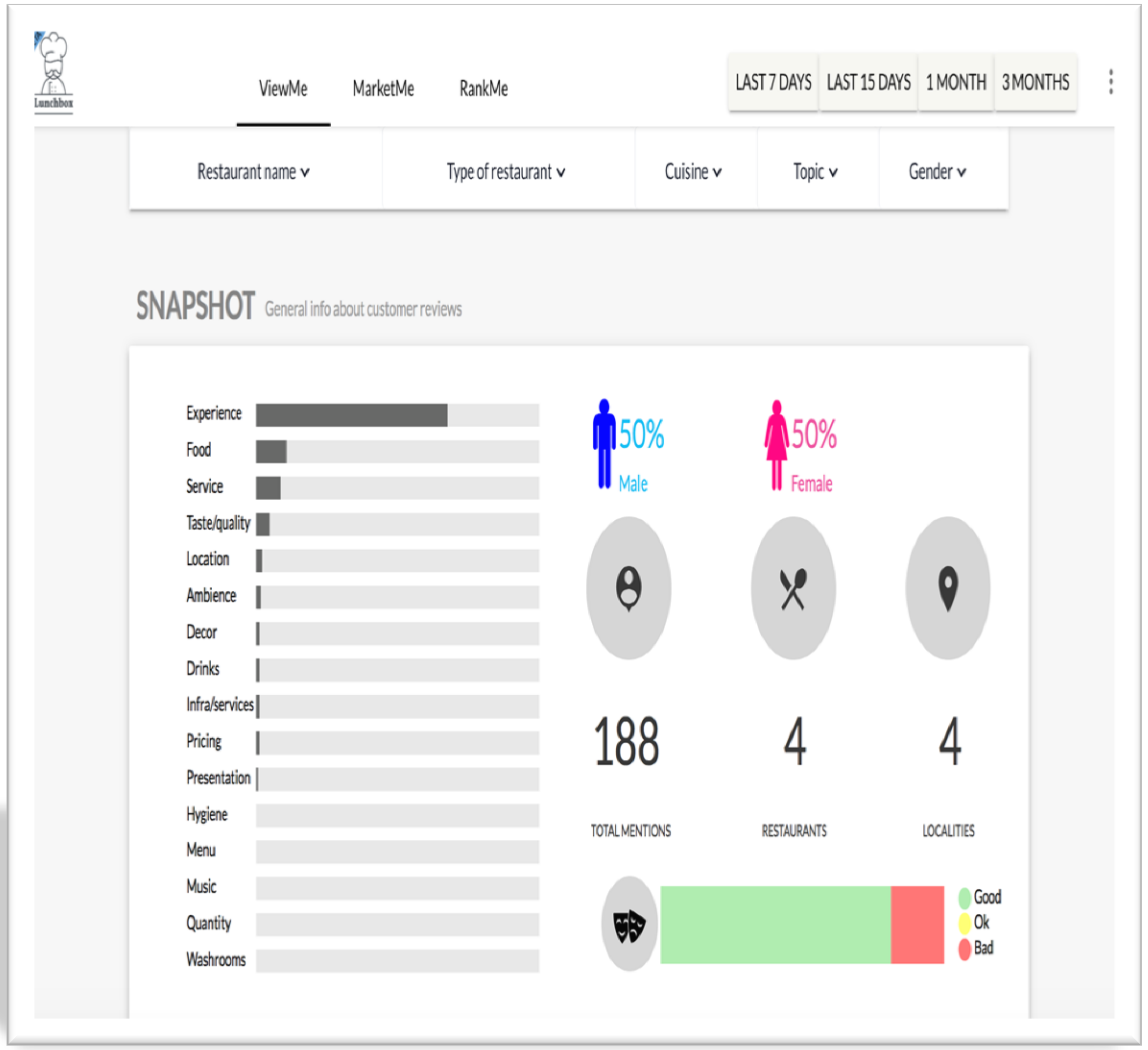
Machine Learning and Deep Learning with NLP

Practice NLP Problems

- Twitter Sentiment Analysis
- Hate Speech Classification
- Classification of Toxic Comments on Wikipedia
- Classification of E-commerce products into their categories
- Classification of news articles into the respective categories
- Email Spam - Non Spam classifier
- Building a chatbot to automatically reply on customer reviews

MINING CUSTOMER REVIEWS

CASE STUDY



Challenge

Offering critical insights into brand perception and competitive scenarios to experience brands wanting to manage and learn from data on review sites that is viewed by millions.

Solution

We scrapped over 10 million customer reviews for more than 100000 restaurants across India.

We focused on the following questions for analysing the reviews:

- What are the keywords and topics of discussion across the comments?
- What elements of the restaurant would they want improved – service, staff behaviour, ambience etc.?
- What is the overall mood of the customers visiting the restaurant?

Using advanced techniques of text analytics - machine learning for topics classification, sentiment analysis, and more, we developed a module where restaurant collected reviews could also be analysed and compared for taking better decisions.

- Handson