NATURAL LANGUAGE PROCESSING

Niharika Krishnan

Faculty Development Program, Auxilium College (06 Jun 2020)

Before we get started....

NIHARIKA KRISHNAN

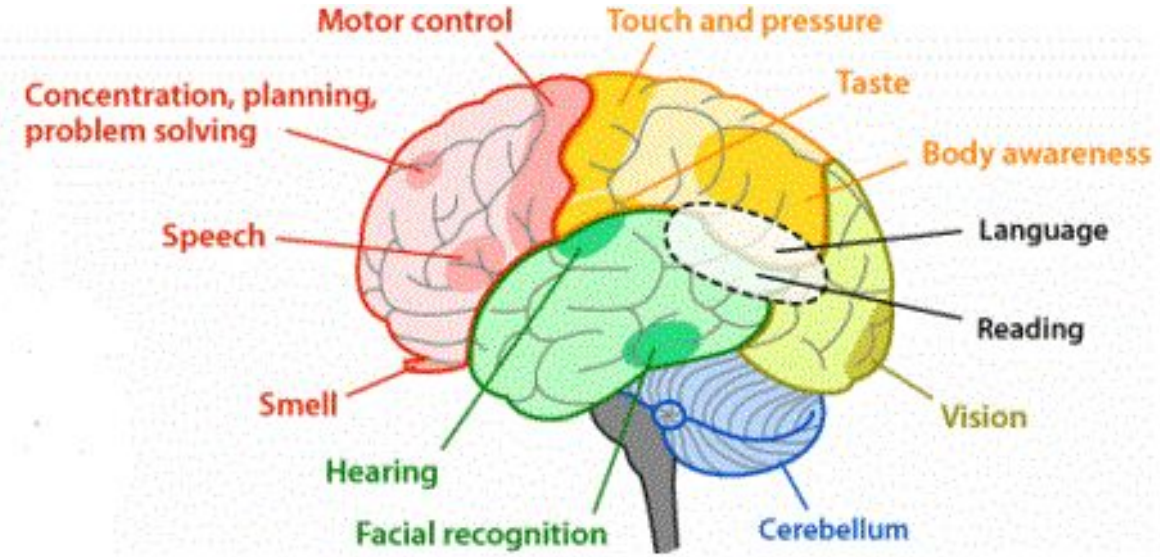
- Machine Learning Engineer, TCS
  - Build Chatbots for a living!
- Lead at PyLadies Chennai
  - Community of 100+ women tech enthusiasts
- Tech speaker
  - PyCon India'19, Canada'19
  - Google Women Techmakers,
- AI and NLP enthusiast
- Always willing to learn more!

WHAT TO EXPECT ?

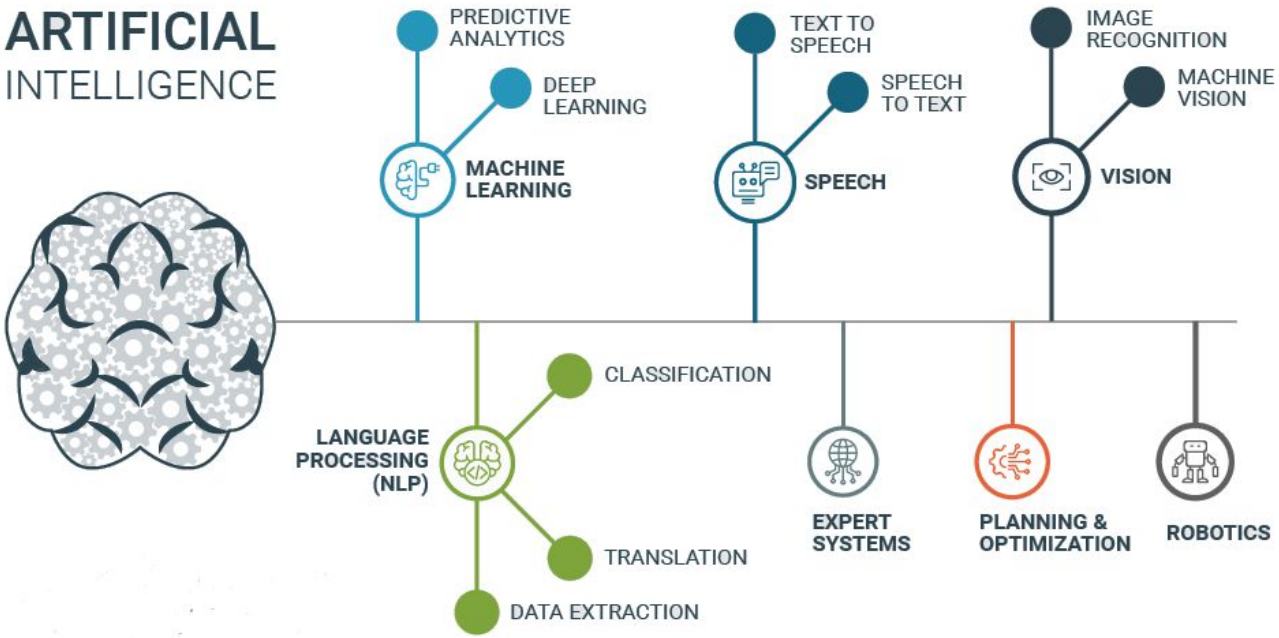
TECH-TALK	HANDS-ON	TECH-TALK	HANDS-ON	INDUSTRY CONNECT
Cognitive Computing Natural Language Processing NLP Around Us	NLP Techniques + Sentiment Analysis	CHATBOTS Why, What, How?	Building a chatbot using Python	Upskilling and staying relevant

Cognitive Computing

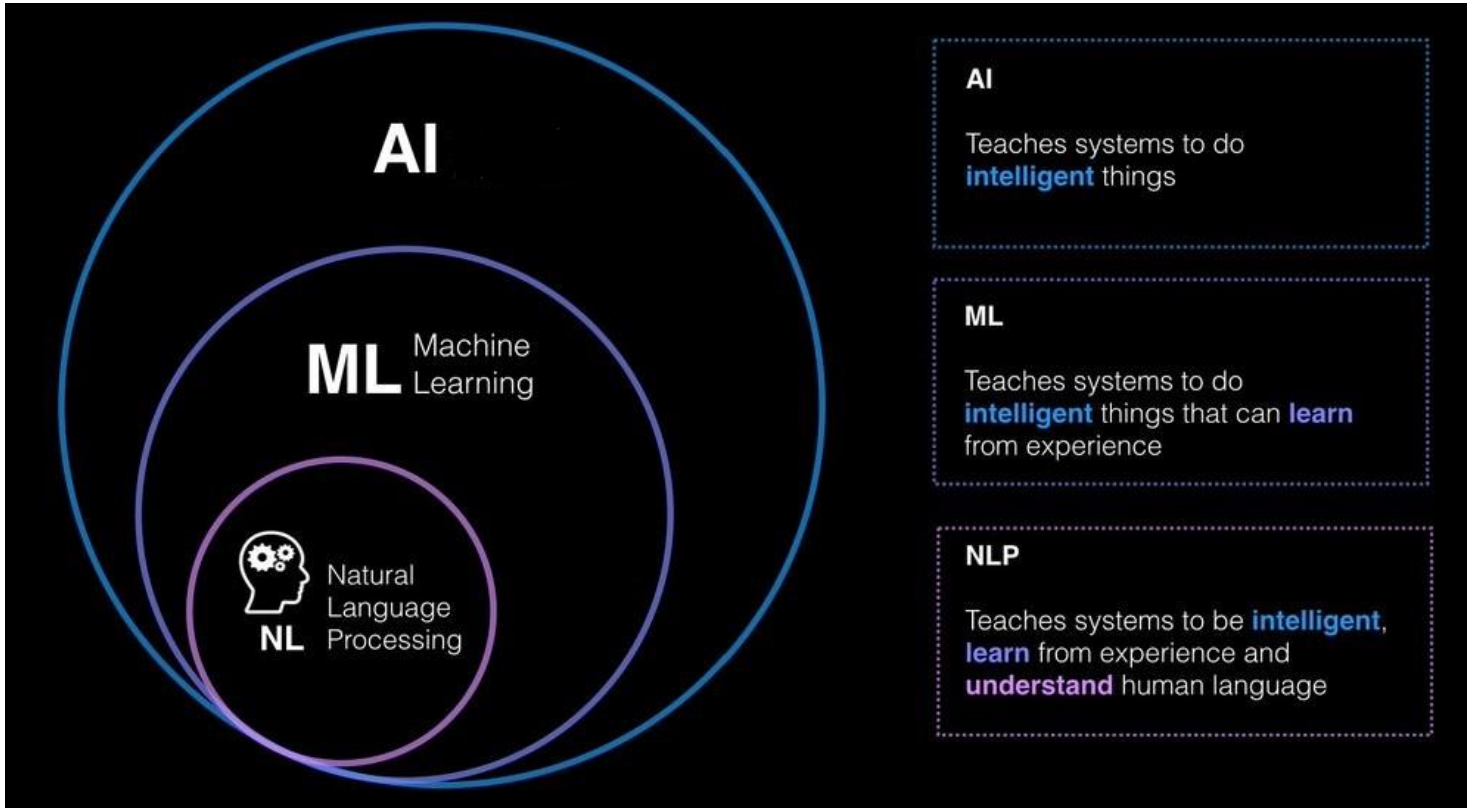
200,000 years and counting....



70 years and counting....



AI vs ML vs NLP ?

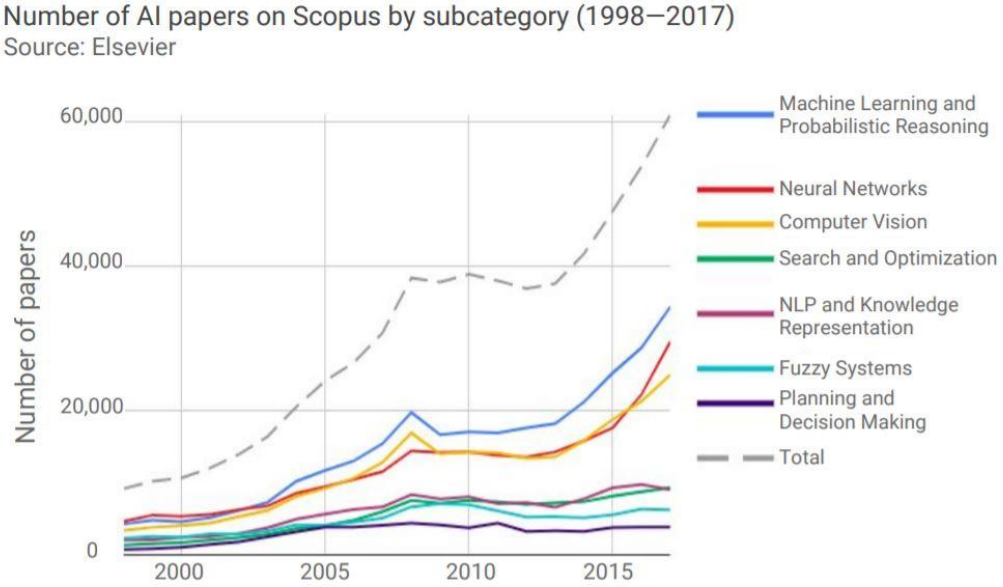


What is Natural Language Processing?

- Natural Language Processing (NLP) is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

- **Objective:** Read, decipher, understand, and make sense

- NLP = Computer Science + AI + Computational Linguistics



What makes NLP so hard?

Standard \ Nonstandard
Examples

Standard	Nonstandard
• am not, is not, has not	• ain't
• very good	• cool
• very	• damn
• to play a trick	• to pull one's leg
• you all	• ya'll

"jaguar" can refer to a car or to an animal (Disambiguation)

I Scream vs Ice Cream (Phonetics)

AMBIGUITY

How we deal with text data.

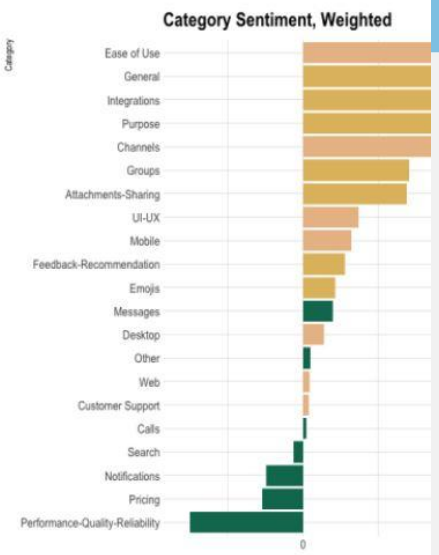
“The key to NLP is data”

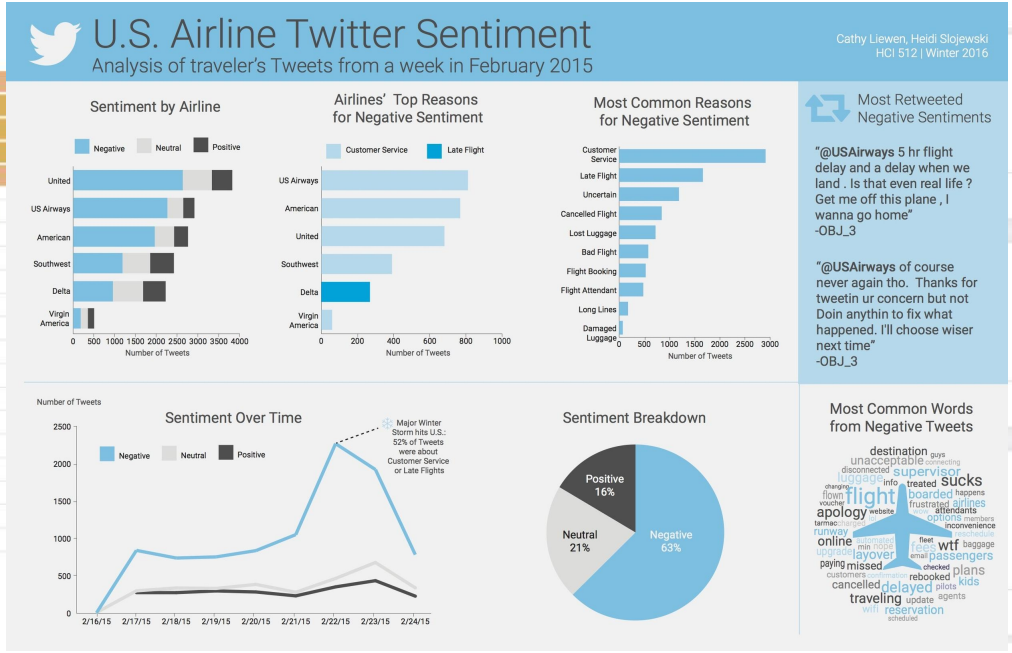
The more data you collect, the more you can correct your algorithm’s mistakes, and reinforce its correct answers. With unlimited data and unlimited compute, we would have perfect NLP today.

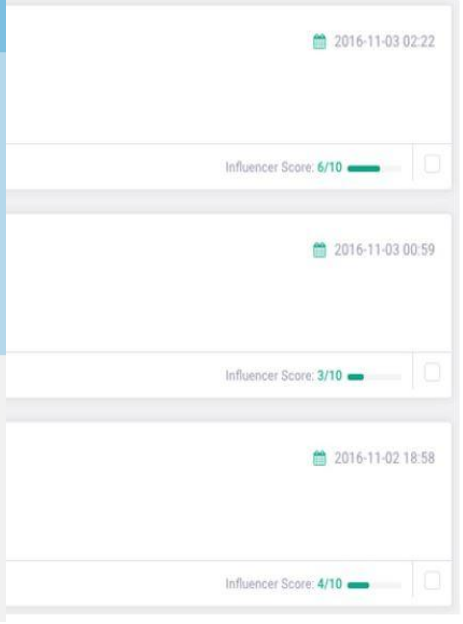
NLP is all around us

Sentiment Analysis

- **3.5 billion people** are active social media users - 45% of the world's population
- Every single minute of the day, **500,000 Tweets** and **510,000 Facebook comments**
- Checked Zomato before dining out? Read amazon reviews before buying?

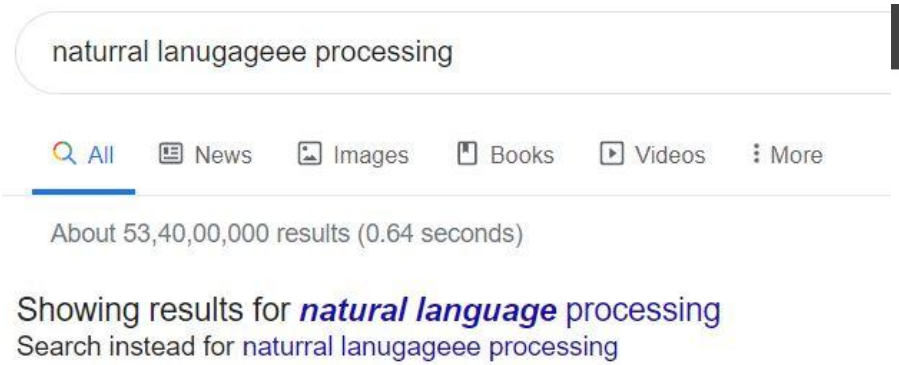


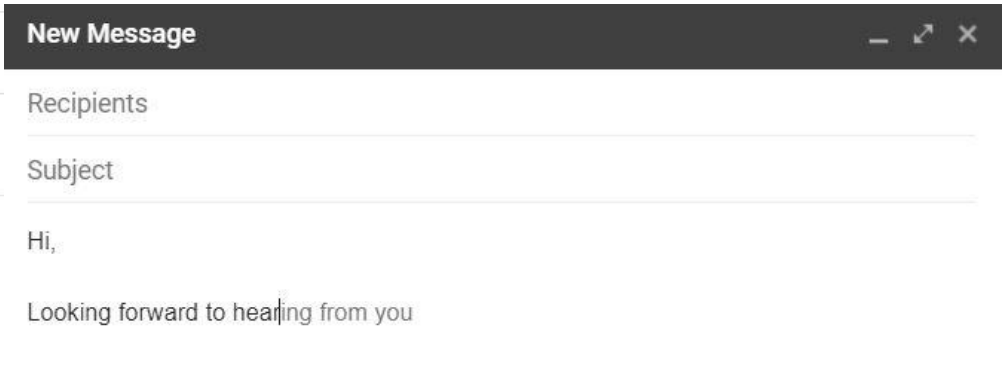




Spell Check / Auto Complete

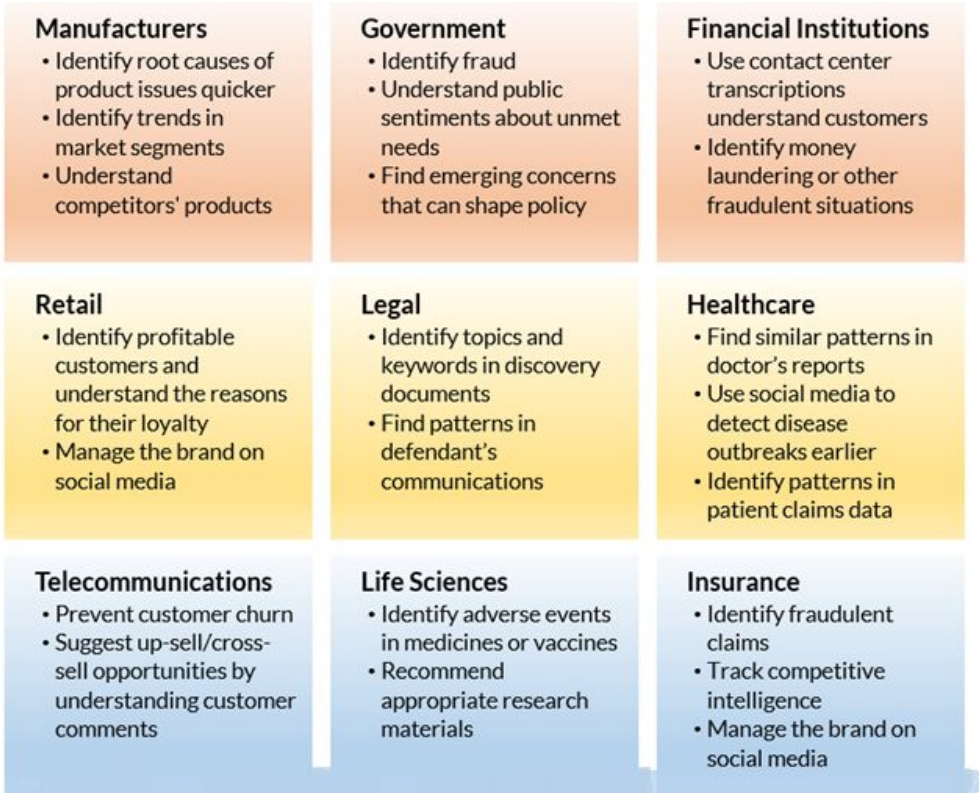
- ➢ Spellchecker points to spelling errors and possibly suggests alternatives.
- ➢ Autocorrector automatically picks the most likely word
- ➢ Auto complete: Understands context + grammar + provides suggestions
- ➢ Levenshtein Distance Algorithm - Insertion, Deletion, Substitution (Edit Distance)
- ➢ Python Packages: Symspell vs Jamspell vs KenLM
- ➢ Grammarly, Google Doc, Plagiarism Detection





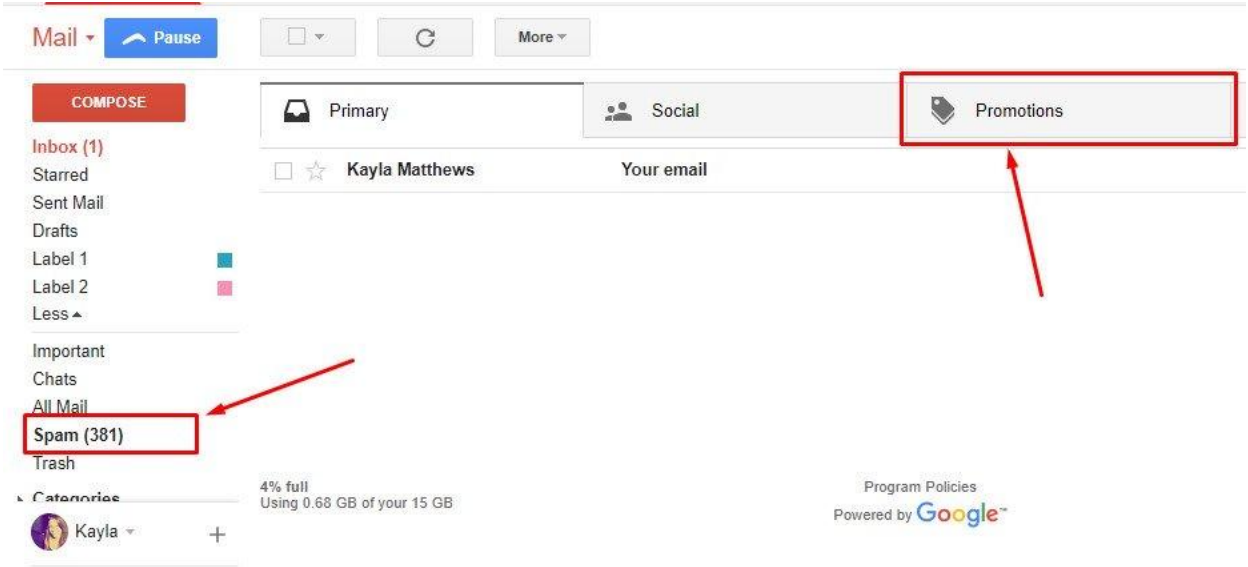
Text Mining

- Process of exploring sizeable textual data and find patterns generate valuable insights, enabling companies to make data-driven decisions
- Unstructured Data to Structured Data
- Packages: PyPDF, pydocx
- Information Retrieval
- Summarization - Semantic Map



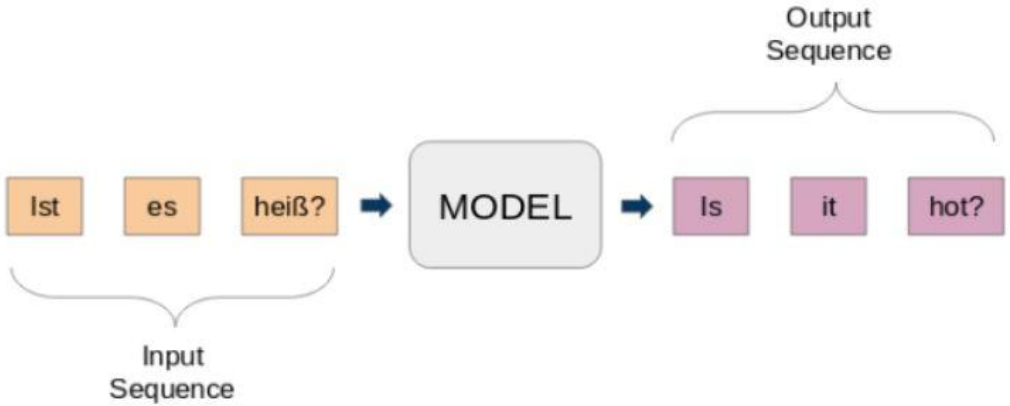
Text Classification

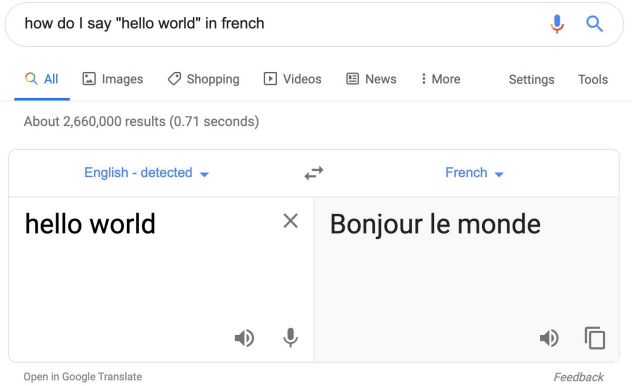
- Process of assigning tags or categories to text according to its content
- Document Classification, Sentence Classification
- Customer Support, Sentiment Analysis, Content Recommendation
- Python Packages: Fasttext, Flair



Machine Translation

- Translating one source language or text into another language
- Interpret & analyze all of the elements in the text and how each word may influence another. Consider grammar, syntax (sentence structure), semantics (meanings) in the source and target languages, as well as familiarity with each local region.
- Concepts of Deep Neural Networks - Recurrent Neural Networks - LSTM
- Seq2seq: Encoder ~ Decoder





Python Packages


Natural
Language
ToolKit
python


For NLP Preprocessing
TextBlob


scikit
learn


flair

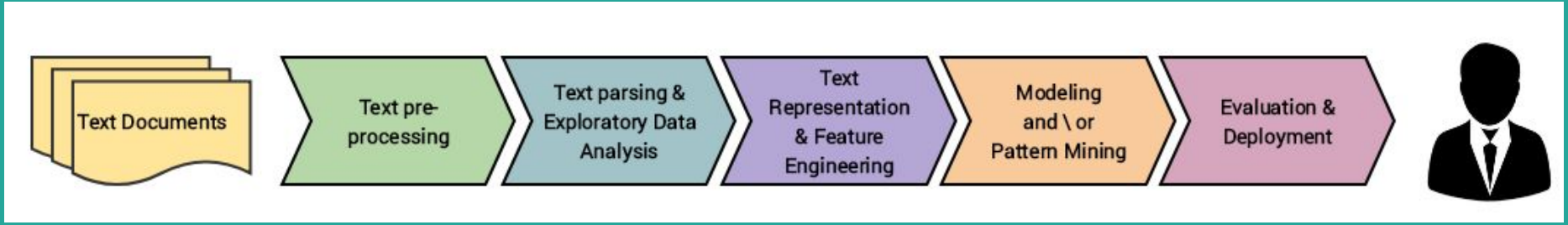

fastText


spaCy


GENSIM
topic modelling for humans


Stanford University
S
NLP
Natural Language Processing

Lifecycle


Text Documents
Text pre-processing
Text parsing & Exploratory Data Analysis
Text Representation & Feature Engineering
Modeling and \ or Pattern Mining
Evaluation & Deployment

“A machine learning model is only as good as the data it is fed.”

Text Preprocessing

Text Normalization

NLP is very Interesting to learn!!!

- Converting all letters to lower or upper case - str.tolower() / str.toupper()
- Converting numbers into words or removing numbers - replace/regex
- Removing punctuations, accent marks and other diacritics - regex
- Removing white spaces - str.strip()
- Expanding abbreviations - str.replace()

Natural language processing is very interesting to learn

Stopword Removal

- Common language articles, pronouns and prepositions such as “and”, “the” or “to”
- Has no impact towards NLP
- There is no Universal list of stopwords → Domain specific
- Beware of NOTs

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Tokenization

- count all words in a piece of text
- occurrence matrix for the sentence or document, disregarding grammar and word order
- word frequencies or occurrences are then used as features for training a classifier.

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"
print(word_tokenize(data))
```

['All', 'work', 'and', 'no', 'play', 'makes', 'jack', 'dull', 'boy', ',', 'all', 'work', 'and', 'no', 'play']

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack dull boy. All work and no play makes jack a dull boy."
print(sent_tokenize(data))
```

['All work and no play makes jack dull boy.', 'All work and no play makes jack a dull boy.']

- End up removing punctuations: Dr. → Dr
- Hyphens? Parenthesis? → PROBLEM!!

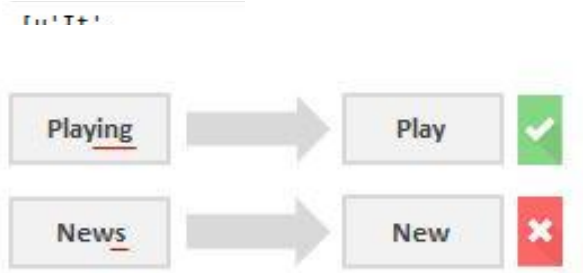
Stemming

- Reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language
- Integral to search queries and information retrieval
- Stem words, sentences, documents
- English Stemmers, Non English Stemmers
- Shorten the lookup & normalize

```
# stemming
from nltk.stem.porter import PorterStemmer
porter = PorterStemmer()

def tokenizer_porter(text):
    return [porter.stem(word) for word in text.split()]

tokenizer_porter('It has become appallingly obvious \
                  that our technology has exceeded our humanity.')
```



Lemmatization

- Reduces the inflected words properly ensuring that the root word belongs to the language. Root word is called Lemma.
- Consideration the context of the word
- NLTK: WordNetLemmatizer built on WordNet Database
- Part-of-speech parameter to a word (whether it is a noun, a verb, and so on) it's possible to define a role for that word in the sentence

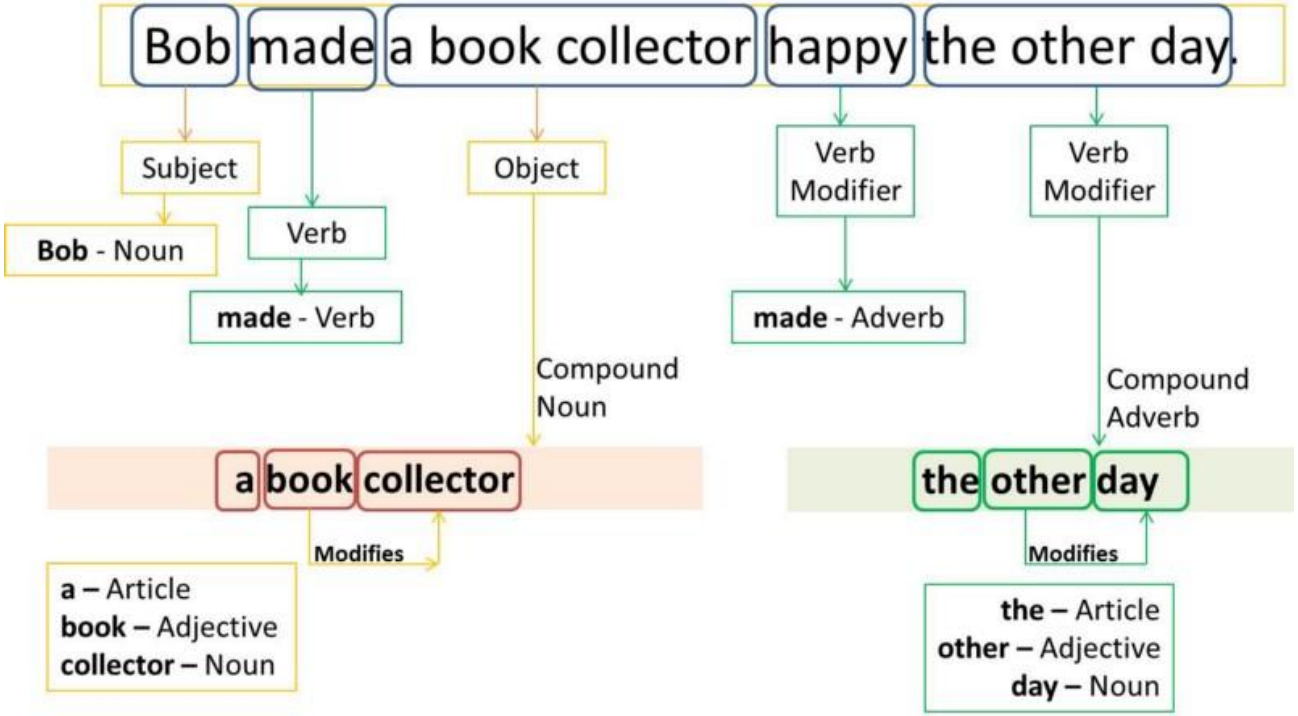
```
# import these modules
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print("rocks :", lemmatizer.lemmatize("rocks"))
print("corpora :", lemmatizer.lemmatize("corpora"))
```

POS Tagging

- Part-of-speech tagging aims to assign parts of speech to each word of a given text (such as nouns, verbs, adjectives, and others) based on its definition and its context



N-Grams

- n-gram is a contiguous sequence of n items from a given sample of text or speech.
- ngram model models sequence, i.e., predicts next word (n) given previous words (1, 2, 3, ..., n-1). multiple gram (bigram and above) captures context
- **Rule of thumb:** trigram is a common choice with large training corpora (millions of words), whereas a bigram is often used with smaller ones.

FEATURE ENGINEERING

Bag of Words/One Hot Encoding/Document Term Matrix

- count all words in a piece of text
- occurrence matrix for the sentence or document, disregarding grammar and word order
- word frequencies or occurrences are then used as features for training a classifier

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

- Absence of semantic meaning & context
- What if “the”, “a” are the most repeated?
- **Solution:** Stop word + TF-IDF

Term Frequency - Inverse Document Frequency

- Numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- TF = Bag of Words; IDF = Measures the importance of the term across the document. Closer to 0, the more common the word is
- TF–IDF = TF * IDF (Float value)
- 83% of text-based recommender systems in digital libraries use tf–idf.

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{N}{\mathrm{df}_i}\right)$$

tf_{ij} = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	**0.48**	0.18							
Doc 2	0.18		0.18	**0.48**	0.18	0.18				
Doc 3					0.18	0.18	**0.48**	**0.95**	**0.48**	**0.48**

Word Vectors

- Convert input into numerical form that deep neural network can process as inputs

“How to plot dataframe bar graph”

- CBOW - Features: {how,to,plot,bar,graph}, Predict: {dataframe}
  - Predicts current word with the help of neighbouring words
- Skip Gram: Features: {dataframe}, Predict: {how,to,plot,bar,graph}
  - Predicts neighbouring words based on current word

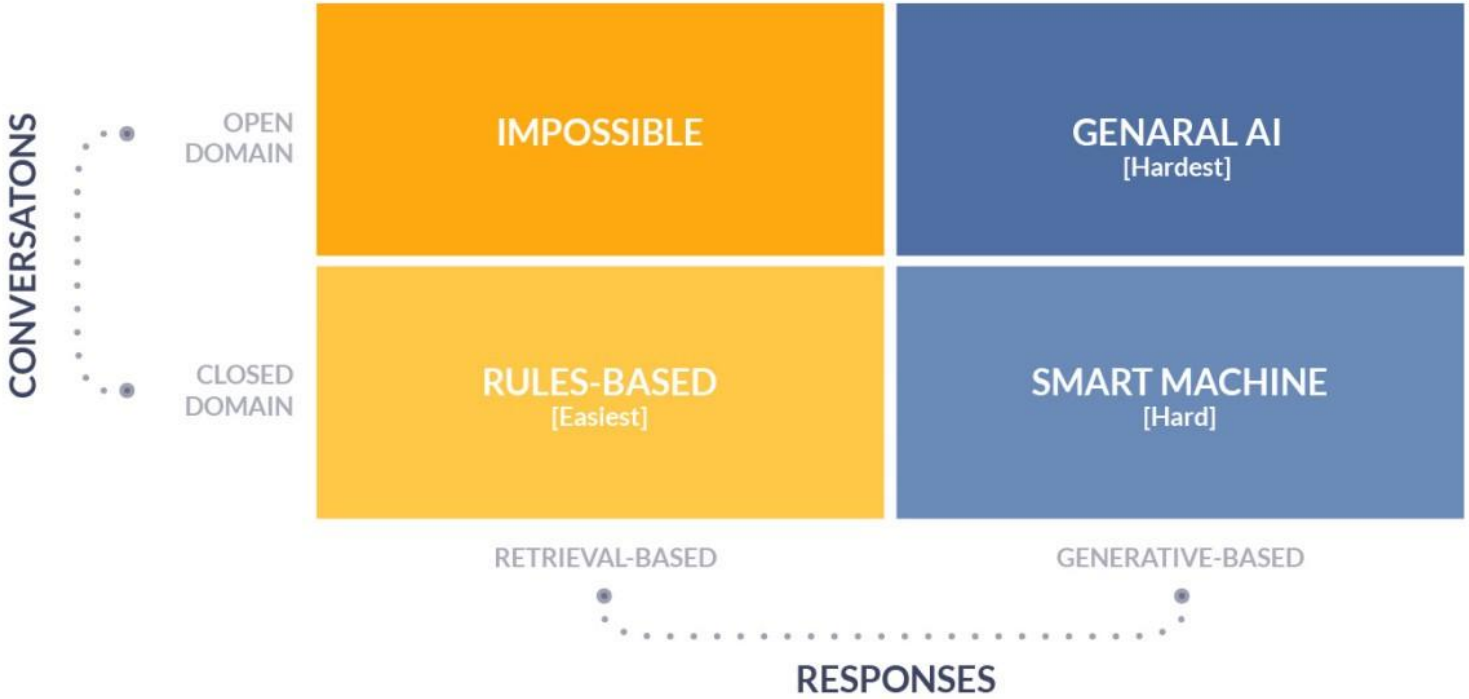
TUTORIAL

CHATBOTS

Chatbots - What, Why, Who, How?

- Artificial intelligence (AI) software that can simulate a conversation with a user
- FAQ/Simple Information Lookup
  - User: Asks Question, Chatbot: Gives Answers, No Machine Learning
- Guided Conversation Form
  - Chatbot: Asks Question, User: Gives Answers, Machine Learning → Extract Entities
- Information Lookup
  - User: Asks Question, Chatbot: Gives Answers, Information from DB/ Web services
  - Machine Learning → Extract Entities
- Smart Home / Simple Information Lookup
  - Performs Predefined Tasks

CHATBOT CONVERSATION FRAMEWORK



Terminologies


Utterance
"Show me yesterday's financial news"
Entity
Entity
Intent: showNews
Verb
Noun

Utterances, Intents, Entities

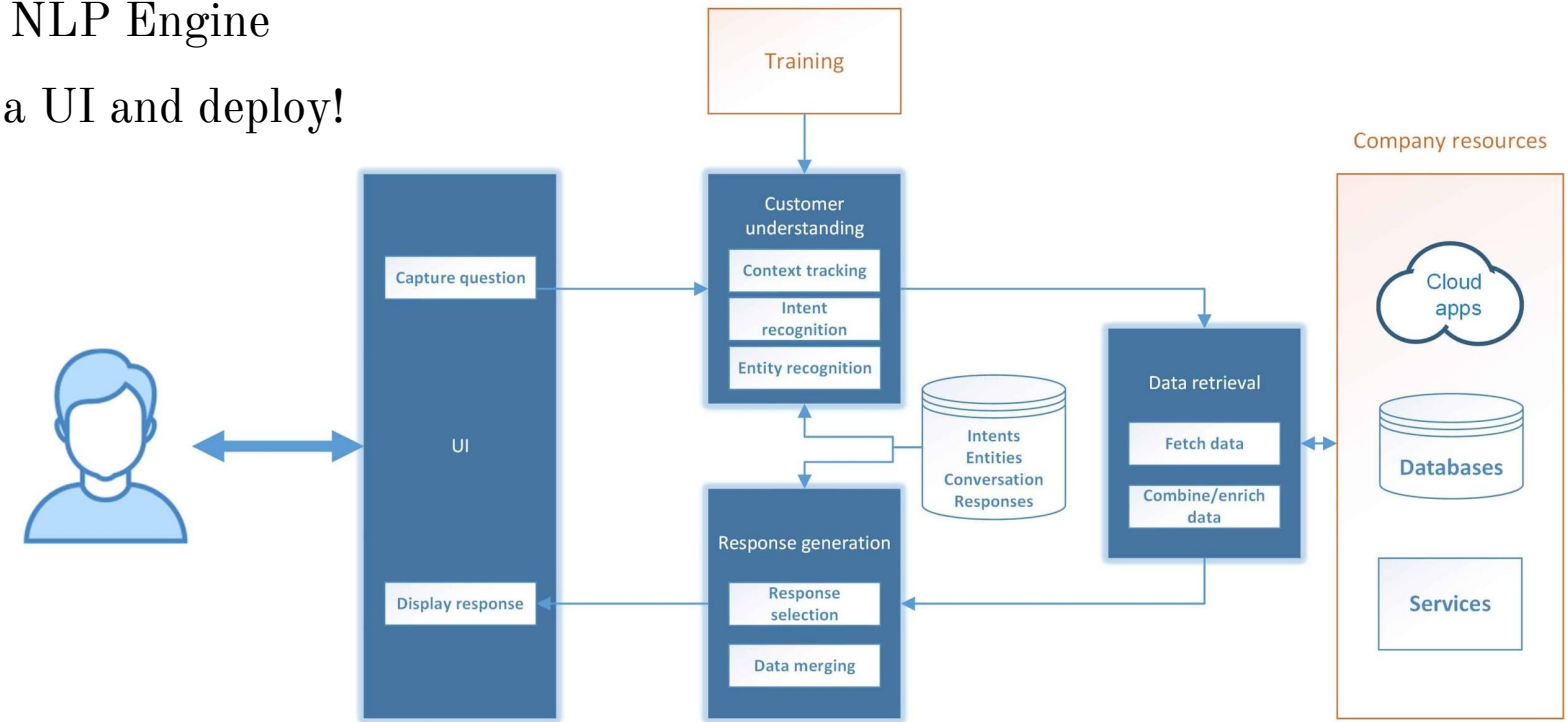
- Utterances: Query Entered by the user
- Intents: An intent is the user's intention
  - Fasttext: Facebook AI Research Lab
  - Intent Classification: Greeting, Schedule, Resource, Thanking
- Entities: An entity modifies an intent
  - SpaCy: Named Entity Recognition

PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)

LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.

Steps to build a chatbot

- Build corpus / knowledge base
- Intents, entities, actions
- Train NLP Engine
- Build a UI and deploy!



amazon Google Microsoft SAMSUNG

amazon alexa

Google ASSISTANT

Cortana

Siri

Bixby

Let’s build a FDP ML chatbot!

> NLP made it possible for machines to become our friend.

BILL GATES
MICROSOFT CO-FOUNDER
AT HUNTER COLLEGE IN NYC

INDUSTRY CONNECT

Staying relevant

Meetups

Chennaipy

YDPY

Conferences

PYCON INDIA Chennai, 2019

Open Source Contribution

Hackathons

SIH

SMART INDIA HACKATHON 2020

Google Summer of Code

Women in Tech

pyladies CHENNAI

QUESTIONS

Want to explore further? Let's connect!

linkedin.com/in/niharikakrishnan

@Nihaaarika

niharikakrishnan

Slide Deck: https://github.com/niharikakrishnan/Talks