NLP-101

Introduction to Natural Language processing

Srikanth Phalgun Pyaraka
(Consultant - Data Analytics)
07/07/2020

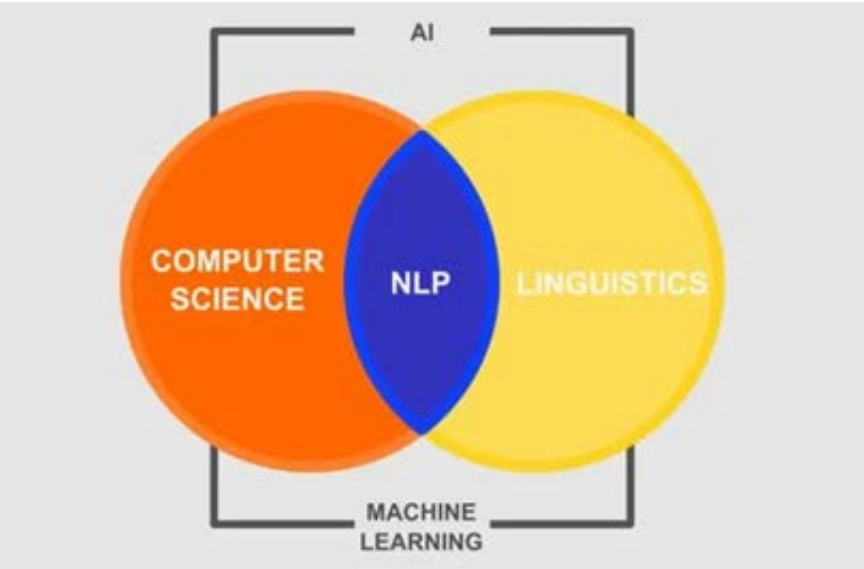
Agenda

- ✓ What is Natural language ?
- ✓ What is NLP?
- ✓ Why NLP?
- ✓ NLP Applications
- ✓ NLP Workflow
- ✓ Hands-on
- ✓ References
- ✓ Q&A

What is Natural Language ?

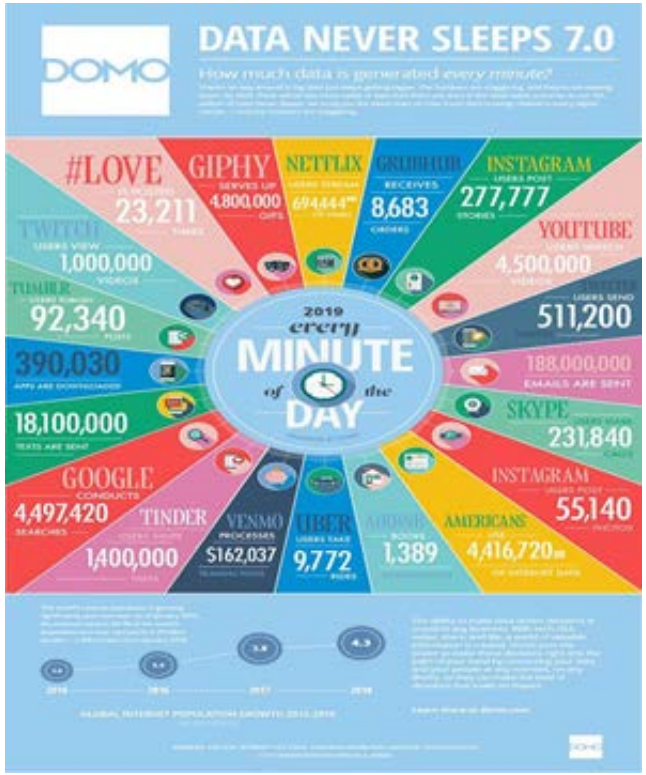
- Mechanism used for Human interaction.
- Not like programming language.
- Highly unstructured in nature – Text & Speech.
- Very Difficult to parse and comprehend by machines.
- Natural language can be communicated in different ways, like speech, writings or even using signs.

What is NLP?



- NLP is area of research in computer science, linguistics, and AI.
- NLP is concerned with interaction between computers and human (natural) language.
- Key focus is to train computers to process, analyse, and model large amounts of natural language data.
- NLP flow often referred as pipeline
- Goal for NLP is make computers understand natural language in order to perform useful tasks.

Why NLP is important?



- Handling large volume of text data.
- 80% of enterprise relevant information originates in unstructured form.
- Structuring highly unstructured data source.
- NLP models are realistic and affordable.
- Most powerful applications uses NLP.
  - Google Translate (Machine Translations)
  - Google Assistant/Alexa(Speech Recognition)
  - Google search
  - Chat bots providing better service experience

NLP Pyramid(Morphology)

Morphology:

- Analyse how words are formed/originated.
- Most of operations are at word level.
- Word is viewed as sequence of characters
- NLP tasks at word level:
  - Prefixes/Suffixes
  - Singularization/Pluralization
  - Gender detection
  - Word inflection
  - Lemmatization(Base form of word)
  - Spell checking etc

NLP Pyramid (Syntax)

Syntax:

- Analyse how sentences which are formed are grammatically valid.
- Most of operations are at sentence level.
- Sentence is viewed as sequence of words.
- NLP tasks at sentence level:
  - Parts-of-speech tagging (Assigning tags to the words like Noun/Verb/Adj etc.)
  - Building syntax trees.
  - Building Dependency trees.

NLP Pyramid (Semantics)

Semantics:

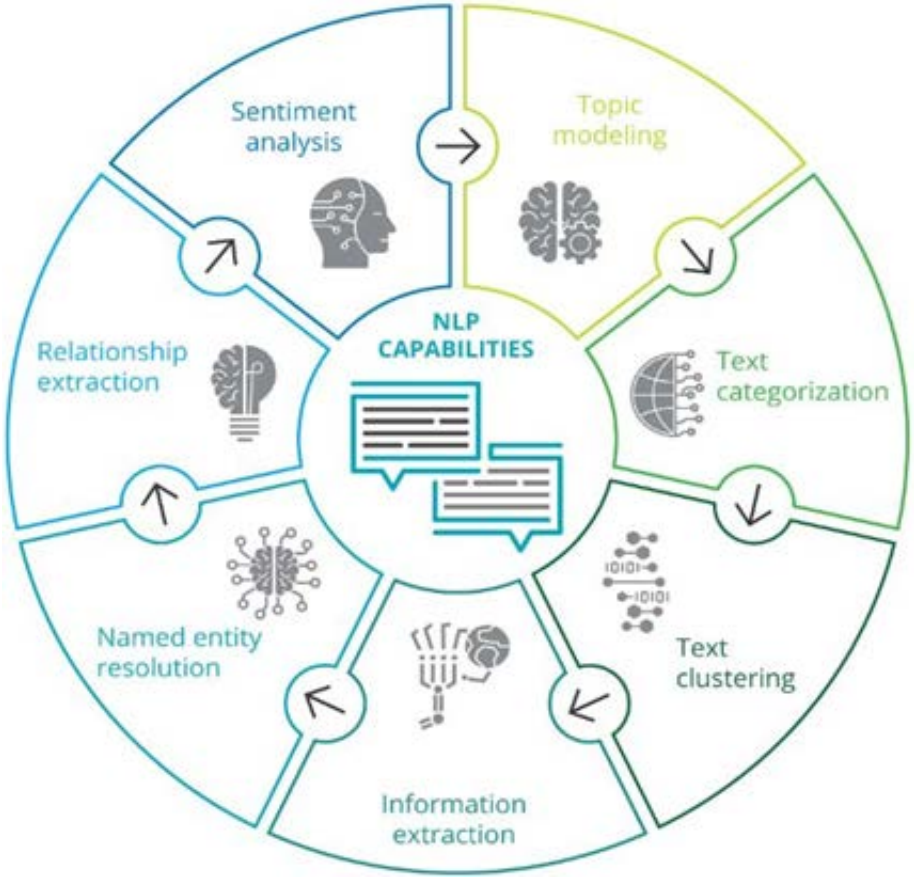
- It derives meaning from text.
- This branch deals with actual understanding of natural language.
- Most of operations are at text level.
- Text is viewed as sequence of sentences.
- NLP tasks at Text level:
  - Named entity extraction
  - Relation extraction
  - Semantic Role labelling
  - Word sense disambiguation

NLP Pyramid (Pragmatics)

Pragmatics:

- It analyses text as whole.
- It studies the way in which context contributes to meaning.
- It usually works on text with contexts.
- NLP tasks at pragmatic level:
  - Coreference/Anaphora resolution (find out what word refers what. E.g John is fine. He [John] in no danger.
  - Topic segmentation
  - Lexical chains
  - Summarization

NLP Applications



Common NLP Use cases

- **Text Classification**

  Support Ticket Classification, News Article Categorization

- **Text Clustering & Similarity**

  Recommender systems, Duplicate Detection with Fuzzy Matching

- **Search and Information Retrieval**

  Search Engines, Document Ranking

- **Parsing and Named Entity Recognition**

  Entities from health records, legal documents

- **Text Summarization**

  Topic models, summarizing entire documents

- **Machine Translation**

  Speech to Text, Language Translation

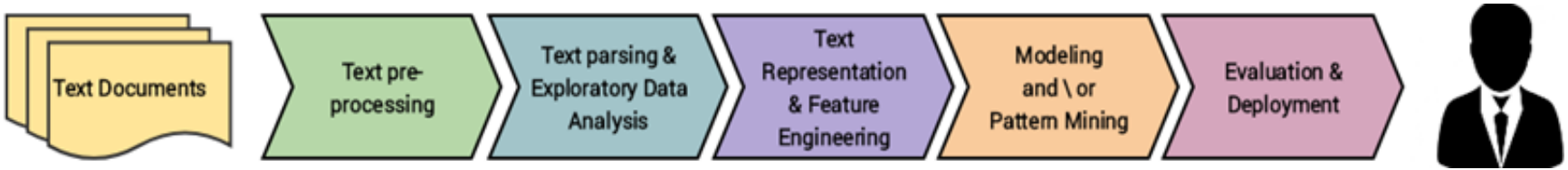
- **Conversational Interfaces**

  Chatbots, Personal Assistants, Q&A Systems

- **Sentiment Analysis**

  Survey result analysis, NPI analysis

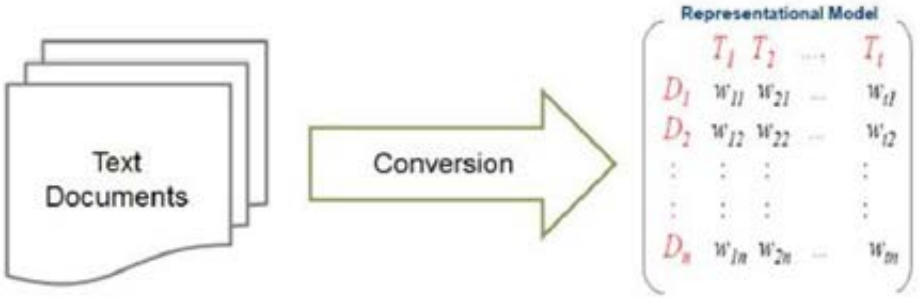
Standard NLP work flow



Text wrangling / Pre processing

- Removing HTML tags
- Remove Extra Whitespace and Newlines
- Remove special characters and symbols (optionally numbers)
- Convert accented characters to ASCII
- Stemming OR Lemmatization
- Removing Stop Words
- Tokenization if needed
- Spell Check & Grammar Check

Text Representation Model



- ML\DL models at heart are mathematical functions and cannot understand unstructured text
- Hence we need to convert text into some numeric representations which can be understood by machines
- Commonly known as Vector Space Models where text is converted to numeric vectors
  - Bag of Words, TF-IDF
  - Topic Models
  - Similarity
  - Word Embeddings - Word2Vec, GloVe, FastText etc.

Traditional Text Representation Model

1 Bag of Words

- Each document is represented by a vector (bag) of words
- Depicts the number of times each word occurs in that document

2 Bag of N-grams

- Same as the Bag of Words model
- Instead of words, we also have n-grams and counts for them in the vector

3 TF-IDF

- Similar to the basic Bag of Words (TF) model
- Normalizes counts using the inverse document frequency (IDF) to downplay effect of frequently occuring words

4 Document Similarity

- Derived attribute \ feature from bag of words based features
- Assign scores to each document w.r.t how similar it is to other documents (based on their BOW vectors)

Hands-on : Movie Recommender System

- Get Movie Dataset
- Clean Movie Descriptions
- Build TF-IDF Features per Movie Description
- Compute Document Similarity (pairwise)
- Recommend Similar Movies based on Movie Description

Do you want to try ??

bit.ly/31O7L3N

References :

- nlpforhackers.io
- https://machinelearningmastery.com/natural-language-processing/
- https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72
- https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/

Q&A

Thank You

www.linkedin.com/in/srikanthpyaraka