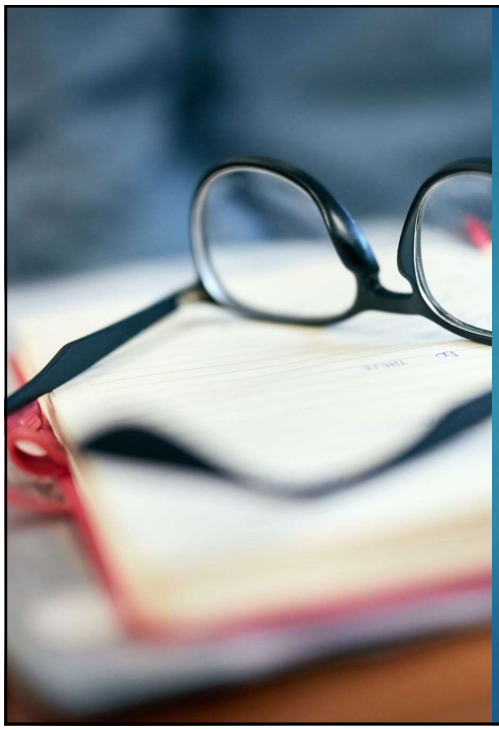
Natural Language Processing

DR FAKHRELDIN SAEED

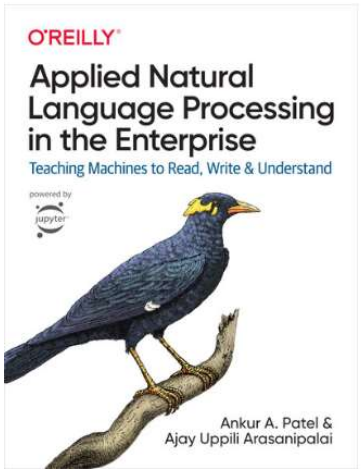
1



Content

- Introduction
- Application
- Python NLP libraries
- NLP tasks
- Pretrained Word Embeddings
- Sequential and Transformer models

2



TIME TO COMPLETE:
8h 45m

TOPICS:
Natural Language Processing

PUBLISHED BY:
O'Reilly Media, Inc.

PUBLICATION DATE:
May 2021

PRINT LENGTH:
333 pages

Reference

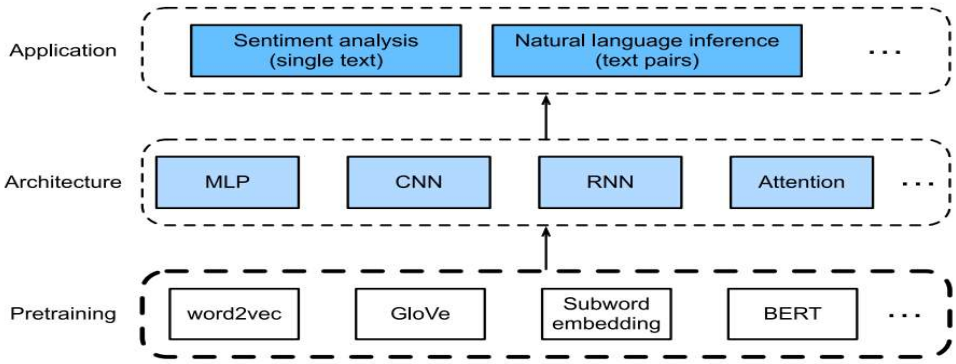
APPLIED NATURAL LANGUAGE PROCESSING IN THE ENTERPRISE (OREILLY.COM)

3

Introduction to NLP

- Natural language processing (NLP) is a subfield of **linguistics**, **computer science**, **information engineering**, and **artificial intelligence** concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data.
- **Challenges** in natural language processing frequently involve **speech recognition**, natural language **understanding**, and natural language **generation**.
- NLP teaches computers to process and analyse natural language data in order to perform tasks such as machine translation, sentiment analysis, natural language generation, and so on.

4



NLP roadmap

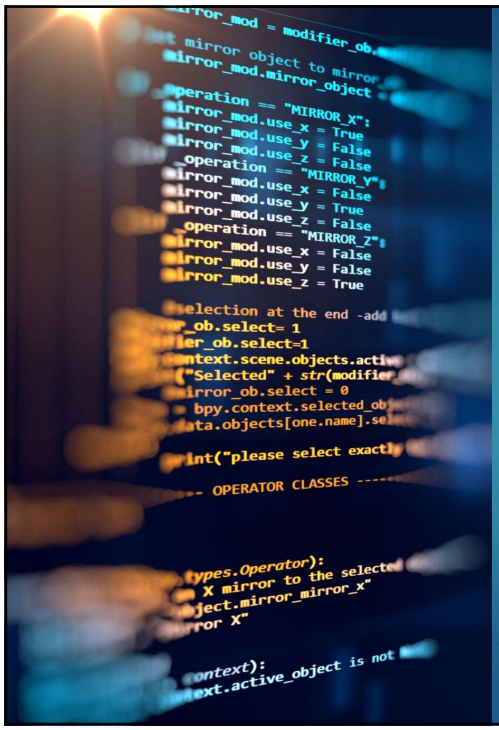
Source: https://d2l.ai/

5

Popular Applications

Sentiment Analysis	Document Summarization	Named Entity Recognition	Caption Generation
Text Classification	Machine Translation	Question Answering	Speech recognition
Chatbots and Voicebots	Text-to-speech and speech-to-text conversion	Text and audio generation	Information extraction

6



Python NLP libraries

- Natural Language Toolkit (NLTK)
- spaCy
- fast.ai
- Hugging Face
- PyNLPl
- Stanford CoreNLP
- Scikit-learn
- Pattern
- Textblob

7

NLP tasks

- In order to build NLP applications, we must master the NLP tasks that serve as building blocks for those applications.
  - **Tokenization** is the process of splitting text into minimal meaningful units such as words, punctuation marks, symbols, etc.
  - **Part-of-speech** (POS) tagging is the process of assigning word types to tokens, such as noun, pronoun, verb, adverb, adjective, conjunction, preposition, interjection, etc.
  - **Dependency parsing** involves labelling the relationships between individual tokens, assigning a syntactic structure to the sentence.
  - **Chunking** involves combining related tokens into a single token, creating related noun groups, related verb groups, etc.
  - **Lemmatization** is the process of converting words into their base forms.
  - **Stemming** is a process related to lemmatization, but simpler. Stemming reduces words to their word stems.
  - **Named entity recognition** (NER), is the process of assigning labels to known objects (or entities) such as person, organization, location, date, currency, etc.
  - **Entity linking** is the process of disambiguating entities to an external database, linking text in one form to another.

8

Pretrained Word Embeddings

- The first steps in NLP is tokenization, while Learning how to represent each token is generally the second step.
- This process is called learning word embeddings. (i.e., word vectors)
- Moreover, the word embeddings trained by **Word2Vec**, **GloVe**, and **fastText** store semantic information for each word, unlike one-hot encoding.
- Words such as "queen" and "king" have vectors that are closer together in space, implying that there is some semantic relationship/similarity between the two.

9

Sequential and Transformer models

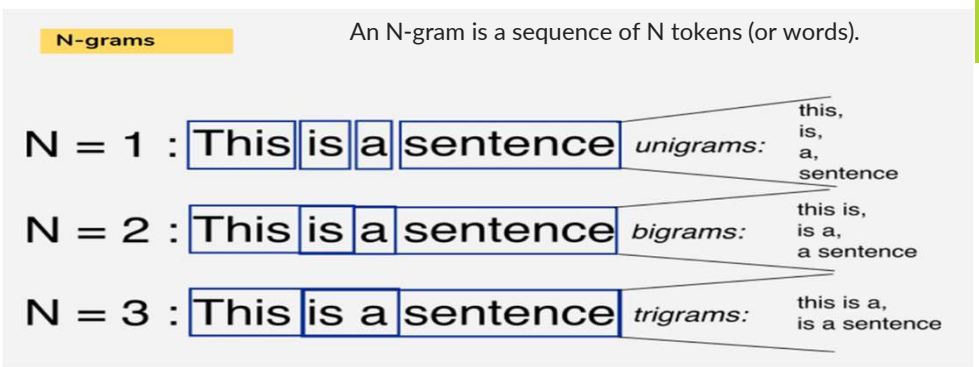
- **Sequential models** are machine learning models that input or output sequences of data, such as text, audio, and time series data.
  - Include RNNs, LSTMs, and GRUs.
- **Attention mechanisms** are a type of neural network component that allow a model to dynamically weight and combine different input elements, based on the task at hand
- **Attention mechanisms** allow the model to focus on relevant parts of the input while processing it.
- **Transformer model** uses a type of attention mechanism called self-attention, which allows the model to directly model relationships between different input elements without the need for recurrence or convolutions.
- **Universal Language Model Fine-Tuning:** first pre-train on a large dataset of unannotated text (Wikipedia), and then fine-tune this pre-trained model on a smaller dataset for a specific task. This allows the model to benefit from the general knowledge and language understanding learned during the pre-training phase, while also adapting to the specifics of the target task.
  - ELMo, BERT, BERTology, GPT-1, GPT-2 and GPT-3

10

Language Model

- A language model is trained to predict the likelihood of a sequence of words.
- An important component in many natural language processing (NLP) systems, such as **machine translation, summarization, and question answering**.
- Types of Language Models
  - Probabilistic language models
    - n-gram modelling
  - Neural language models
    - feed-forward
    - RNNs and LSTMs
    - Transforms

11



N-gram modelling

12

Reading

- 14. Selected Topics in Natural Language Processing | Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow (oreilly.com)