Artificial Intelligence (Machine Learning & Deep Learning) [Course]

Week 6 – Natural Language Processing

[See examples / code in GitHub code repository]

It is not about Theory, it is 20% Theory and 80% Practical – Technical/Development/Programming [Mostly Python based]

NLP | What is Natural Language Processing (NLP)

Natural Language Processing (NLP) enables computers to understand and interpret human language. While computers excel at processing structured data, such as spreadsheets or databases, natural language in its unstructured form (text, speech, etc.) presents a unique challenge. NLP bridges this gap by allowing machines to process and understand human languages, making it an essential tool in modern AI systems.

Natural language processing (NLP) is the discipline of building machines that can manipulate human language — or data that resembles human language — in the way that it is written, spoken, and organized. It evolved from computational linguistics, which uses computer science to understand the principles of language, but rather than developing theoretical frameworks,

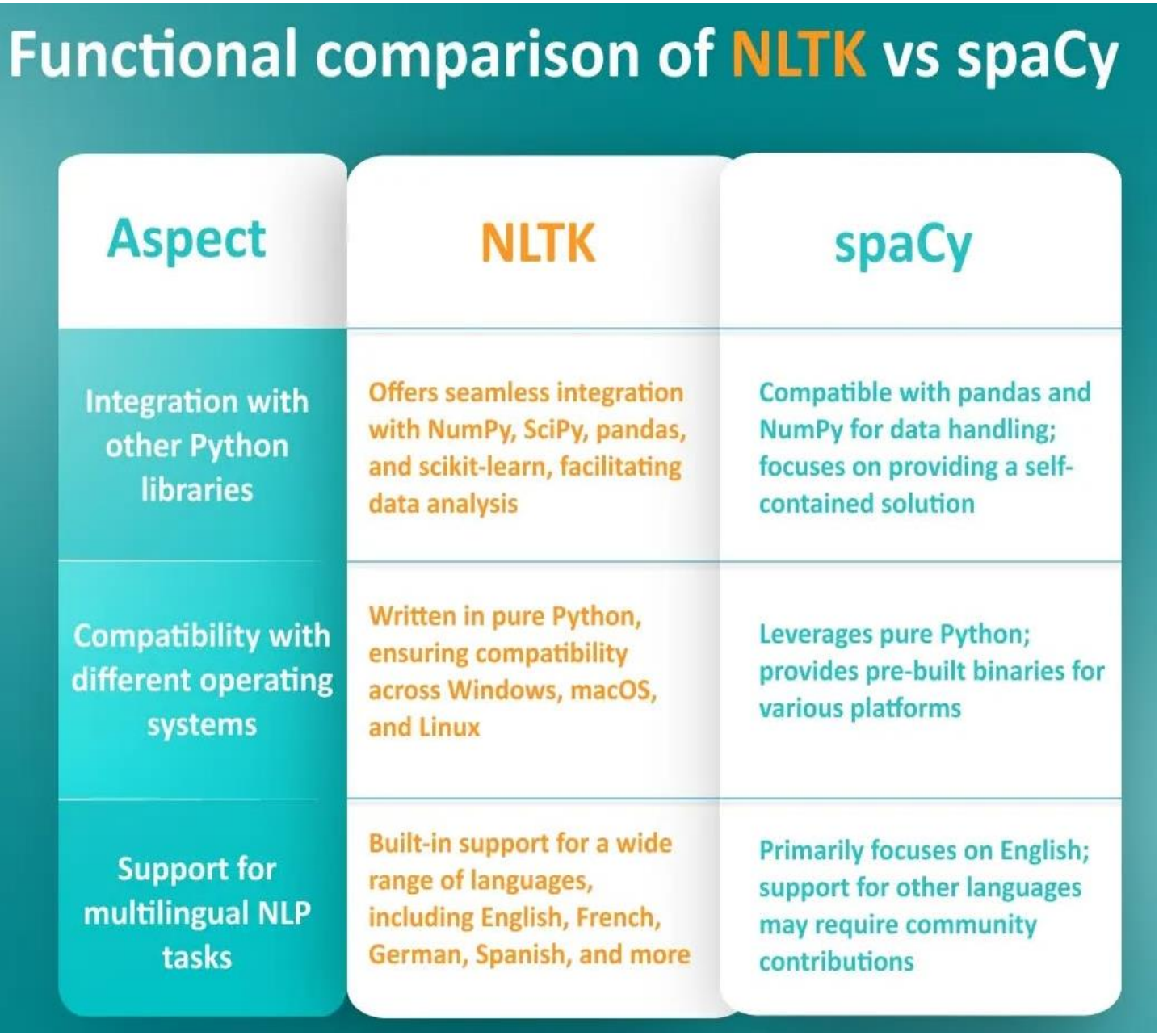
NLP can be divided into two overlapping subfields: **Natural Language Understanding (NLU)**, which focuses on semantic analysis or determining the intended meaning of text, and **Natural Language Generation (NLG)**, which focuses on text generation by a machine. NLP is separate from — but often used in conjunction with — speech recognition, which seeks to parse spoken language into words, turning sound into text and vice versa.

Major tasks in natural language processing are speech recognition, text classification, natural-language understanding, and natural-language generation

Reference:

https://www.geeksforgeeks.org/introduction-to-natural-language-processing/
https://www.deeplearning.ai/resources/natural-language-processing/
https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_introduction.htm

NLP | NLTK vs SpaCy



Reference:

https://www.seaflux.tech/blogs/NLP-libraries-spaCy-NLTK-differences
https://medium.com/@prabhuss73/spacy-vs-nltk-a-comprehensive-comparison-of-two-popular-nlp-libraries-in-python-b66dc477a689
https://konfuzio.com/en/spacy-vs-nltk/

NLP | Core Foundation Elements - NLP Pipeline

An **NLP (Natural Language Processing) pipeline** is a sequence of processing steps that take raw text as input and convert it into a format suitable for analysis or downstream applications (e.g., search engines, chatbots, sentiment analysis).

1. Data Acquisition
2. Text Preprocessing
3. Feature Engineering
4. Modelling
5. Evaluation
6. Deployment

Points to remember:

- ❑ This pipeline is not universal.
- ❑ This is ML pipeline and deep learning pipelines are slightly different.
- ❑ NLP pipeline is non-linear (that means stages can have more dynamic connections, allowing for branching and iteration).

Resources:

https://medium.com/@asjad_ali/understanding-the-nlp-pipeline-a-comprehensive-guide-828b2b3cd4e2
https://www.restack.io/p/natural-language-understanding-answer-nlp-pipeline-cat-ai
https://dataqoil.com/2022/07/22/natural-language-processing-pipeline/
https://monicamundada5.medium.com/basic-steps-in-natural-language-processing-pipeline-763cd299dd99
https://www.geeksforgeeks.org/natural-language-processing-nlp-pipeline/

NLP | Core Foundation Elements - NLP Pipeline

1. Data Acquisition

Data acquisition involves obtaining raw textual data from various sources to create a robust dataset for NLP tasks. It involves assessing the availability and accessibility of data, whether it's readily available, needs supplementation, or requires creation from scratch.

2. Text Preprocessing

(i) Basic Cleaning:

This initial stage focuses on eliminating irrelevant or disruptive elements from the text: [HTML Tag Removal , Handling Emojis, Basic Spell Checks]

(ii) Basic Preprocessing

- ❑ Tokenization: Segmenting text into smaller units such as words or sentences (word tokenization and sentence tokenization). This step breaks down the text into manageable chunks.
- ❑ Stop Word Removal: Eliminating common and less meaningful words (stop words) like "the," "is," etc., which don't contribute significantly to the meaning of the text.
- ❑ Stemming/Lemmatization: Reducing words to their root forms — stemming removes prefixes/suffixes, while lemmatization maps words to their base or dictionary form, aiding in standardization.
- ❑ Lowercasing: Converting all text to lowercase to ensure uniformity in text analysis, as case sensitivity can affect certain NLP tasks.
- ❑ Language Detection: Identifying the language of the text, is especially useful when dealing with multilingual content.

(iii) Advanced Preprocessing:

Part-of-Speech (POS) Tagging: Assigning grammatical categories (like nouns, verbs, adjectives) to words in the text, providing insights into the syntactic structure.

Parsing: Analyzing the grammatical structure of sentences to identify relationships between words, and determining the syntactic roles and dependencies.

NLP | Core Foundation Elements - NLP Pipeline

3. Feature Engineering

Feature engineering in Natural Language Processing (NLP) involves transforming raw text data into numerical features that machine learning models can comprehend and utilize effectively. The goal is to represent text in a format that captures semantic meaning, contextual information, and relationships between words.

4. Modelling

- ❑ Machine Learning (ML) Approaches
- ❑ Deep Learning (DL) Approaches
- ❑ Cloud APIs - SaaS AI Development Platform : MicroSoft vs Google vs Amazon vs Oracle vs IBM
  - ➢ Microsoft – Azure AI Services: https://azure.microsoft.com/en-us/solutions/ai/
  - ➢ Google AI services: https://cloud.google.com/products/ai/
  - ➢ Amazon AI Services https://aws.amazon.com/ai/services/
  - ➢ Oracle AI Services: https://www.oracle.com/artificial-intelligence/ai-services/
  - ➢ IBM AI Services: https://www.ibm.com/artificial-intelligence

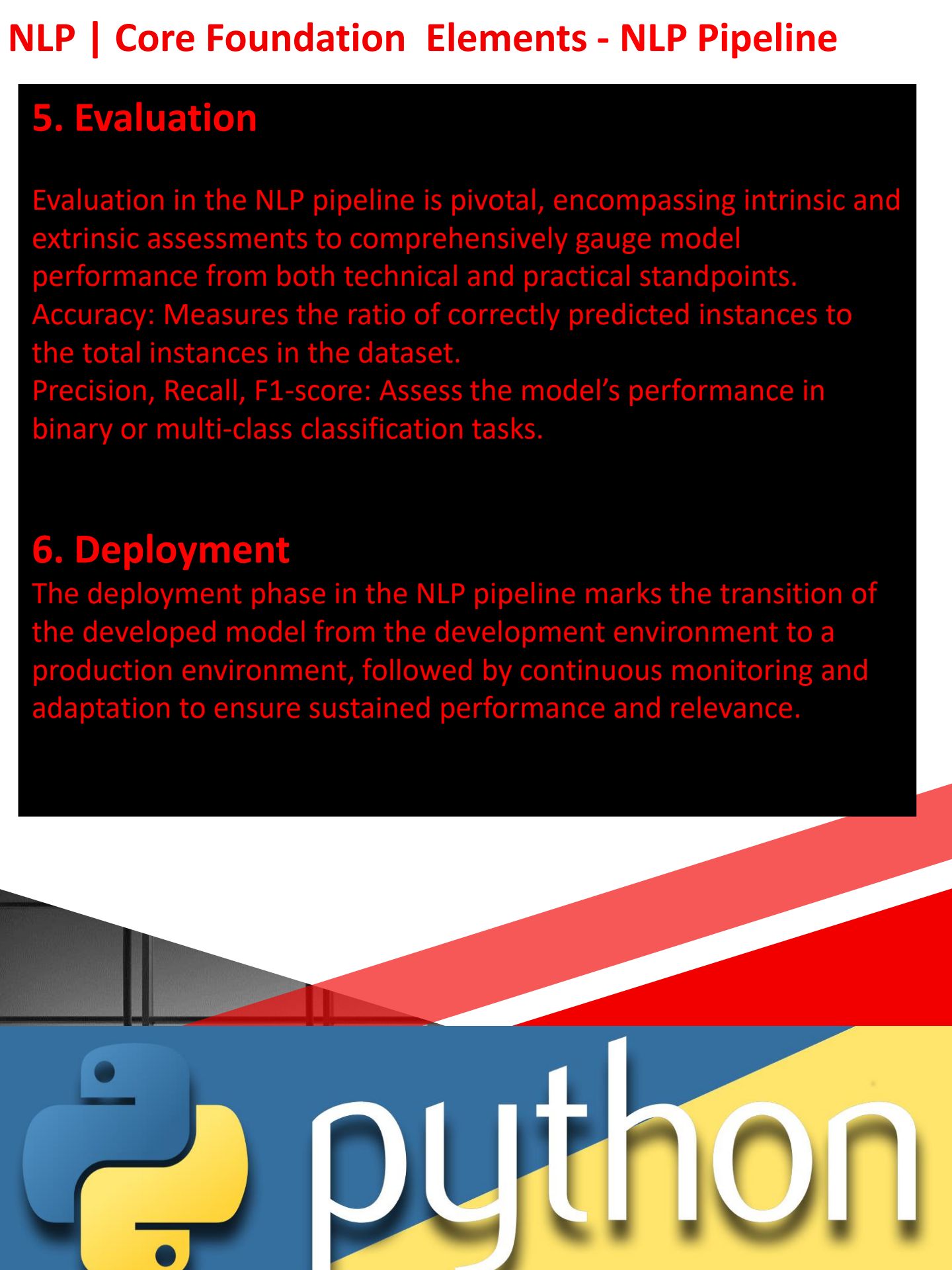
NLP | Core Foundation Elements - NLP Pipeline

5. Evaluation

Evaluation in the NLP pipeline is pivotal, encompassing intrinsic and extrinsic assessments to comprehensively gauge model performance from both technical and practical standpoints.
Accuracy: Measures the ratio of correctly predicted instances to the total instances in the dataset.
Precision, Recall, F1-score: Assess the model's performance in binary or multi-class classification tasks.

6. Deployment

The deployment phase in the NLP pipeline marks the transition of the developed model from the development environment to a production environment, followed by continuous monitoring and adaptation to ensure sustained performance and relevance.



NLP | Core Concepts - Exercises

See code here: https://github.com/ShahzadSarwar10/AI-ML-Explorer/blob/main/Week6/Case6-1-NaturalLanguageProcessing.py

You should be able to analyze – each code statement, you should be able to see trace information – at each step of debugging. "DEBUGGING IS BEST STRATEGY TO LEARN A LANAGUAGE." So debug code files, line by line, analyze the values of variable – changing at each code statement. BEST STRATEGY TO LEARN DEEP.

Let's put best efforts.
Thanks.
Shahzad – Your AI – ML Instructor

Exercises

https://towardsai.net/p/editorial/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0

https://pub.towardsai.net/understanding-semantic-analysis-using-python-nlp-f48016422677

Thank you - for listening and participating

- ❑**Questions / Queries**
- ❑**Suggestions/Recommendation**
- ❑**Ideas…..?**

Shahzad Sarwar
Cognitive Convergence
https://cognitiveconvergence.com
shahzad@cognitiveconvergence.com
voice: +1 4242530744 (USA) +92-3004762901 (Pak)