DeepTech_Ready
by Data Science Nigeria with support from Google.org

3MTT

3MTT/DeepTech_Ready Upskilling Program

Empowerment of young Nigerians with foundational technical skills and job-ready competencies in Data Science and Artificial Intelligence

Week 4: Fundamentals of Natural Language Processing

In this week, you will look at this course;

- Fundamentals of NLP

Course 1: Fundamentals of Natural Language Processing

Learning objectives for course

At the end of this course, you should be able to;

- Understand NLP Fundamentals.
- Preprocess Text and Analysis.
- Apply Text Representation Techniques.

Learning Requirements

To support your learning this week, you will require the following resources;

- Jupyter Notebook
- Google Colab (Recommended)

Provided is a guide on how to use and for your assignment with Google Colab. Google Colab Guide

Course 1: Fundamentals of Natural Language Processing

Link(s) to the course:

- Introduction to Natural Language Processing (NLP)
- Text Preprocessing, Tokenization, Stemming & Lemmatization
- Text Representations in NLP

Week 4: Fundamentals of Natural Language Processing

Learning Resources

Course:

1. **Slide 1** – Introduction to Natural Language Processing (NLP)
2. **Slide 2** - Text Preprocessing, Tokenization, Stemming & Lemmatization
3. **Slide 3** - Text Representations in NLP
4. **Notebook 1**- Regex Colab Notebook
5. **Notebook 2** - String Processing Colab
6. **Notebook 3**- Text tokenization Colab
7. **Notebook 4** - Text Representation Colab

Week 4: Fundamentals of Natural Language Processing

Applied Learning Assignments 1:

1. Define Natural Language Processing (NLP) in your own words.
2. List at least three real-world applications of NLP and explain their significance.
3. Identify and explain two challenges that make NLP complex.
4. Extract the following patterns using **regex:**

   a) All email addresses from the text below:

   "**Contact us at support@company.com or sales@business.org.**
   For more, email info@service.net."

   b) All words that end with "ing" from this sentence:
   "**NLP is amazing for cleaning and processing text while learning new techniques.**"

5. Write a Python program to clean the following text by:
   a) Removing all punctuation.
   b) Converting it to lowercase.
   c) Splitting it into words.

Week 4: Fundamentals of Natural Language Processing

Applied Learning Assignments 2:

1. Text Cleaning Task
   Apply text cleaning techniques to preprocess the following text:

   "**OMG!! NLP is soooo coool 🤩...!!! It costs $1000. Learn it now at https://3mtt.com 😎.**"

   Refer to the course slide for more information

2. Tokenization Task
   Perform both word-level and sentence-level tokenization on the given text.

   "Tokenization is the first step in NLP. It splits text into smaller pieces for analysis."

   - Use NLTK to perform **word tokenization**.
   - Use NLTK to perform **sentence tokenization**

Week 4: Fundamentals of Natural Language Processing

Applied Learning Assignments 2:

3. Stemming and Lemmatization Task

Apply stemming and lemmatization techniques to a list of words:

["running", "flies", "studies", "easily", "studying", "better"]

- Use **Porter Stemmer** to perform stemming on the words.
- Use **spaCy** to perform lemmatization on the same words.

Week 4: Fundamentals of Natural Language Processing

Applied Learning Assignments 3:

1. Define a vocabulary of at least 5 unique words. Write Python code to generate one-hot encoded vectors for your vocabulary.

2. Use the following sentences as your dataset:

   - "The quick brown fox jumps over the lazy dog."
   - "The dog sleeps in the kernel"

   – Write Python code to generate a Bag of Words representation for the dataset using CountVectorizer.

   – Write Python code to compute the TF-IDF representation using TfidfVectorizer.

Week 4: Fundamentals of Natural Language Processing

Applied Learning Assignments 3:

3. Create a small dataset of at least 3 sentences related to animals.

Example: "The cat meows. The dog barks. The bird sings."

- Write Python code to:
  - Train a Word2Vec model using gensim.
  - Retrieve the embedding for the word "dog".

4. Load the pretrained GloVe model (glove-wiki-gigaword-50) using gensim.

- Write Python code to:
  - Retrieve the embedding for the word "king".
  - Find the 5 most similar words to "king".

DeepTech_Ready

by **Data Science Nigeria** with support from Google.org

3MTT