**Internship Progress Report – Week 1 & 2**

**Fabeha Zahid Mahmood**

This document outlines the topics and concepts covered during the first two weeks of my internship. The work involved both theoretical understanding and practical applications, particularly focused on preparing for modern natural language processing tasks.

**1. NLP Overview**

Natural Language Processing (NLP) is a subfield of artificial intelligence that enables machines to read, interpret, understand, and generate human language. It forms the basis of many AI-powered applications like chatbots, translation systems, and information retrieval.

**2. Analysis vs Synthesis in NLP**

• **Analysis**: Focuses on breaking down and understanding the structure of language. Tasks include POS tagging, parsing, and sentiment analysis.  
• **Synthesis**: Involves generating new language output. Examples include language translation, summarization, and text generation.

**3. Core NLP Techniques**

The following key techniques were explored:  
• Sentiment Analysis  
• Named Entity Recognition (NER)  
• Text Classification  
• Topic Modeling  
• Text Summarization  
• Question Answering  
• Language Translation

**4. NLP Pipeline**

The standard NLP workflow consists of the following steps:

1. Data acquisition
2. Text cleaning
3. Tokenization
4. Stopword removal
5. Stemming or Lemmatization
6. Embedding or vectorization
7. Modeling and inference

**Tools Used:**  
• **NLTK** – Tokenization, stopword removal, stemming  
• **spaCy** – Tokenization, lemmatization, POS tagging, NER  
• **Gensim** – Word2Vec embedding and similarity  
• **NumPy** – Vector operations and similarity computation  
• **Pandas** – Data loading, transformation, and handling

**5. Data Cleaning with Regex**

Regular expressions (regex) were used to clean noisy or unstructured text, such as removing URLs, special characters, numbers, and HTML tags.

**6. Data Preprocessing Techniques**

• **Tokenization**: Splitting text into individual words or tokens.  
• **Stemming**: Trimming words to their root form (less context-aware).  
• **Lemmatization**: Converting words to their base form using grammatical context.

**7. N-Grams**

N-grams help capture local word patterns and context. Bigrams (e.g., "new york") and trigrams ("new york city") are used for capturing phrase-level information.

**8. Word Embeddings and Similarity**

• Word embeddings represent words as continuous vectors encoding semantic meaning.  
• **Word2Vec** models (CBOW and Skip-Gram) were explored.  
• **Cosine similarity** was used to compare semantic closeness of vectors for recommendation tasks.

**9. spaCy Overview**

spaCy is an efficient NLP library for industrial use.  
Key features:  
• Tokenization  
• POS Tagging  
• Dependency Parsing  
• Lemmatization  
• Named Entity Recognition

**10. Basics of Neural Networks in NLP**

Neural networks are fundamental in modern NLP applications.  
Topics included:  
• Neurons and activation functions (e.g., ReLU)  
• Layered architecture (Input → Hidden → Output)  
• Backpropagation  
• Applications: Text classification, embedding generation, sequence modeling

**11. Transformers and Architectures**

Transformers are foundational to models like BERT and GPT.  
Key Concepts:  
• Self-attention, positional encoding  
• Encoder-decoder architecture  
• Applications: Classification, QA, summarization  
• Benefits: Long-range dependency handling, parallelism

**12. Vector Databases and Pinecone**

Vector databases enable fast similarity search on embeddings.  
**Pinecone** was used to store high-dimensional vectors and retrieve similar results efficiently for recommendation tasks. It supports scalable, real-time search for use cases like book recommendation systems.

**13. Prompt Engineering (Side Learning)**

The **Google AI Essentials** course introduced fundamental concepts of prompt engineering.  
Topics included:  
• Writing clear, specific prompts  
• Using examples effectively  
• Structuring instructions to reduce ambiguity  
• Adapting prompts to improve output reliability