ASSIGNMENT - NLP and Sentimental Analysis, SoC 25’

Dataset: Twitter US Airline Sentiment (positive/neutral/negative)

Suggested libraries:

- `nltk`, `spaCy`, `scikit-learn`, `pandas`, `matplotlib`, `seaborn`
- Sentiment tools: **VADER** (`nltk.sentiment.vader`) or **TextBlob**
- *(Optional)* Embeddings: Gensim, Transformers (BERT)

Q1. Pre-processing & Token Analysis

Select a sample text (5–6 movie reviews or a short paragraph). Perform:

- a) Tokenization
- b) Stop-word removal
- c) Stemming vs. Lemmatization (show side-by-side outputs for ≥5 words)
- d) POS tagging

➢ Present a table comparing each step’s output to the original.

Q2. Vectorization Comparison

Using 20 text samples:

- a) Create features using BoW, TF-IDF, and *(optional)* word embeddings
- b) Show matrix shapes (samples × features)

- c) Discuss which captures semantics better and why

➢ Include code and a brief explanation.

Q3. Text Classification: Logistic Regression vs Naive Bayes

Using 30–50 labeled text samples:

- a) Preprocess & vectorize (BoW or TF-IDF)
- b) Train Naive Bayes & Logistic Regression models
- c) Evaluate with Accuracy, F1-score, and Confusion Matrix

➢ Conclude which model performed better and why.

Q4. Emotional Trajectory in a Passage

Take a 3–4 paragraph text (e.g., from Harry Potter):

- a) Split into 5 segments
- b) Compute sentiment for each using VADER or TextBlob
- c) Plot sentiment vs segment number

➢ In 3–4 sentences, interpret the emotional journey.

Q5. Conceptual Reflection (1–2 lines each)

1. Why is *lemmatization* often preferred over *stemming*?
2. How does *TF-IDF* down-weight common words?
3. Describe the *curse of dimensionality* in text data.

4. When should you use *word embeddings* instead of BoW/TF-IDF?

5. How can *POS tagging* enhance NLP pipelines?

✅ Submission Checklist

- **Google Doc/PDF** with:
  - Answers to Q1–Q5
  - Tables, code snippets, and plots
- **Jupyter Notebook** including:
  - Data loading and preprocessing
  - Vectorization
  - Model training and evaluation
  - Sentiment plot (Q4)
- Well-commented source code

🛠️ Notes

- Use `CountVectorizer`, `TfidfVectorizer`, `MultinomialNB`, and `LogisticRegression` from scikit-learn.
- VADER example: `from nltk.sentiment.vader import SentimentIntensityAnalyzer`
- TextBlob example: `from textblob import TextBlob`