Skill Share - Weekly Assignment (Week 3)

NLP Core Preprocessing & N-Grams Assignment

Deadline: Sunday, 11:59 PM
Format: Jupyter Notebook + Screenshots + GitHub Link (optional)

This week's assignment covers all major NLP preprocessing steps learned in class:

- Tokenization
- Stopword Removal
- Stemming
- Lemmatization
- POS Tagging
- Named Entity Recognition
- Bag of Words (BoW)
- N-Grams (1,2,3)

Use the following paragraph for all tasks:

```
Skill Share students are learning Natural Language Processing in Hyderabad.
Rohit teaches NLP with clarity, examples, and real-world scenarios. Many
learners from Chennai and Trivandrum attend these sessions to improve their AI
and ML skills.
```

📝Task 1 – Tokenization (Sentence & Word)

1. Perform sentence tokenization.
2. Perform word tokenization.
3. Count:
4. Total sentences
5. Total words
6. Unique words

📝Task 2 – Stopword Removal

1. Load NLTK stopwords.
2. Remove stopwords from tokenized words.
3. Print:
4. Removed stopwords list
5. Final filtered word list

6. Word reduction percentage

📝Task 3 – Stemming (Porter & Snowball)

1. Apply Porter stemmer.
2. Apply Snowball stemmer.
3. Create comparison table:

Word	Porter Stem	Snowball Stem

1. Write 3–4 lines about differences.

📝Task 4 – Lemmatization (Basic + POS-based)

1. Apply simple lemmatization.
2. Apply POS-tag-aware lemmatization.
3. Print both results.
4. Write 4–5 lines on why lemmatization is more accurate.

📝Task 5 – POS Tagging

1. Generate POS tags for the filtered words.
2. Group words into:
3. Nouns
4. Verbs
5. Adjectives
6. Adverbs
7. Visualize counts using a bar or pie chart.

📝Task 6 – Named Entity Recognition (NER)

1. Perform NER using NLTK.
2. Extract:
3. PERSON
4. ORGANIZATION
5. GPE
6. DATE (if any)
7. Put results in a table.
8. **Bonus:** Use spaCy for advanced NER.

📝Task 7 – Bag of Words (BoW)

1. Create BoW using CountVectorizer.
2. Display vocabulary list.
3. Print the BoW matrix as a DataFrame.
4. Identify top 5 most frequent words.

📝Task 8 – N-Grams (1, 2, 3)

1. Generate:
2. Unigrams
3. Bigrams
4. Trigrams
5. Print all three lists.
6. Explain (4–5 lines) how meaning improves from unigram → bigram → trigram.

📝Task 9 – Real-World NLP Application Question

Write 8–10 lines answering:

"How can these text preprocessing techniques be used to build a student feedback sentiment analyzer for Skill Share?"

📤Submission Requirements

✔️Notebook with all code executed
✔️Output screenshots
✔️Visualizations (charts/tables)
✔️Summary write-up
✔️GitHub link (optional bonus)

🕙Grading (Out of 40 Marks)

Component	Marks
Tokenization	4
Stopwords Removal	4
Stemming & Lemmatization	6
POS Tagging	5

Component	Marks
Named Entity Recognition	5
Bag of Words	6
N-Grams	6
Real-World Question	4