Assignment 1
NLP Basics & Text Preprocessing

Part A — Foundations

25 questions

★	**Must-know** — asked very frequently in GenAI / ML interviews (8 questions). Prepare these first.
[CODE]	**Hands-on task** — write and run the code (4 tasks).

1. What is NLP? Give 3 real-world applications you have seen in production. ★
2. What are the stages in the lifecycle of an NLP project (data → deployment → monitoring)?
3. What are the main challenges in NLP? ★
4. What is ambiguity in NLP? Explain lexical, syntactic, semantic and pragmatic ambiguity with examples.
5. Define corpus, document, vocabulary and token.
6. What are the common preprocessing steps in NLP? Which of them should you SKIP when using a Transformer/LLM, and why?
7. What is text normalization? List common normalization methods.
8. What is tokenization? Compare word-level, character-level and subword tokenization. ★
9. What is stemming and how is it different from lemmatization? When would you prefer each? ★ · Asked: NLP interview rounds
10. Explain over-stemming vs under-stemming with examples.
11. Why does WordNetLemmatizer need POS tags? What happens if you don't pass them?
12. What are contractions (e.g. "don't", "I'm") and how do you handle them?
13. How do you handle emojis, URLs, hashtags and HTML tags in text?
14. How do you handle spelling mistakes in user input?
15. What are subword tokenization methods? Explain BPE, WordPiece and SentencePiece briefly. ★
16. Explain step by step how the WordPiece tokenizer builds its vocabulary and tokenizes a new word.
17. Which tokenizers are used by GPT-family models and by BERT? ★
18. Why does an LLM sometimes struggle to count letters in a word or do arithmetic? (Hint: tokenization.)
19. Write code for word and sentence tokenization (NLTK / spaCy). [CODE]

20. Write code for stemming and lemmatization and compare the outputs on the same sentence. [CODE]

21. Write code for stopword removal. When can removing stopwords hurt a model? [CODE]

22. In real-time data users type "Lunch", "lunch" and "LUNCH". Why do we lowercase text, and when should you NOT lowercase? ★ · Asked: NLP interview rounds

23. You removed all numeric values from the text. Could numbers contain useful information? How would you handle numbers and punctuation instead? · Asked: NLP interview rounds

24. Users will type stopwords at inference time. If you remove them during training, what happens? When should you remove stopwords and when not? Give practical applications. ★ · Asked: NLP interview rounds

25. Live-coding: clean this messy real-world review end-to-end and explain why the ORDER of steps matters: [CODE] · Asked: NLP interview rounds

```
raw = "Heyyyy!!! I justtt bought this from @Amazon □. TBH it's sooo goodddd!!! Price was Rs.1,999 & I g
```