Sample Dataset (for Preprocessing Assignments)

```
sample_data = [
    "🚀 Euron offers 50+ self-paced AI & Data Science courses! #AI #Learning",
    "Learn Python, NLP, and ML on Euron.one — India's leading AI learning
platform.",
    "What’s new in Euron? 100+ projects, WhatsApp bots, and resume analyzers!
□🔥",
    "Euron = 🔥 AI tools + 📚 Deep content + □ Smart learning. Let’s gooo!!",
    "Join Euron's Focus Mode 💡 - Bye distractions, Hello productivity!
#NoExcuses",
    "I’ve completed 10 projects on Euron already. Can't wait to start the
next one!",
    "Euron’s Resume Analyzer Tool just gave me a 92% job match! #JobReady",
    "Is Euron available in Hindi too? I want to learn AI in my mother
tongue.",
    "Euron = Knowledge + Projects + Certifications + Placement Support 💼",
    "Use code AI2025 to get 30% OFF on all Euron subscriptions! #Deal
#Learning"
]
```

Hands-On Assignment Questions

BASIC TEXT CLEANING

1. **Remove emojis from all Euron sample texts.**
   Hint: Use the `emoji` library.
2. **Convert all text to lowercase and remove punctuation.**
   Use `str.lower()` and `string.punctuation`.
3. **Remove all hashtags from the data.**
   Use a regex like `#\w+`.

TOKENIZATION & STOPWORDS

4. **Tokenize each sentence into words.**
   Use NLTK’s `word_tokenize`.
5. **Tokenize each sentence into individual sentences.**
   Use NLTK’s `sent_tokenize`.
6. **Remove stopwords using NLTK and compare results with spaCy stopwords.**
   Show both side by side.

REGEX CLEANING

7. **Write a function to clean digits, special characters, and extra whitespace using regex.**
   Example: Remove numbers like 92% and 30%.
8. **Replace all URLs (e.g., `euron.one`) with `<LINK>` using regex.**
9. **Write a function to identify and extract all capitalized words (e.g., AI, Python).**

STEMMING & LEMMATIZATION

10. **Apply PorterStemmer on tokenized words and display root forms.**
11. **Apply SnowballStemmer and compare results with PorterStemmer.**
12. **Use spaCy's lemmatizer to lemmatize all words in each sentence.**
13. **Use NLTK's WordNet lemmatizer and compare it with spaCy output.**

ADVANCED CLEANING TASKS

14. **Expand contractions in all text lines.**
    E.g., "Can't" → "Cannot", "I've" → "I have"
15. **Build a full cleaning function that combines: lowercasing, emoji/hashtag removal, punctuation removal, and stopword removal.**
16. **Create a custom stopword list that also removes words like 'euron', 'project', and apply it.**

PIPELINE-STYLE ASSIGNMENTS

17. **Create a custom preprocessing pipeline function that performs all steps: cleaning, tokenization, stemming, lemmatization, stopword removal.**
18. **Build a pipeline that returns both clean sentence and list of clean tokens.**
19. **Generate a frequency count of the most common non-stopword terms across all Euron texts (after cleaning).**
20. **Build a DataFrame with original text, cleaned text, tokenized words, and lemmas in separate columns.**