DATA 622 | HOMEWORK 2

Questions

Proceed as indicated:

1. **Read the File**

Read the content of https://www.cnn.com/2025/06/13/style/why-luxury-brands-are-so-expensive and print the first 700 characters.

2. **Remove HTML Tags**

If any HTML tags are present in the file, remove them so that only the raw text remains.

3. **Lower and Remove Punctuation**

Convert all text to lowercase and remove all punctuation characters.

4. **Remove Stopwords**

Remove English stopwords from the text. (Use NLTK's list of stopwords.)

5. **Lemmatize the Text**

Lemmatize all remaining words (use NLTK's *WordNetLemmatizer)* and print the first 50 lemmatized words. Is there any difference in your output if you stemmed the text?