MIDTERM EXAM

- Copy each part of the task as comment above the solution
- Use Jupyter Notebook to solve tasks
- Upload the .ipynb file to the LMS
- Name file in format: **FirstName_LastName**

1. (48 pts)You will be working on one of the Gutenberg project's text - austen-persuasion.txt. Use NLTK library to finish following tasks:
    - Load text
    - Find the total number of characters, words, and sentences (**using NLTK**)
    - Find the total number of distinct words
    - Find the frequency distribution of the text and the most common words
    - Find how many times the word „love“ appears
    - What percentage of all words is the word „love“?
    - Generate a cumulative frequency plot for the 8 most common words.
    - Tokenize the text and show the first 14 tokens.
    - Find part-of-speech tags for those tokens
    - Find the 50 most frequent bigrams of a text, omitting bigrams that contain stopwords.
    - Use the Porter Stemmer to normalize tokenized text, calling the stemmer on each word.
    - Use WordNet Lemmatizer to find a list of valid lemmas

2. (25 pts) Define a function find_language() that takes a string as its argument and returns a list of languages that have that string as a word. Use the udhr corpus and limit your searches to files in the Latin-1 encoding.

3. (27 pts) You will be working with the NLTK Brown Corpus, which contains text samples from various genres. Your objective is to analyze the distribution of word lengths in different genres. Implement the following tasks:
    - Define a function genre_word_length_distribution(corpus, genres) that takes the Brown Corpus and a list of genres as input. The function should return a Conditional Frequency Distribution where the conditions are genres, and the values are the lengths of words in the corresponding genre.

- Use the function to generate the Conditional Frequency Distribution for a specific set of genres, e.g., ['news', 'romance', 'science_fiction'].
- Plot the Conditional Frequency Distribution to visualize the distribution of word lengths across the selected genres.