Language Identification Assignment

You may use any programming language or tool(s) to do this assignment. Please submit the following documents:

1. A PDF report answering the questions mentioned below
2. Your code and README file in a tarball archive

Download the datasets required for this assignment via this **link**.

Submit a report based on the Czech-Polish language pairs after performing the following operations on the dataset you downloaded:

1. **(5 points)** Collect character-based unigram, bigram, trigram and 4-gram character-based ngram counts from the training dataset.
2. **(5 points)** Using character ngram counts from the training data, predict the language of each sentence in the test set using unigram, bigram, trigram and 4-gram character-based ngram counts. Report the precision and recall of classification of each of the above ngram classes in tabular format. Write a short note on the ngrams with zero count in your test set.
3. **(5 points)** Write a short note on the cases where your system misclassified sentences. Please use linguistic examples to illustrate your points.
4. **(5 points)** Analyze the cases where the system successfully classified sentences and illustrate the linguistic phenomena that are modelled by character ngrams (you may need to refer to additional materials to learn more about the language pairs in question).

Bonus question (10 points): Perform the same experiments using the Spanish-Portugese dataset and answer the above questions.

Note: Please mention any assumptions you make in your report (e.g., tokenization, character encoding, or preprocessing).

Submission format: `rollno_a2.tar`