**Python Challenge Project: Natural Language Processing/AI**

**Goal:** Determine the language of text in a mystery file/document.

**Given:** Several files that are labelled based on their language. (ex: provided files in English, Spanish, German, and a mystery file, determine which of the 3 languages the mystery file is)

**Strategy:** Identify patterns in each of the labelled text files in the form of character sequences, and identify patterns in the mystery file and determine which labelled file matches closest to the mystery file.

* The patterns we will be looking at are known as n-grams, or sequences of characters of a specified length, n
* For example, the 3-grams of the following sentence:
  + “learning python coding is exciting” 🡪 [‘lea’, ‘ear’, ‘arn’, ‘rni’, ……..]
* Certain sequences are very common in English (‘ing’, ‘the’, ‘ion’) and other sequences are more common in other languages (ex: German (‘der’, ‘und’, ‘ein))
* By finding the most common 3-character sequences in each of the documents that the language is known, the language of the mystery document can be predicted based on how closely its most common 3-character sequences match up with each of the labelled documents (for example, if the most common sequences in the mystery document are very similar to the most common sequences in the French document, then French is the predicted language)