Task: Parsing Text Files (%55)

This assessment touches the very first step of analysing textual data, i.e., extracting data from semi-structured text files. Datasets that contain information about COVID-19 related tweets is given. Each text file contains information about the tweets, i.e., “id”, “text”, and “created_at” attributes. The task is to extract the data and transform the data into the **XML** format with the following elements:

1. id: is a 19-digit number and letter.
2. text: is the actual tweet.
3. Created_at: is the date and time that the tweet was created

The following constraints must be satisfied:

1. The “id”s must be unique, so if there are multiple instances of the same tweets, you must only keep one of them in your final XML file.
2. The non-English tweets should be filtered out from the dataset and the final XML should only contain the tweets in English language. For the sake of consistency, you **must** use the langid package to classify the language of a tweet.
3. The **re, os,** and the **langid** packages in **Python** are the only packages that you are allowed to use for the task (e.g., “pandas” is not allowed!). Any other packages that you need to “import” before usage is not allowed.