**Interim report on Machine translation**

**Contents :**

1. Context
2. Datasets
3. Import Dataset
4. Data cleansing
5. Pre processing
6. Data split
7. Test and train the data
8. Evaluating the model
9. Conclusion

**CONTEXT**:

Machine Translation is the automated translation of source material into another language without human intervention. The database comes fromACL2014 Ninth workshop on Statistical Machine Translation. This workshop mainly focuses on language translation between European language pairs. The idea behind the workshop is to provide the ability for two parties to communicate and exchange the ideas from different countries

The database is basically sentences in German/English of various events.

**Datasets:**

Three datasets are obtained from Statistical Machine Translation workshop.

* Europarl v7
* Common Crawl corpus
* News Commentary

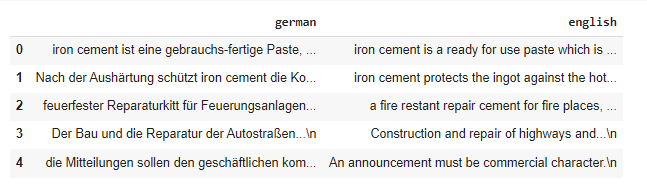
**Import Datasets :**

We're using dataset containing pairs of English - German sentences

The three datasets are present in .txt file.

* The Europarl v7 has a .txt file containing **1920209** pairs of English to German phrases.
* The Common Crawl corpus has a .txt file containing **2399123** pairs of English to German phrases.
* The News Commentary has a .txt file containing **201288** pairs of English to German phrases

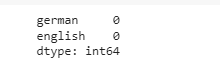
We have now mereged this into a dataframe with English and German columns in it.The total Dataset has **4520620** records



**DataCleansing:**

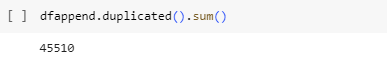
* **Check for nulls**

We have checked if there are any null values in the data set and we see there is no such values



* **Check for duplicates :**

Check if there are any duplicates and remove them as it is not needed and in this way we can reduce the size and increase the test results.

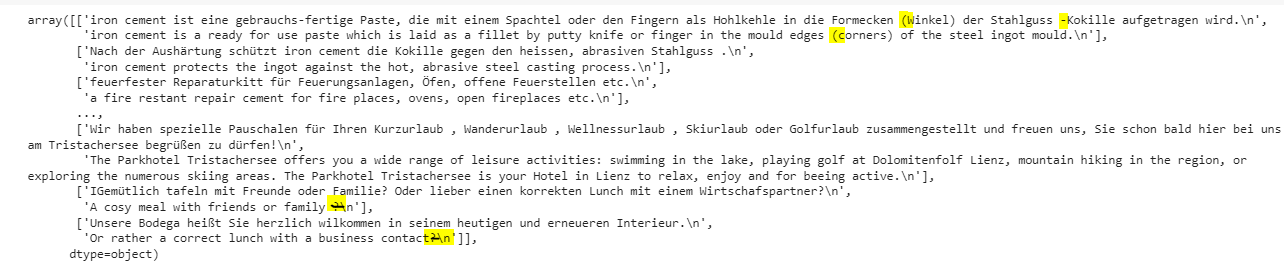


* **Drop the duplicates:**

We have 45510 dupliactes, post removing this the size of the dataframe is **4475110**

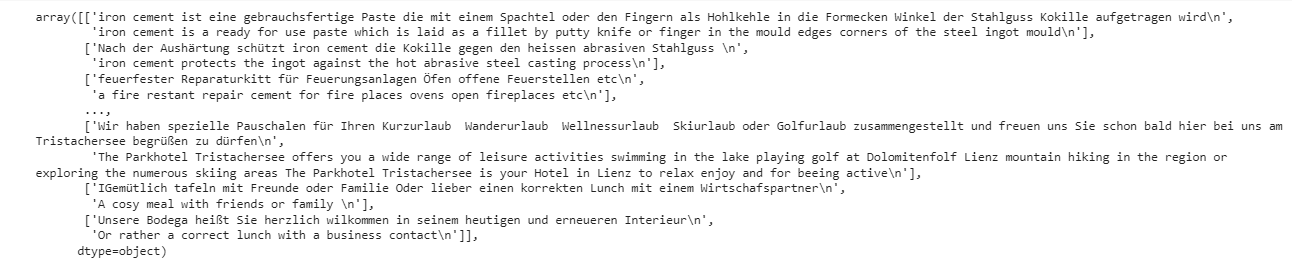
* **Check for punctuations:**

Checking the datasets if there are any punctuation marks or special characters.



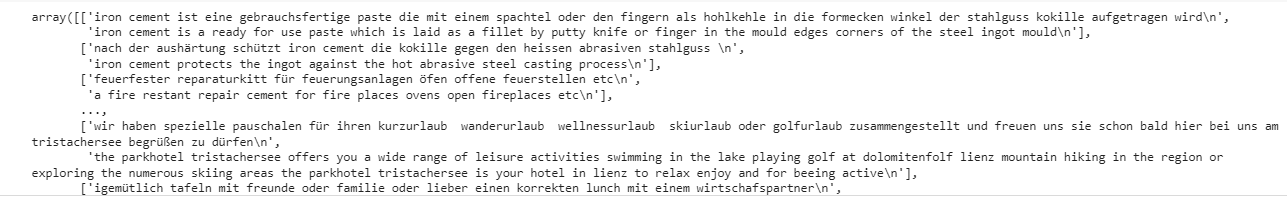
* **Remove the punctuations:**

We see the data has the same and we are removing them



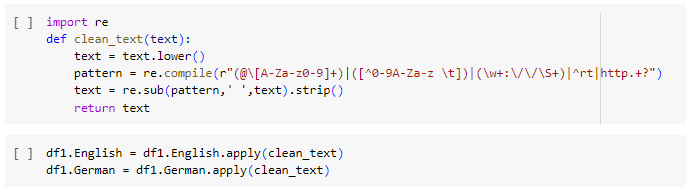
* **Convert to Lower case:**

Converted all the text to lower cases for both the languages

****

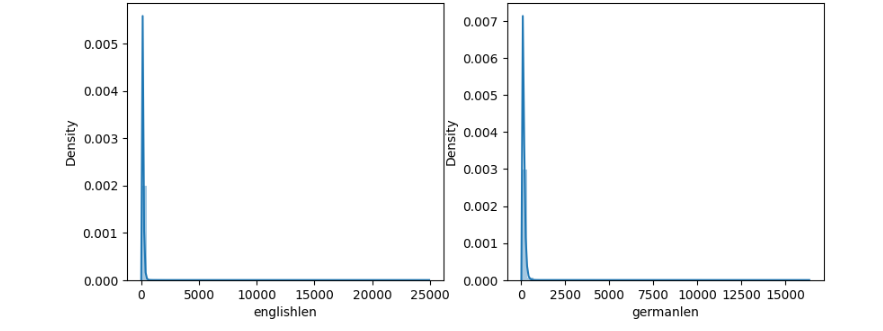
* **Remove Html tags and Urls**

Have written a function to clean the html tags, url if any present in between the text in both the languages



* **Plot the length of the sentence**

Have plotted the length of the sentence using displot



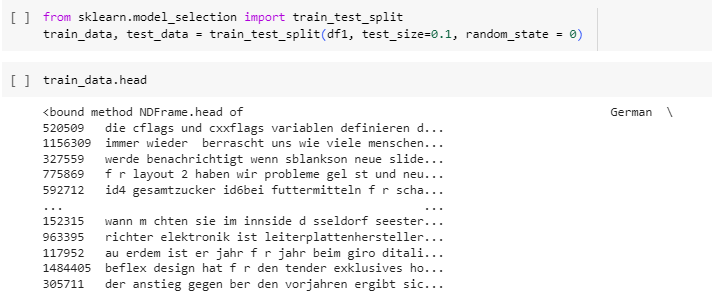
From the plot it is visible that the length of German sentences are bigger than English.

**Pre processing:**

Let us vectorize our text data by using Keras Tokenizer class. It will turn our sentences into sequences of integers. We can then pad those sequences with zeros to make all the sequences of the same length

**Data Split :**

We will now split the data into train and test set for model training and evaluation,



Also since we checked the length and added them in the df , letz drop those.

**Test and train:**

**Evaluate the model:**

**Conclusion :**

With a simple RNN and LSTM model, the results are good. We can improve on the performance by using a more encoder-decoder model on a larger dataset. We can also try our hands on considering longer sentences.