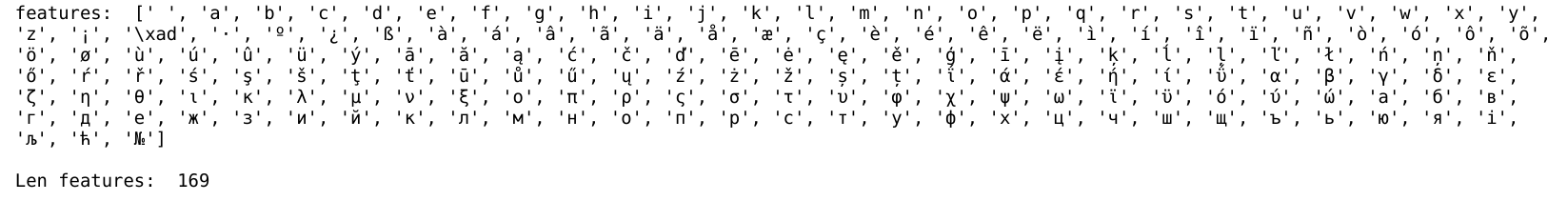
**REPORT**

**LANGUAGES IDENTIFICATION**

* **Approaching:**
  1. Read and understand the paper
  2. Intent to the terminologies in Languages Identification, synonym terms
  3. Research the methods to solve this problem in the github, tutoria; in medium for example:
     + Deep learning methods:
       - <https://github.com/bhargav265/Language-Detection>
     + Using RNN - recurrent neural network:
       - <https://github.com/rand0wn/LanguageDetection>
     + Using machine learning classifier algorithms: SVC, Gaussian, Logistic regression, Random Forest.
       - <https://github.com/rand0wn/LanguageDetection>
       - <https://github.com/NightFantom/LanguageDetection>

In this approaching, I choose the third approaching which spends less time for training ( I use CPU I7, 6GB, 12 threads, GTX 1050Ti - just 768 cudas). Beside that I use scikit-learn for training process, it optimized for CPU, so it is one of my advantages.

* 1. Evaluate models: I used Gridsearch CV to optimize the parametters of model, the accuracy up to **95 %** with **Random Forest** Algorithm. (train/test = 0.8/0.2). You can see all of them in my source code in main.ipynb file.
* **Preprocessing:**
  + The first step is remove special characters like '\t', '!', '"', '%', '&', '\*',.. and the number out of the dataset. You can see all in my code, it is to make sure that the feature engineering step will be good. Then, I use CountVectorizer of sklearn to extract features to save the coding time. The result, I got 169 features to represent for this database.



* + Finally, I give it to classificasions. It is SVC, Gaussian, Logistic regression, Random Forest classifier.
* **Evaluation model:**
  + Advantages: I tried Random Forest, SVC, Gausian, Logistic regression algorithms and i see that Random Forest give us the best accuracy In additionally, i use Grid search CV to optimize the parametters of the algorithm. It give me n\_estimators=300, max\_features= 'log2' is the best parametters.



* + Disadvantages: The time for training is very fast with the classification but to find the optimal parametters for random forest algorithm, it took more time ( near 30 mins). It the system need to learn online or realtime training and prediction, i think Gaussian or Logistic Regression is better because the accuracy is not much different but the time for training is very fast.
* **Imrovement:**
  + We can use other aproaching like RNN, reinforcement learning to learn and remember the characters, I use my personal laptop/CPU to train the model. So it is not good to use this way because the parametter model is very large, the time training will so long and hot GPU,.. But i think this method will better my solution.