Technical Report

This is a Java program for detecting which language is written in a document.

Title: LanguageDetector2
Date: 2016.01.09 ~ 2016.01.11
Author: KyungTae Lim
Email: kyungtaelim@kisti.re.kr

[About system flows]:

1. Generating plain text corpus from Wikipedia for each language (just copy that in a java file for every languages)

2. Extracting features for each language

    a. **Term based feature extraction**:
    - From Wikipedia articles which are plain text, generating term(chunk) dictionary with frequency in the articles for each language
    - Extracting Top N terms from the dictionary in each language. For example, extracted results can be "and, or, he, Obama" in English and "der, Ist, die" in German. It is decided for feature about language.

    b. **Character based feature extraction**:
    - From Wikipedia articles which are plain text, generating character dictionary with frequency in the articles for each language.
    - Extracting Top N character from the dictionary in each language. (normally 25 percent of character should be deleted because it is negligible for feature)
    - From all language's character dictionary, it generate a new global dictionary, the example of data set like below:

char	frequency	language	Picked
a	3	eng, ger, fr	
b	3	eng, ger,fr	
z	2	eng, ger	
ä	1	ger	Picked for feature
à	1	fr	Picked for feature

[char] [frequency] [language]
a 3 eng,ger,fr
b 3 eng,ger,fr
z 2 eng,ger
ä 1 ger --> picked for feature
à 1 fr --> picked for feature

    - From the global dictionary, extracting features by low frequency then the

feature would be unique for each language.

3. Generating SVM training corpus by using feature dictionary

4. Machine Learning (Support Vector Machine)

5. How to use?
   - For testing: The class Main.java is the example of test, you just put the text on the "testSetGenerator" method

[Additional information]

1. How to add another language such as chinese?
   - Generating corpus in LanguageResources.class
   - Make new language process in LanguageDetectorHandler

2. How can i check the feature lists and feature numbers?
   - You just add these codes in the main class

```
HashMap<String, ArrayList<TermVo>> featureList =  ldh.getFeatureDic().featureTermDictionary;
HashMap<String, Integer> featureNumberList = ldh.getFeatureDic().featureNumberDictionary;
```