Language Identifier

Description

Implementation of the method to identify the language of a given document. The algorithm covers 14 languages including Azeri, Dutch, English, French, German, Icelandic, Italian, Norwegian, Polish, Portuguese, Romanian, Spanish, Swedish and Walloon. The idea is partially taken from this Wikipedia page [1] that describes a language recognition chart to identify a language. As per my knowledge, no other researcher have used this technique until know. The algorithm is based on commonly used words (20 words per language), commonly used suffixes (maximum 20 per language) and count of special characters used in a language other than standard English alphabets. These words, suffixes and special characters for each language are defined in LanguageRules.java file.

For a given document, the score for each language is computed and language with maximum scores is returned as the language of the document. For example, if the score distribution is following for an input document, then the language of the document is Azeri**.**

Language	**CommonWordsScore + SuffixScore + SpecialCharactersScore = Final Scores**
English	1 + 1 + 0 = 2
Italian	0 + 2 + 0 = 2
French	0 + 0 + 17 = 17
Spanish	0 + 1 + 12 = 13
Dutch	0 + 0 + 18 = 18
German	0 + 0 + 18 = 18
Romanian	0 + 0 + 0 = 0
Portuguese	0 + 0 + 5 = 5
Walloon	0 + 0 + 11 = 11
Swedish	0 + 0 + 6 = 6
Norwegian	0 + 0 + 0 = 0
Azeri	**11 + 26 + 32 = 69**
Icelandic	0 + 3 + 6 = 9
Polish	0 + 0 + 0 = 0

There can be the cases where scores of 2 or more languages are the same. In that case, all these languages are returned as a result.

Evaluation

Because of the unavailability of a single dataset containing test data for all above languages, multiple datasets are used for evaluation.

LIGA [2] dataset:

LIGA dataset is used for evaluation of German, English, Spanish, French, Italian and Dutch. The LIGA dataset contains 2 groups of files for each language; 10 files with large text (Approx. 2000 words per document), and 10 files with the text of length 300 words approximately. Along with these files, another set of documents was created by taking the random text (Approx. 30 words per document) from the large documents. For the above-mentioned languages, there are total 30 files for each language. Following table shows the results:

Language	**Large text Files**		**Medium text files**		**Small text files**	
	Correctly identified/total files	**Accuracy**	**Correctly identified/total files**	**Accuracy**	**Correctly identified/total files**	**Accuracy**
German	10/10	100%	10/10	100%	10/10	100%
English	10/10	100%	10/10	100%	10/10	100%
Spanish	10/10	100%	10/10	100%	9/10	90%
French	10/10	100%	10/10	100%	9/10	90%
Italian*	10/10	100%	10/10	100%	10/10	100%
Dutch**	10/10	100%	10/10	100%	10/10	100%

*: Algorithm incorrectly detected other languages too along with Italian for 4 files in the Italian dataset, detailed results can be seen in Output/output.txt file

**: Algorithm incorrectly detected French too along with Dutch for 1 file in the Dutch dataset, detailed results can be seen in Output/output.txt file

DLI32 and DLI32-2 corpora [3]:

The DLI32 corpus contains 10 files per language, in which the text length ranges between 93 and 146 words. The DLI32-2 have 20 texts per language and the text length ranges between 43 and 67 words. The texts may contain any kind of the following noises*: URLs, Citations in other language, Tags, Abbreviations, Unaccented characters. DLI32 and DLI32-2 corpora are used for evaluation of Portuguese, Swedish, Norwegian, Icelandic, and Polish. Following table shows the results:

Language	**Medium text files**		**Small text files**	
	Correctly identified/total files	**Accuracy**	**Correctly identified/total files**	**Accuracy**
Portuguese	5/5	100%	15/15	100%
Swedish	8/8	100%	18/18	100%
Norwegian**	10/10	100%	20/20	100%
Icelandic	1/1	100%	6/6	100%
Polish	10/10	100%	20/20	100%

*: Files with parsing issues were removed

**: Algorithm incorrectly detected German language along with Norwegian for 1 file in the Norwegian dataset, detailed results can be seen in Output/output.txt file

For the rest of the languages; text is taken from multiple online resources. Following table shows the results for Romanian, Walloon and Azeri languages:

Language	**Large text files (ranges b/w 100 and 150)**		**Small text files (Approx. 30 words)**	
	Correctly identified/total files	**Accuracy**	**Correctly identified/total files**	**Accuracy**
Romanian	5/5	100%	5/5	100%
Walloon*	3/3	100%	3/3	100%
Azeri	5/5	100%	5/5	100%

*: Algorithm incorrectly detected English language along with Walloon for 1 file in the Walloon dataset, detailed results can be seen in Output/output.txt file

Possible next steps

1. Can be extended to detected languages of documents contain text from multiple languages
2. Testing with other languages that are not given in LanguageRules.java file

[1]: https://en.wikipedia.org/wiki/Wikipedia:Language_recognition_chart

[2]: https://github.com/llaisdy/liga/tree/master/datasets

[3]: https://github.com/xprogramer/DLI32-corpus