Language detection

This `python` library can be used to detect the natural language of a given document. It is based on the relative frequency of n-grams (n-gram model) to create sets of character n-grams and use them as a feature to train the classifier.

The classifier uses the n-grams frequency (one to three) from the testing and training docments, which is weighted with “TF-IDF” (Term Frequency–Inverse Document Frequency), and then apply the classical cosine similarity measure to obtain a distance between the docments.

All the testing documents were submitted to the classifier method. The proposed method achieved 100% accuracy in language detection.

Getting started

- `resources\train` folder has 110 language docments used for training. The docments selected are based on the Universal Declaration of Human Rights (UDHR) because it is available in a large number of languages.
- `resources\test` folder has 10 language docments used for testing. The docments selected are extracted from Wikipedia articles written in different languages.
- `TrainingTesting.ipynb` is the algorithm that creates the n-gram models for testing and training data. Each document is submitted to pre-processing tasks, such as segmentation, remove punctuation, lowercase letters, and tokenization.
- `Classifier.ipynb` is the proposed method to identify the language a document is written in.

Supported languages

This library supports the identification of multilingual documents as shown in the table below. It is based on a plain text version prepared by the “UDHR in Unicode” project. [https://www.unicode.org/udhr]

Language	Language Code	Language	Language Code
Abkhaz	ab	Inuktitut	iu
Afrikaans	af	Italian	it
Albanian	sq	Japanese	ja
Amharic	am	Javanese	jv
Arabic	ar	Kanuri	kr
Armenian	hy	Khmer	km
Aymara	ay	Korean	ko
Azerbaijani, North (Cyrillic)	az-Cyrl	Kurdish	ku
Azerbaijani, North (Latin)	az-Latn	Lao	lo
Basque	eu	Latin	la
Belarusan	be	Latvian	lv
Bengali	bn	Lingala	ln
Bislama	bi	Lithuanian	lt
Bosnian (Cyrillic)	bs-Cyrl	Malay (Arabic)	ms-Arab
Bosnian (Latin)	bs-Latn	Malay (Latin)	ms-Latn
Breton	br	Maltese	mt
Bulgarian	bg	Marshallese	mh
Catalan	ca	Mongolian, Halh (Cyrillic)	mn-Cyrl
Chamorro	ch	Navajo	nv
Chinese, Mandarin (Simplified)	zh-Hans	Ndonga	ng
Chinese, Mandarin (Traditional)	zh-Hant	Norwegian, Bokmål	nb
Corsican	co	Norwegian, Nynorsk	nn
Cree	cr	Persian	fa
Croatian	hr	Polish	pl
Czech	cs	Portuguese (Brazil)	pt-BR
Danish	da	Portuguese (Portugal)	pt-PT
Dutch	nl	Romanian	ro
Dzongkha	dz	Russian	ru
English	en	Sanskrit	sa
Esperanto	eo	Slovak	sk
Estonian	et	Slovene	sl
Faroese	fo	Somali	so
Fijian	fj	Spanish	es
Finnish	fi	Swati	ss
French	fr	Swedish	sv
Frisian	fy	Tagalog	tl
Gaelic, Irish	ga	Tahitian	ty
Gaelic, Scottish	gd	Tamil	ta
Galician	gl	Tatar	tt
Ganda	lg	Thai	th
Georgian	ka	Tibetan	bo
German	de	Tonga	to
Greek (monotonic)	el-monoton	Turkish	tr
Greek (polytonic)	el-polyton	Ukrainian	uk
Guarani	gn	Urdu	ur
Gujarati	gu	Uyghur (Arabic)	ug-Arab
Hausa	ha	Uyghur (Latin)	ug-Latn
Hebrew	he	Uzbek	uz
Hindi	hi	Venda	ve
Hungarian	hu	Vietnamese	vi
Icelandic	is	Walloon	wa
Ido	io	Welsh	cy
Igbo	ig	Wolof	wo
Indonesian	id	Xhosa	xh
Interlingua	ia	Yoruba	yo

In [ ]: