Foysal87 / **sbnltk** Public

Bangla NLP toolkit. Bangla NER, POStag, Stemmer, Word embedding, sentence embedding, summarization, preprocessor, sentiment analysis, etc.

MIT License

9 stars **1** fork

Star Watch

Code Issues Pull requests Actions Projects Wiki Security Insights

main

ahmed.foysal@orbitax.com … on Jun 1

View code

pypi-download-stats

pypi v1.0.5 license MIT python 3 downloads 22/month downloads 4/week

SBNLTK

SUST-Bangla Natural Language toolkit. A python module for Bangla NLP tasks.
Demo Version : 1.0
NEED python 3.6+ vesrion!! Use virtual Environment for not getting unessessary Issues!!

INSTALLATION

PYPI INSTALLATION

```
pip3 install sbnltk
pip3 install simpletransformers
```

MANUAL INSTALLATION FROM GITHUB

- Clone this project
- Install all the requirements
- Call the setup.py from terminal

What will you get here?

- Bangla Text Preprocessor
- Bangla word dust,punctuation,stop word removal
- Bangla word sorting according to Bangla or English alphabet
- Bangla word normalization
- Bangla word stemmer
- Bangla Sentiment analysis(logisticRegression,LinearSVC,Multilnomial_naive_bayes,Random_Forst)
- Bangla Sentiment analysis with Bert
- Bangla sentence pos tagger (static, sklearn)
- Bangla sentence pos tagger with BERT(Multilingual-cased,Multilingual uncased)
- Bangla sentence NER(Static,sklearn)
- Bangla sentence NER with BERT(Bert-Cased, Multilingual Cased/Uncased)
- Bangla word word2vec(gensim,glove,fasttext)
- Bangla sentence embedding(Contexual,Transformer/Bert)
- Bangla Document Summarization(Feature based, Contexual, sementic Based)
- Bangla Bi-lingual project(Bangla to english google translator without blocking IP)
- Bangla document information Extraction

SEE THE CODE DOCS FOR USES!

TASKS, MODELS, ACCURACY, DATASET AND DOCS

README.md

Preprocessor	Punctuation, Stop Word, DUST removal Word normalization, others..	------	-----	
Word tokenizers	basic tokenizers Customized tokenizers	----	----	

TASK	MODEL	ACCURACY	DATASET	About
Sentence tokenizers	Basic tokenizers Customized tokenizers Sentence Cluster	-----	-----	
Stemmer	StemmerOP	85.5%	----	
Sentiment Analysis	logisticRegression	88.5%	20,000+	
	LinearSVC	82.3%	20,000+	
	Multilnomial_naive_bayes	84.1%	20,000+	
	Random Forest	86.9%	20,000+	
	BERT	93.2%	20,000+	
POS tagger	Static method	55.5%	1,40,973 words	
	SK-LEARN classification	81.2%	6,000+ sentences	
	BERT-Multilingual-Cased	69.2%	6,000+	
	BERT-Multilingual-Uncased	78.7%	6,000+	
NER tagger	Static method	65.3%	4,08,837 Entity	
	SK-LEARN classification	81.2%	65,000+	
	BERT-Cased	79.2%	65,000+	
	BERT-Mutilingual-Cased	75.5%	65,000+	
	BERT-Multilingual-Uncased	90.5%	65,000+	
Word Embedding	Gensim-word2vec-100D-1,00,00,000+ tokens	-	2,00,00,000+ sentences	
	Glove-word2vec-100D-2,30,000+ tokens	-	5,00,000 sentences	
	fastext-word2vec-200D 3,00,000+	-	5,00,000 sentences	

TASK	MODEL	ACCURACY	DATASET	About
Sentence Embedding	Contextual sentence embedding	-	-----	
	Transformer embedding_hd	-	3,00,000+ human data	
	Transformer embedding_gd	-	3,00,000+ google data	
Extractive Summarization	Feature-based based	70.0% f1 score	------	
	Transformer sentence sentiment Based	67.0%	------	
	Word2vec--sentences contextual Based	60.0%	-----	
Bi-lingual projects	google translator with large data detector	----	----	
Information Extraction	Static word features	-		
	Semantic and contextual	-		

Next releases after testing this demo

Task	Version
Coreference Resolution	v1.1
Language translation	V1.1
Masked Language model	V1.1
Information retrieval Projects	V1.1
Entity Segmentation	v1.3
Factoid Question Answering	v1.2
Question Classification	v1.2
sentiment Word embedding	v1.3

Task	Version
So many others features	---

Package Installation

You have to install these packages manually, if you get any module error.

- simpletransformers
- fasttext

Models

Everything is automated here. when you call a model for the first time, it will be downloaded automatically.

With GPU or Without GPU

- With GPU, you can run any models without getting any warnings.
- Without GPU, You will get some warnings. But this will not affect in result.

Motivation

With approximately 228 million native speakers and another 37 million as second language speakers,Bengali is the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world. But still it is a low resource language. Why?

Dataset

For all sbnltk dataset and existing Dataset, see this link Bangla NLP Dataset

Trainer

For training, You can see this Colab Trainer . In future i will make a Trainer module!

When will full version come?

Very soon. We are working on paper and improvement our modules. It will be released sequentially.

About accuracy

Accuracy can be varied for the different datasets. We measure our model with random datasets but small scale. As human resources for this project are not so large.

Contribute Here

- If you found any issue, please create an issue or contact with me.

Releases

No releases published

Packages

No packages published

Languages

● **Python** 100.0%