Language Identification

SHUBHAM PATIL
ECKOVATION
PYTHON PROGRAMMING

INTRODUCTION

Intro

Language identification or language guessing is the problem of determining which natural language given content is in.

In this project, we have used **Textblob Library** for processing textual data.

It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Features of Textblob

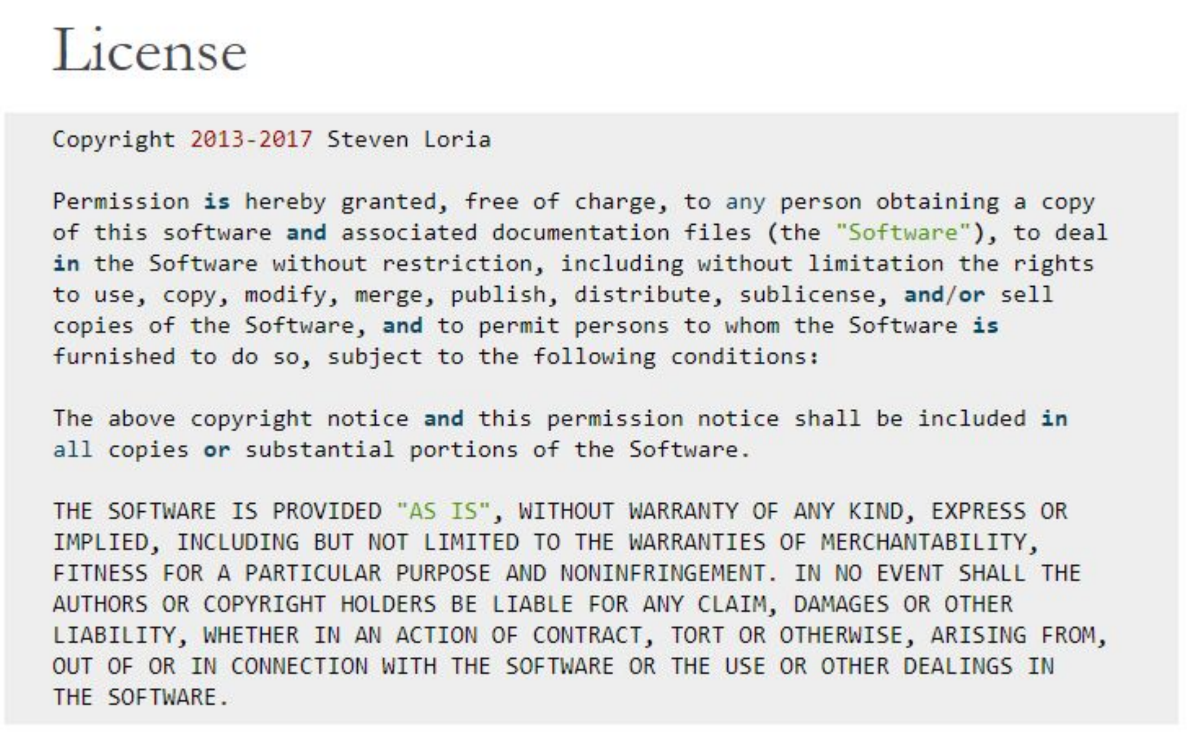
- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Language translation and detection powered by Google Translate
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- n-grams
- Word infection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions

License

TextBlob stands on the giant shoulders of NLTK and pattern.

The data sets are in JSON format, to be able to read in pandas dataframe.

Next slide will give you brief information of how textblob works.

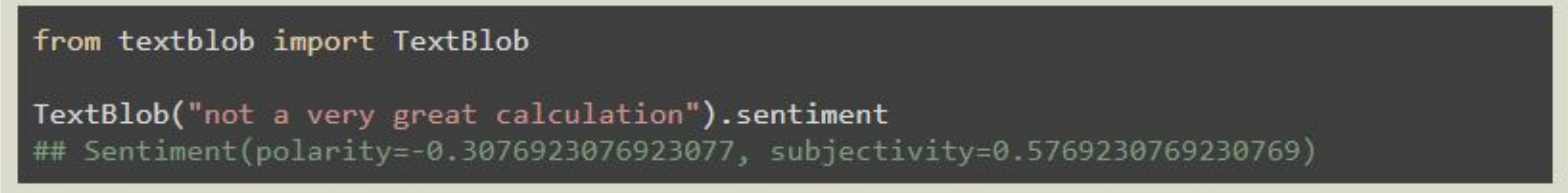


License

```
Copyright 2013-2017 Steven Loria

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in
all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN
THE SOFTWARE.
```

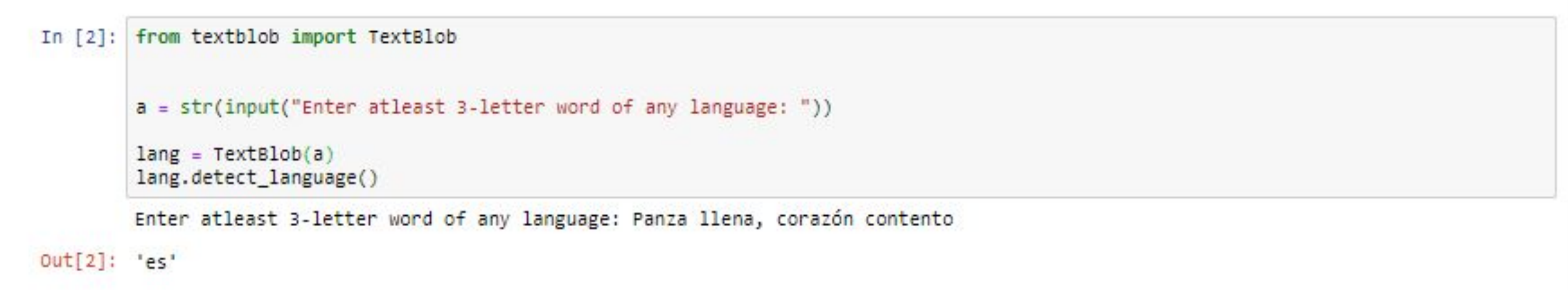
How does Textblob calculate sentiment?

Based on the polarity and subjectivity, you determine whether it is a positive text or negative or neutral. For TextBlob, if the polarity is >0, it is considered positive, <0 - is considered negative and ==0 is considered neutral.



```
from textblob import TextBlob

TextBlob("not a very great calculation").sentiment
## Sentiment(polarity=-0.3076923076923077, subjectivity=0.5769230769230769)
```

This tells us that the English phrase “not a very great calculation” has a *polarity* of about -0.3, meaning it is slightly negative, and a *subjectivity* of about 0.6, meaning it is fairly subjective.



```
In [2]:  from textblob import TextBlob


         a = str(input("Enter atleast 3-letter word of any language: "))

         lang = TextBlob(a)
         lang.detect_language()

         Enter atleast 3-letter word of any language: Panza llena, corazón contento

Out[2]:  'es'
```

Above, “Panza llena, corazón contento” is a Spanish sentence.

Hence, Textblob gives an output as “es” which determines Espanol(Spanish).

Conclusion

We showed that this project accurately estimate the proportion of the document written in each of the languages identified.

We showed that the system out performs alternative approaches from the literature on synthetic data as well as on real-world data from research on languages using the web as a resource.

Thank You!

SHUBHAM PATIL
ECKOVATION
PYTHON PROGRAMMING