Tech Review

Jingtian Yao (yao28@illinois.edu)

This review is to address the functionalities and use cases provided by NLTK and MeTA toolkit in natural language processing, compares the capability and difference between the two tools.

NLTK language toolkit is a widely used platform to work with language data in industry. It provides user friendly interface and readable documentations. With an adequate knowledge in the text information system and basic NLP concepts, NLTK is a very powerful tool for researchers and professionals. I'd like to review some of the functionalities below covered by NLTK in the article.

1. Pre-processing of Text

NLTK provides easy-to-use Python API to perform pre-processing tasks for text data. NLTK Tokenizer package can divide strings into list of substrings to find words and punctuation, or sentences in a string. NLTK Stemmers interface can be utilized to do stemming for words, with support of multiple methods.

NLTK also provides a collection of corpus (https://www.nltk.org/nltk_data/) to be imported for further usage. For example, the Stop Words corpus can be integrated with the per-processing step of text data.

2. Text Categorization

The task can be performed by calling the Classifier module. It supports classifiers including conditional exponential classifier, decision tree classifier, maxent classifier, naïve bayes classifier and Weka classifier. The logistic of training a text categorization model includes:

a. Provide the features and labels of the training dataset
b. Provide the testing dataset with features
c. Introduce the classifier from the package, train the classifier model with the training data as input
d. Classify the testing dataset, and evaluate the model

3. Text Clustering

Text clustering can be implemented by *nltk.cluster* package. It is an unsupervised machine learning problem. The package supports multiple clustering algorithms, including K-means clusterer, E-M clusterer and group average agglomerative clusterer. After the model is being trained, we can perform the below operations to an input. The operations include:

a. Cluster a sequence of vectors
b. Assign a vector to a cluster
c. Provide the probability distribution (of a vector) over cluster memberships

4. Language Model

NLTK package supports N-gram language models. It wraps up the functionalities of preprocessing such as padding both ends for a sentence in a N-gram model. It can also train model, such as using Maximum Likelihood Estimator, with the pre-processed data. We can get a list of n-gram tuples, and a list of unique vocabularies from the text. The package has the capability to auto-fill the words not occurred during training, and ignore words as we intend, like words that did not occur frequently enough during the training process.

To use a trained language model, we can do the operations below easily:

a. Get the counts for a unigram or bigram
b. Get a word's relative frequency as a score or log score
c. Evaluate the model's cross-entropy and perplexity with respect to sequences of n-grams
d. Generate text with the input of a random seed

MeTA is targeting a provide a unifying framework for text indexing and analysis methods. It is designed with the goal of performing a specific task. It modularizes the concept of feature generation, data manipulation and algorithms used, and makes the settings configurable in a configuration file, to perform a task. Similar with NLTK, it has the capability to do basic text analysis like tokenization, stemming and filtering, also tasks like text categorization. Regarding creating search engine, MeTA also wraps up algorithms such as Okapi BM25, Pivoted Length normalization for documents etc. Targeting at topic modeling, MeTA has wrapped up topic model algorithms like LDA in the package as well, makes it a desirable tool to do topic mining.