Technology Review

Text Information Systems 2021 Fall

Zhengkai Zhang(Net ID: zz68)

Toolkit: Apache OpenNLP

Introduction:

Apache OpenNLP is a machine learning based natural language processing tool that supports most of the common NLP tasks. The learning algorithm is very generic and supports most languages. It can perform tasks such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks usually required advanced text processing services. Apache OpenNLP is mostly all free and it is a framework where users can train their own model to work on their own tasks.

It is an open source framework in Java language and has a strong community support for it. It is well documented and nice for beginners to learn. The language was initially released in 2004, and entered Apache incubation in 2010. It became a top level apache project in 2011 at version 1.5.2.

Features of OpenNLP:

- **Named Entity Recognition (NER)**

  Named entity recognition also know as entity identification/entity chunking/entity extraction is a subtask of information extraction. It is classifying unstructured text of named entities into categories such as names, orgs, countries, area code, and others.

  − Open NLP supports NER, using which you can extract names of locations, people and things even while processing queries.

- **Summarize**

  Text summarization in NLP is the process of summarizing the information in large texts for quicker consumption.

  − Open NLP Using the summarize feature, user can summarize Paragraphs, articles, documents or their collection in NLP.

- **Searching**

  OpenNLP supports a powerful feature, that with a given user input string or synonyms it can be identified even if it is misspelled or altered.

- **Tagging (POS)**

  Part of speech tagging is also called grammatical tagging. It is the process of marking up a word in text as a particular part of speech. It identify of word usch as nouns, verbs adjectives and adverbs.

In OpenNLP, it supports POS to divide text into various grammatical elements and support further analysis.

- **Information grouping**

  This option in NLP groups the textual information in the content of the document, just like Parts of speech which supported in the Apache Open NLP.

- **Natural Language Generation**

  Natural language generation is a subfield of artificial intelligence, it is a method that automatically transforms data into plain English Content. OpenNLP uses it to generate information from a database and automate the report such as weather or medical reports.

- **Feedback Analysis**

  Feedback analysis is use to find the needs and frustrations of users therefore it can improve the user satisfaction. OpenNLP support various types of feedbacks from user and collect them. It use NLP to analyze how well the product success.

- **Speech recognition**

Speech recognition is the ability to identify words and speech into readable text. It is difficult to analyze human speech, but open NLP has some powerful built-in features that support this requirement.

Conclusion:

Apache OpenNLP is a great open source toolkit that can provide a lot of value to text retrieval and text mining. It has a lot of features in NLP. Above content gives a brief introduction and feature list for this tooling. It is good for our students to learn and use it. It could enable us to do a lot with the concept of this text information system course, and it could be useful in the final project.