CS410 Technology Review

Topic: Apache OpenNLP

Author: Sagar Dalwadi (Net ID: sagardd2)

What is OpenNLP?

Apache **OpenNLP** is an open-source Java library which is used to process natural language text. You can build an efficient text processing service using this library.

OpenNLP provides services such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution, etc.

Features of OpenNLP

Following are the notable features of OpenNLP −

- **Named Entity Recognition (NER)** − Open NLP supports NER, using which you can extract names of locations, people and things even while processing queries.
- **Summarize** − Using the **summarize** feature, you can summarize Paragraphs, articles, documents or their collection in NLP.
- **Searching** − In OpenNLP, a given search string or its synonyms can be identified in given text, even though the given word is altered or misspelled.
- **Tagging (POS)** − Tagging in NLP is used to divide the text into various grammatical elements for further analysis.
- **Translation** − In NLP, Translation helps in translating one language into another.
- **Information grouping** − This option in NLP groups the textual information in the content of the document, just like Parts of speech.
- **Natural Language Generation** − It is used for generating information from a database and automating the information reports such as weather analysis or medical reports.

- **Feedback Analysis** − As the name implies, various types of feedbacks from people are collected, regarding the products, by NLP to analyze how well the product is successful in winning their hearts.
- **Speech recognition** − Though it is difficult to analyze human speech, NLP has some built-in features for this requirement.

Different ways to use OpenNLP

- **OpenNLP API -** The Apache OpenNLP library provides classes and interfaces to perform various tasks of natural language processing such as sentence detection, tokenization, finding a name, tagging the parts of speech, chunking a sentence, parsing, co-reference resolution, and document categorization.

- **OpenNLP CLI -** In addition to the library, OpenNLP also provides a Command Line Interface (CLI), where we can train and evaluate models.

Pricing

OpenNLP is freely available under the Apache 2.0 open source license. This includes the right to freely modify and distribute anything built with OpenNLP and is also compatible with the GPL 3 license. However, it's important to note that training data may be licensed separately. Training data meant for use with OpenNLP can range from free Apache licensed packages to a standard paid license. The exact costs of training data will vary on a case by case basis.

Conclusion

OpenNLP offers a powerful and low-cost natural language processing API. And it's distributed under a license which allows for any type of modification or expansion. It's well suited to environments where programmers can create custom solutions as needed. Below are some of the Pros & Cons of Apache OpenNLP:

Pros

- It's an easy to use API and understand
- Shallow learning curve and detailed documentation with lots of examples
- Covers a lot of NLP functionality.

- Lots of resources available to learn more about NLP

Cons

- It seems the development is slow or has been stagnated
- A few models are missing when going through the examples in the documentation
- The current models provided may need further training as per your use case(s).

References

- https://www.javaadvent.com/2019/12/exploring-nlp-concepts-using-apache-opennlp.html
- https://en.wikipedia.org/wiki/Apache_OpenNLP
- https://www.tutorialspoint.com/opennlp/opennlp_overview.htm