CS 410 Technology Review

Introduction

Natural Language Processing entails processing and extracting information from natural language such as text or speech. There are many frameworks to do Natural Language Processing (NLP), two examples of which are Apache OpenNLP and Amazon Comprehend. In this review I am going to review each of these frameworks and compare the two to determine which is better for text-based NLP tasks.

Overview of Apache OpenNLP

OpenNLP is an open-source Java based framework which is licensed for use under the Apache License 2.0. It supports the following NLP tasks:

- Tokenization
- Sentence Segmentation
- Part-of-speech tagging
- Named entity extraction
- Chunking
- Parsing
- Language detection
- Coreference resolution

The OpenNLP framework makes tools accessible from the Java based APIs as well as a command line interface for testing. Since it is open source, it is free to use and modify. OpenNLP has its own format for models called OpenNLP models but also supports ONNX models which can be trained in other frameworks such as PyTorch or Tensorflow. OpenNLP may require having domain knowledge of NLP to support training models which can then be used to perform the NLP tasks especially in domains where pre-trained models are not as helpful.

Overview of Amazon Comprehend

Amazon Comprehend is an NLP service provided by Amazon Web Services (AWS). It closed source and follows a pay per use model where the user is charged depending on the NLP task done. Comprehend supports the following NLP tasks.

- Custom entity recognition
- Custom classification
- Entity recognition
- Sentiment analysis
- Targeted Sentiment
- PII Identification and Redaction
- Keyphrase extraction
- Events detection

- Language detection
- Syntax Analysis
- Topic Modeling
- Multi-language support

Amazon comprehend supports a wide variety of NLP tasks without needing much domain knowledge in NLP. It provides APIs which can be called from a wide variety of methods from SDKs for most popular programming languages and AWS CLI commands.

Comparison of OpenNLP vs Comprehend

Functionality:

In terms of functionality both OpenNLP and Amazon Comprehend perform a wide variety of NLP tasks but Comprehend outperforms OpenNLP with the ability to detect and redact Personally Identifiable Information (PII) as well as perform key phrase extraction to get the main talking points of text. Comprehend is also constantly improving the model performance and adding new features so it gets better over time where the OpenNLP model needs to be changed to see improvements.

Ease of use:

Amazon Comprehend requires no upfront training or machine learning skills. It provides APIs accessible from every popular programming language and a command line interface as well as easy integration with other AWS services such as S3 and AutoML. On the other hand, OpenNLP APIs are only accessible through Java but there is also a command line interface like Comprehend. Additionally, it is harder to get up and running because it requires more machine learning knowledge to train a model to support custom domains. However, with Comprehend if the application that is performing these NLP tasks does not have connectivity to the AWS endpoints it will not be able to perform them unlike with OpenNLP which will work even without internet access.

Cost:

OpenNLP is free and open source which means it is free to use and modify. Comprehend on the other hand charges per API call with a price depending on the NLP task. For most of the NLP tasks comprehend charges 0.0001$ per API call with discounts for high volumes of API calls. Using Amazon Comprehend also requires creating an AWS account unlike with OpenNLP.

Conclusion

Overall, both OpenNLP and Amazon Comprehend are both great frameworks for performing NLP tasks on text and documents. Amazon Comprehend has the edge in terms of functionality but the best framework depends on the application. For simple NLP tasks running on personal hardware OpenNLP is the clear winner in addition to being free but for larger NLP tasks with integration to other applications Amazon Comprehend is better for the job.

References

- https://opennlp.apache.org/docs/2.0.0/manual/opennlp.html#intro.models
- https://en.wikipedia.org/wiki/Apache_OpenNLP
- https://www.baeldung.com/apache-open-nlp
- https://aws.amazon.com/comprehend/
- https://aws.amazon.com/comprehend/features/
- https://aws.amazon.com/comprehend/pricing/
- https://docs.aws.amazon.com/comprehend/latest/dg/what-is.html
- https://docs.aws.amazon.com/comprehend/latest/APIReference/API_Operations.html