Mathew Shaw
NetID: mcshaw2

AWS Comprehend

Amazon offers many different cloud software services through Amazon Web Services (AWS) that can be integrated into existing or greenfield applications to add functionality that could take a sizable software team at your company months, or years, to develop independently. The AWS Comprehend service provides access to machine learning based NLP services via common web APIs that can easily be accessed through SDKs in multiple programming languages. Your team can supply your own text collection generated by your business to the Comprehend system and, without requiring developers to train machine learning models or even have machine learning knowledge, extract value, insights or optimizations usable across your applications or team. Comprehend can take in unstructured text data and analyze documents for sentiment analysis; generate keywords and entities to enhance your text based search; and identify topics across your corpus, assisting in providing recommendations for your users on related documents. Being a cloud SaaS, there are some potential downsides regarding costs and data security that your team will have to accept in order to use the service.

Sentiment analysis is one feature provided by Comprehend. If your team works with products that collect user reports, feedback or interactions, this could allow your developers to add functionality to track user opinion of your products/services over time. Submitting a text sample to the Comprehend sentiment analysis API returns a response with confidence scores for Positive, Neutral, Negative, and Mixed sentiments, with the confidence scores summing to one. This provides a straightforward, quantitative estimate for either display in user dashboards or further processing/computation in your application.

Providing efficient text search over a large body of unstructured text corpus is a common challenge. Comprehend can help enhance your application's text search via keyphrase extraction and entity recognition. Each document can be processed with these two services to generate lists of keyphrases and entity labels with associated confidence scores. You can then sort or filter by a desired confidence level to obtain the top N phrases/entity labels and assign them to the document. This could provide the basis for a hashtag style keyword search in your application as an alternative or addition to other types of direct text search.

Recommendation systems are another common feature in many modern applications, and this functionality requires the ability to relate a particular item to others in your collection. This functionality can be enhanced by using Comprehend to generate topic categories for your corpus, as well as labels for individual documents. These

labels could then be used as a factor for a similarity function in your recommendation system, or as a means to allow users to browse your entire corpus by topic. Classifying incoming new documents into topics as they arrive could also help with routing these new documents to different departments in your organization, for example incoming help desk tickets could be sent to departments based on the most likely topic(s) generated for the document.

The largest selling point of adapting Comprehend for your team is to leverage the service as an NLP machine learning based SaaS. The value this provides for your team is largely offloading the responsibility of developing, training, and maintaining text based machine learning models for your use cases. This is a significant effort for any team, likely requiring a dedicated team, potentially requiring data scientists, data engineers, and machine learning aware software developers, to execute in house. AWS has access to massive amounts of unstructured text based data to continually train and evaluate its models with. By choosing to leverage Comprehend to satisfy these parts of the requirement for your project, you reduce the required team size, as well as removing or reducing the need for the team to have specialized machine learning knowledge. This can all be done with the unstructured text data collection your team already has without requiring any customizations to the Comprehend models.

Being a cloud based SaaS, Comprehend has two major drawbacks: cost and data security. Comprehend requires paying for use of the services in “Units” (defined by AWS as 100 characters for 1 Unit). The pricing model is pay as you use, with price per unit decreasing the more you use per month. A cost-benefit analysis will have to be done for your team’s situation to determine if this indefinite monthly cost is preferable to a larger up front in house development effort to build and train machine learning models. The second downside comes from the nature of a cloud based service: to use any of the Comprehend services you must upload your text data to the Amazon cloud, generally S3 data buckets, in order to use the service. This necessitates exposing your data to a cloud your company does not own. Amazon has extensive data security and privacy policies outlined on their website, and is also trusted by many large companies and the US Government for data security, but if your company policy prohibits exposing your data in any manner this may disqualify Comprehend from consideration.

AWS Comprehend provides several benefits to teams that require text based machine learning to satisfy the requirements of their applications. Comprehend APIs allow developers with no machine learning knowledge to integrate sentiment analysis, expand text search effectiveness, and perform topic identification/classification into applications. While Comprehend is a SaaS with a required monthly payment, the cost is pay-as-you-go and could potentially provide a high amount of value when compared with in house development and training of machine learning models, as long as your team has permission to expose your text corpus to the Amazon cloud.