

Microsoft Text Information System Tools

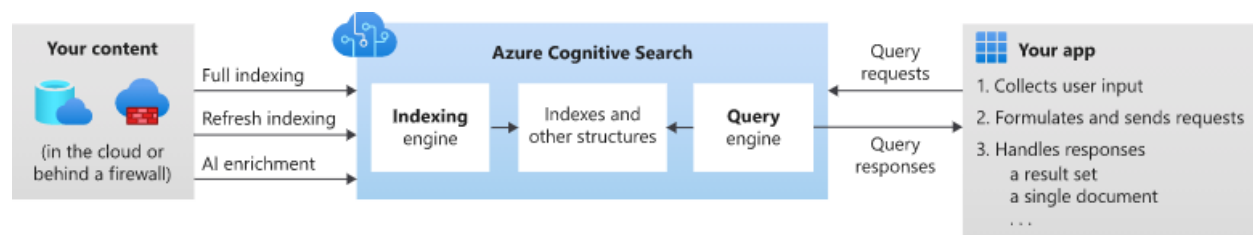
Intro

Microsoft has a wide range of services that process text data and help create text information systems. While there are many open source and competitor offerings for text retrieval, text mining, and NLP tools from Google and Amazon, researchers and engineers in industry might benefit from Microsoft's offerings if they are already invested in the Azure Cloud or utilize a Microsoft framework such as .NET. Unfortunately, it can be a bit overwhelming understanding the function of all of Microsoft's tools given the marketing influenced titles of services and unclear overlap between their own services. This paper seeks to clarify the use cases of each tool. Effectively, Microsoft's text information system tools can be categorized into 1) Azure Cognitive Search, 2) Azure Cognitive Services, and 3) Azure Machine Learning.

Body

Azure Cognitive Search

The most holistic tool Microsoft provides is Azure Cognitive Search (also known as "Azure Search"). The Azure Cognitive Search tool provides a framework for a search engine for applications. Unlike some of their other services, there is no stipulation for the content needing to be on the Azure cloud itself, rather it utilizes Azure resources to create the index and query engine of the text retrieval system:



Data in the system is persisted as an inverted index, though the implementation of this is not transparent to users. This service also appears to be one of the most granular offerings, allowing developers to create scoring profiles, which is an abstraction of the weighting BM25 or other ranking algorithms would use – for instance, one can specify newer items get higher rankings in their scoring profile. While the algorithm implementation is obfuscated to users, the tool does claim to use Okapi BM25 and provides control over $k1$ and b parameters for term frequency and document length score adjustment. Creation of the index is up to the developer to setup, though the majority of the work is automated. Developers simply need to use the REST APIs Azure provides to define the data structure of the text data, then upload the text data itself. One downside of their service is that the text data must be uploaded via JSON format, which means users would have to transpose existing text data that could be XML or plain text into JSON format. Once the data is indexed, querying is as simple as utilizing GET requests with a variety of parameters, such as filters, ranges, and expressions more akin to database queries like SQL. Nevertheless, it is distinct from pure database search tools by Microsoft such as SQL Server and Cosmos DB because it consists granular features like stemming, lemmatization, and scoring controls. It also is different than the similarly named Bing Search APIs – notably Bing Web Search and Bing Custom Search that leverage Bing.com's indexes to create custom search engines specifically for the web or a subset of web pages.

Azure Cognitive Services

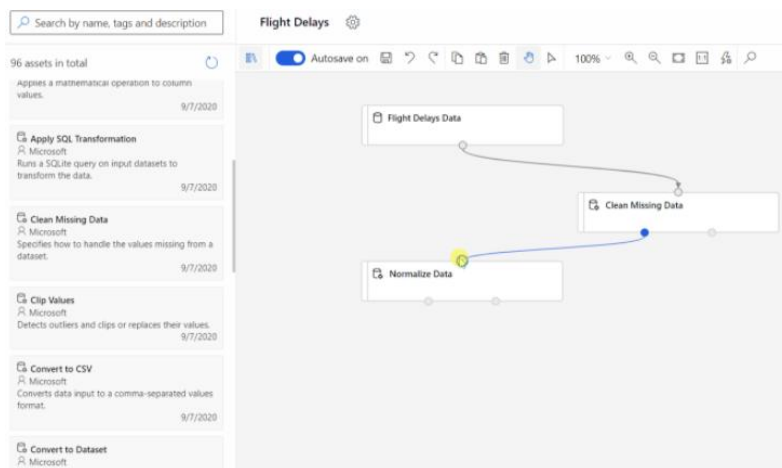
Azure Cognitive Services, unlike Azure Cognitive Search, is a suite of REST APIs and SDKs that leverage on-cloud algorithms not transparent to users to process text data along with other data, but does not create an end-to-end text retrieval system. Their five main categories of services include Vision, Speech, Language, Decision, and Search APIs, though focus will only remain on Language and Search as the others are not specialized for text data processing.

The Search APIs under Azure Cognitive Services include the aforementioned Bing Web Search and Bing Custom Search, alongside other Bing APIs for News, Business data, and spell check. These APIs are naturally integrated into Microsoft SDKs for C#, Java, JavaScript, and Python, as well as their SaaS offerings such as Power Automate.

The Language APIs are distinctly specialized as NLP algorithms for users to invoke and process their data. For example, the Translator API can be integrated into websites and applications to create multi-language user experiences. On the other hand, the Text Analytics API includes sentiment analysis, key phrase extraction, named entity recognition, and language detection, which may be more useful for data analytics applications. Like the Search APIs, these Language APIs do not reveal implementation details of their algorithms but are rather simply invoked through REST or various language's SDKs that Microsoft has developed. The main downside to this though is that a developer cannot adjust the algorithm's parameters, such as the granularity of sentiments provided by a sentiment analysis.

Azure Machine Learning

Azure Machine Learning is the most generic of the 3 main categories, as it seeks to be a general cloud-based Machine Learning solution. It provides tools to “to train, deploy, automate, manage, and track ML models.” It also enables developers to use either a Microsoft SDK in a language such as Python or utilize Azure Machine Learning Studio, which is a drag and drop interface that users can use to connect data sources and create models without writing code:



Consequently, it can be a powerful tool for text information system tool that supports supervised, unsupervised, and deep learning on text data, although developers will have even less insight into how data is processed since the model will learn by training instead of explicit programming of an algorithm. One downside to this is that one cannot train their model offline and must provision and use Azure servers to train and use their model.

Conclusion

Through their various services, Microsoft can in fact be an appropriate source of tools for text information systems. It seems though that these SDKs and APIs have all become tightly integrated with Azure and the algorithms obfuscation can be a downside for developers in need of fine-tuning. However, the company utilizes state of the art algorithms such as Okapi BM25 and can significantly decrease development time with their libraries. If one seeks to create a search engine, recommendation system, or other text retrieval model, Azure Cognitive Search provides all the necessary pipelines and framework from data ingestion to indexing, ranking, and querying. On the other hand, if one requires more text mining and analytics, the Azure Cognitive Services APIs provide a way to leverage Microsoft's proprietary sentiment analysis, entity recognition, and other algorithms alongside Bing's sizable technologies for web site ranking and crawling. If a user requires classification or other text processing tasks that go beyond Azure Cognitive Search and Azure Cognitive Services, they can leverage the Azure Machine Learning Studio to create, train, and deploy scalable machine learning models. Microsoft has done a good job with documentation of each of these individual categories, providing tutorials and videos, even though they do not possess a page that succinctly distinguishes the services. It is a great offering for those already invested in the Azure ecosystem as well as for those seeking low code but well-supported tools for text retrieval, mining, and machine learning.

Citations

<https://docs.microsoft.com/en-us/azure/?product=ai-machine-learning>

<https://docs.microsoft.com/en-us/azure/search/search-what-is-azure-search>

<https://docs.microsoft.com/en-us/azure/search/index-ranking-similarity>

<https://docs.microsoft.com/en-us/azure/search/search-query-simple-examples>

<https://docs.microsoft.com/en-us/azure/cognitive-services/Bing-Custom-Search/overview>

<https://docs.microsoft.com/en-us/azure/cognitive-services/what-are-cognitive-services>

<https://docs.microsoft.com/en-us/azure/cognitive-services/bing-web-search/>

<https://docs.microsoft.com/en-us/azure/cognitive-services/translator/translator-info-overview>

<https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview>

<https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml>

<https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-machine-learning-studio>