Choosing a natural language processing technology in Azure

02/12/2018 • 2 minutes to read • Contributors

In this article

Free-form text processing is performed against documents containing paragraphs of text, typically for the purpose of supporting search, but is also used to perform other natural language processing (NLP) tasks such as sentiment analysis, topic detection, language detection, key phrase extraction, and document categorization. This article focuses on the technology choices that act in support of the NLP tasks.

What are your options when choosing an NLP service?

In Azure, the following services provide natural language processing (NLP) capabilities:

- Azure HDInsight with Spark and Spark MLlib
- Azure Databricks
- Microsoft Cognitive Services

Key selection criteria

To narrow the choices, start by answering these questions:

- Do you want to use prebuilt models? If yes, consider using the APIs offered by Microsoft Cognitive Services.
- Do you need to train custom models against a large corpus of text data? If yes, consider using Azure HDInsight with Spark MLlib and Spark NLP.
- Do you need low-level NLP capabilities like tokenization, stemming, lemmatization, and term frequency/inverse document frequency (TF/IDF)? If yes, consider using Azure HDInsight with Spark MLlib and Spark NLP.
- Do you need simple, high-level NLP capabilities like entity and intent identification, topic detection, spell check, or sentiment analysis? If yes, consider using the APIs offered by Microsoft Cognitive Services.

Capability matrix

The following tables summarize the key differences in capabilities.

General capabilities

Capability	Azure HDInsight	Microsoft Cognitive Services
Provides pretrained models as a service	No	Yes

Capability	Azure HDInsight	Microsoft Cognitive Services
REST API	Yes	Yes
Programmability	Python, Scala, Java	C#, Java, Node.js, Python, PHP, Ruby
Support processing of big data sets and large documents	Yes	No

Low-level natural language processing capabilities

Capability	Azure HDInsight	Microsoft Cognitive Services
Tokenizer	Yes (Spark NLP)	Yes (Linguistic Analysis API)
Stemmer	Yes (Spark NLP)	No
Lemmatizer	Yes (Spark NLP)	No
Part of speech tagging	Yes (Spark NLP)	Yes (Linguistic Analysis API)
Term frequency/inverse-document frequency (TF/IDF)	Yes (Spark MLlib)	No
String similarity—edit distance calculation	Yes (Spark MLlib)	No
N-gram calculation	Yes (Spark MLlib)	No
Stop word removal	Yes (Spark MLlib)	No

High-level natural language processing capabilities

Capability	Azure HDInsight	Microsoft Cognitive Services
Entity/intent identification and extraction	No	Yes (Language Understanding Intelligent Service (LUIS) API)
Topic detection	Yes (Spark NLP)	Yes (Text Analytics API)
Spell checking	Yes (Spark NLP)	Yes (Bing Spell Check API)
Sentiment analysis	Yes (Spark NLP)	Yes (Text Analytics API)
Language detection	No	Yes (Text Analytics API)
Supports multiple languages besides English	No	Yes (varies by API)

See also

Natural language processing