Azure Cognitive Search- analyzer

1. 词语分析器：

Analyzers 负责在索引和查询执行期间处理字符串。文本处理中，analyzers 用于标记术语。

[Analyzers for linguistic and text processing - Azure Cognitive Search | Microsoft Learn](https://learn.microsoft.com/en-us/azure/search/search-analyzers#language-analyzer-example)

通过postman的调用方式，查看不同分词器，在处理过程中的区别。

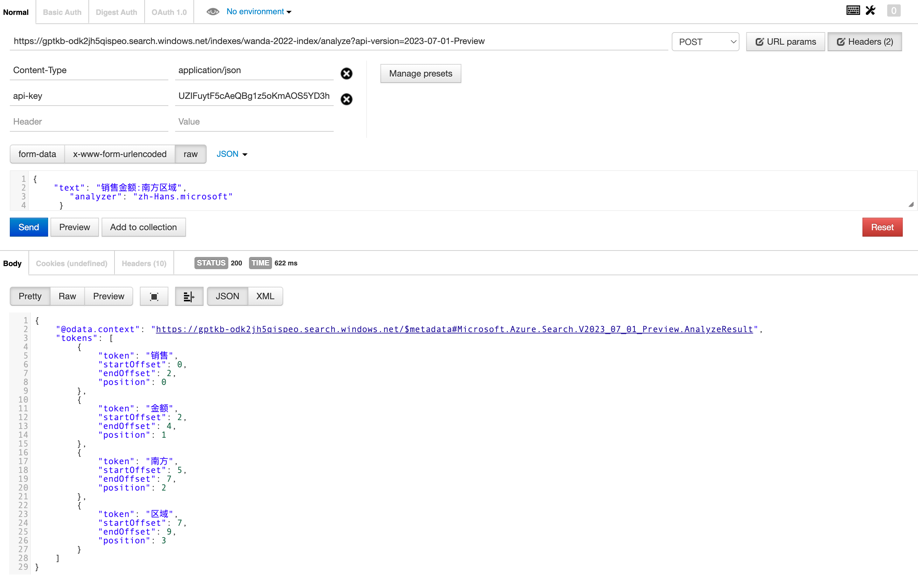
默认的Lucene Standard对中文的分词处理：

因为中文字没有空格的概念，所以它就是每一个字作为一个单位标记处理：

A screenshot of a computer

Description automatically generated

如果选择中文的处理词：（它有词汇的概念，这样词汇变成一个单元。在查询的时候能更准确）



1. 分词器与语义搜索的关系

语义搜索是一项功能，通过利用语言理解能力，可以显著提高搜索结果的相关性。要更好地了解这个功能。

[Semantic search - Azure Cognitive Search | Microsoft Learn](https://learn.microsoft.com/en-us/azure/search/semantic-search-overview)

1. 它单独一个功能，但是它是基于基本搜索的top50重新排序（[BM25-ranked search result](https://learn.microsoft.com/en-us/azure/search/index-ranking-similarity) ）
2. 它对给定的文档，进行总结提取256 token
3. 基于256 token 和 查询内容机型语义比较

它是基于Bing and Microsoft Research的实现。

1. 支持的语言分析器

[Add language analyzers to string fields - Azure Cognitive Search | Microsoft Learn](https://learn.microsoft.com/en-us/azure/search/index-add-language-analyzers)

A screenshot of a computer

Description automatically generated

1. Microsoft 和 Lucene分析器的比较

Azure 认知搜索支持 35 个受 Lucene 支持的语言分析器，以及 Office 和必应中使用的专有 Microsoft 自然语言处理技术支持的 50 个语言分析器。

某些开发人员可能首选更熟悉、简单的开源 Lucene 解决方案。 Lucene 语言分析器速度更快，但 Microsoft 分析器具有高级功能，如德语、丹麦语、荷兰语、瑞典语、挪威语 (、爱沙尼亚语、芬兰语、匈牙利语、斯洛伐克语) 和实体识别 (URL、电子邮件、日期、数字) 。 如果可能，应对 Microsoft 和 Lucene 分析器进行比较以确定哪一个更合适。 可以使用分析 API 查看使用特定分析器从给定文本生成的标记。

Microsoft 分析器的索引平均比 Lucene 的索引慢两到三倍，具体取决于语言。 对于平均大小的查询，搜索性能应该不会受到显著影响。

1. 默认的Lucene分析器

默认分析器为 Standard Lucene，它适用于英语，但可能不如 Lucene 的英语分析器或 Microsoft 的英语分析器那样适用。

* Lucene 的英语分析器扩展了标准分析器。 它从字词中删除所有格（尾部的 's）、根据 Porter 词干分解算法应用词干分解，并删除英语非索引字。
* Microsoft 的英语分析器执行词形还原，而不是词干分解。 这意味着它可以更好地处理发生了词尾变化的字词形式以及不规则的字词形式，从而产生相关度更高的搜索结果。

***它最中文的处理，如上截图，这样就减少了token里面包含的文本数量，同时其他其他高级功能，拼音，同义词等就没有办法处理了。***

[用于语言和文本处理的分析器 - Azure Cognitive Search | Microsoft Learn](https://learn.microsoft.com/zh-cn/azure/search/search-analyzers#default-analyzer)

在 Azure 认知搜索中，分析器会自动调用标记为“可搜索”的所有字符串字段。

Azure 认知搜索默认使用 Apache Lucene 标准分析器 (standard lucene)，该分析器按照“Unicode 文本分段”规则将文本分解成多个元素。 标准分析器会将所有字符转换为其小写形式。 已编入索引的文档和搜索词在索引和查询处理期间完成分析。



A close-up of a white background

Description automatically generated

1. 修改分析器

[向字符串字段添加自定义分析器 - Azure Cognitive Search | Microsoft Learn](https://learn.microsoft.com/zh-cn/azure/search/index-add-custom-analyzers)

定义分析器、tokenizer、标记筛选器或字符筛选器后，便无法修改它。 仅当 allowIndexDowntime 标志在索引更新请求中设置为 true 时，才可向现有索引添加新的上述内容：

PUT https://[search service name].search.windows.net/indexes/[index name]?api-version=[api-version]&allowIndexDowntime=true