Creating an Azure AI Search Solution

Sunday, November 24, 2024 8:16 PM

Reference

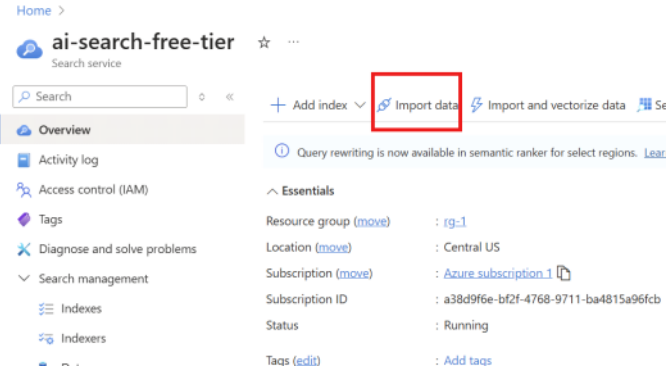
- https://microsoftlearning.github.io/mslearn-knowledge-mining/Instructions/Exercises/01-azure-search.html

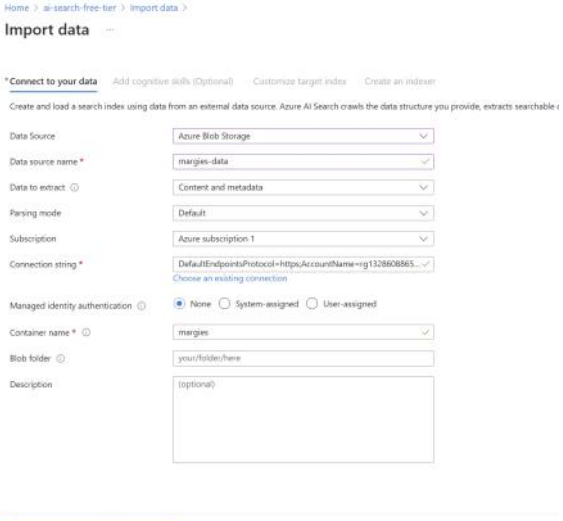
Upload Data to AZ storage account

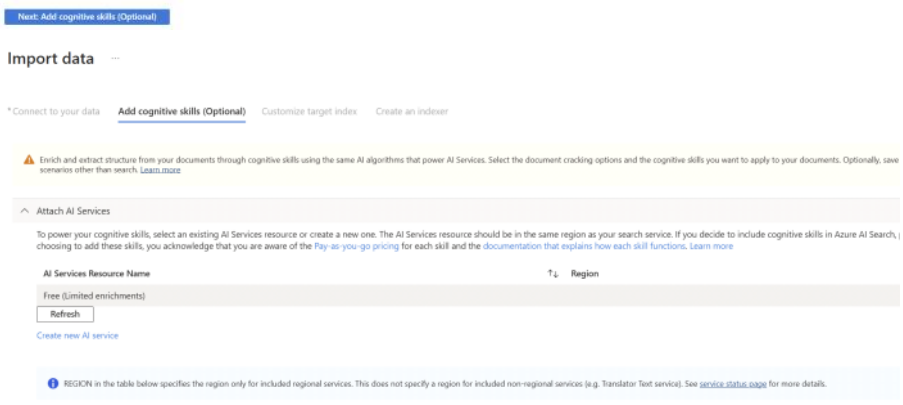
- Run the below script
  - It creates a container called margies
  - It uploads data from local to this container

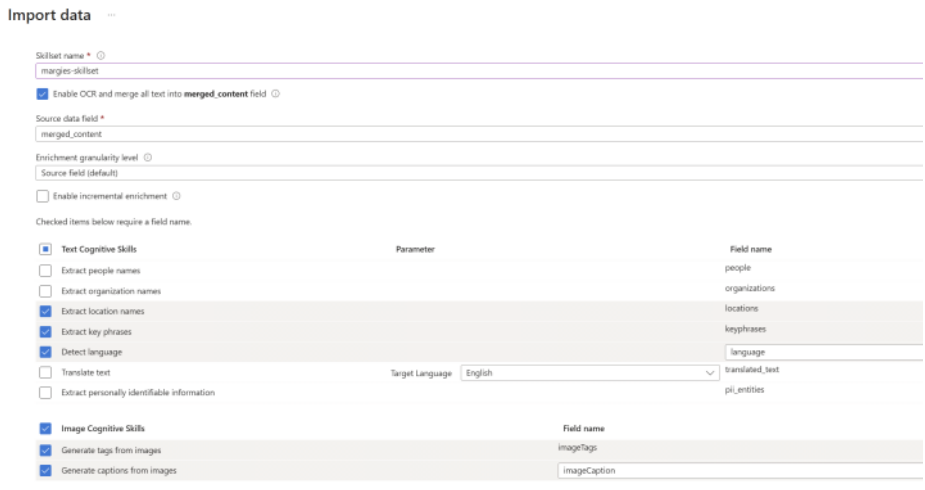
```
@echo off
SETLOCAL ENABLEDELAYEDEXPANSION
rem Set values for your storage account
set subscription_id=a38d9f6e-bf2f-4768-9711-ba4815a96fcb
set azure_storage_account=rg13286088654
set
azure_storage_key=UaLKAGfjNmI0LfesRtDJUQvDZykrehPSujINxq2quHt4yxtEa3FcEy0aC6a3Y4I
IC/jtYstUQGb5+ASt8B0W1g==

echo Creating container...
call az storage container create --account-name !azure_storage_account! --
subscription !subscription_id! --name margies --auth-mode key --account-key !
azure_storage_key! --output none
echo Uploading files...
call az storage blob upload-batch -d margies -s data --account-name !
azure_storage_account! --auth-mode key --account-key !azure_storage_key!  --
output none
```

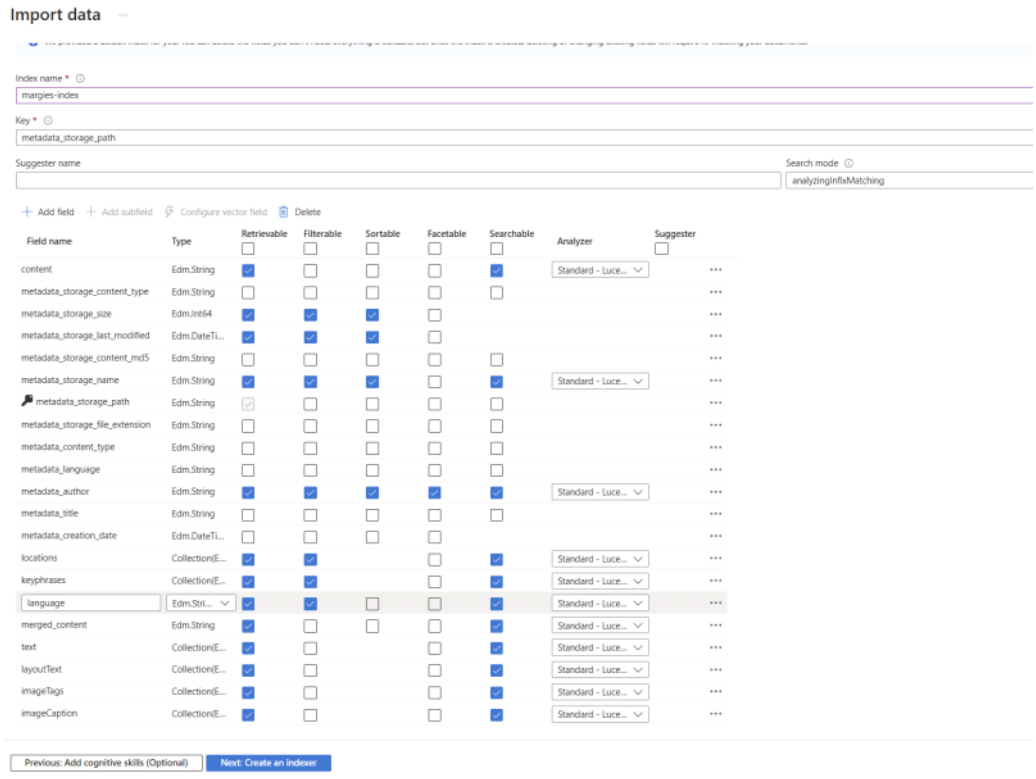
Index the Documents

- 

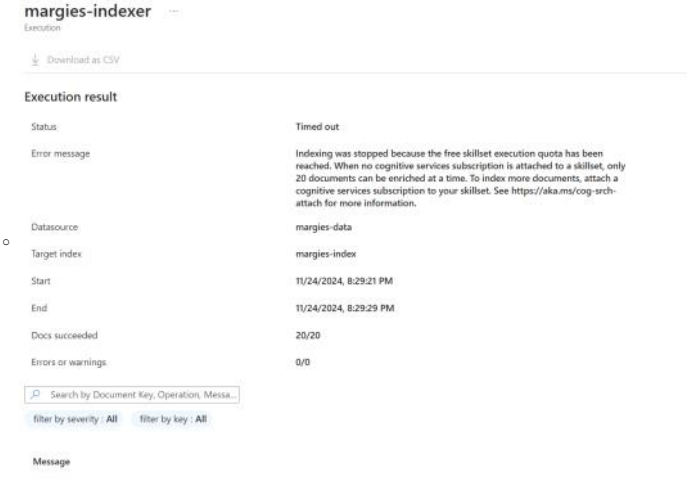
- 

- 





- Since I am on free tier - skillset is limited to 20 documents thus the indexer times out with the above settings on free tier



  - If I perform the same without adding skillset - it succeeds on free tier

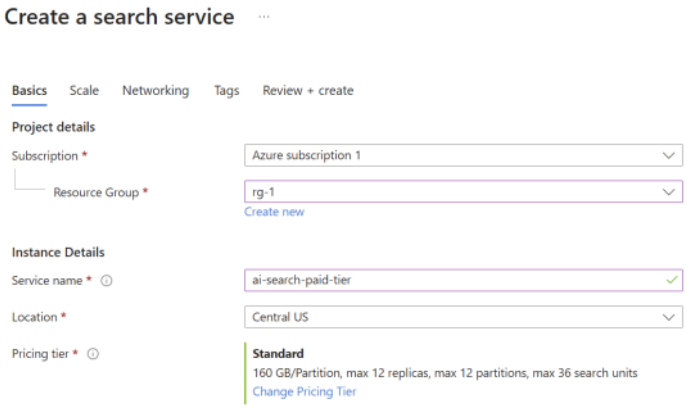
Search the Index

- Can be followed here - https://microsoftlearning.github.io/mslearn-knowledge-

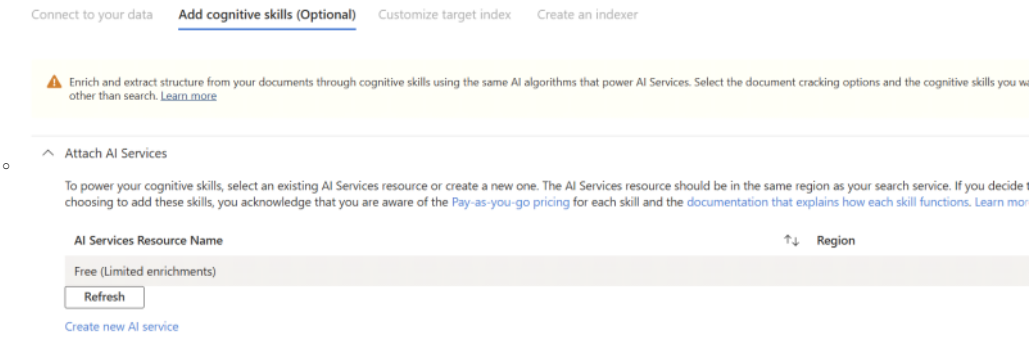
mining/Instructions/Exercises/01-azure-search.html#search-the-index

Trying the above with a Paid tier AI Search Service

- Estimated cost per month = $250

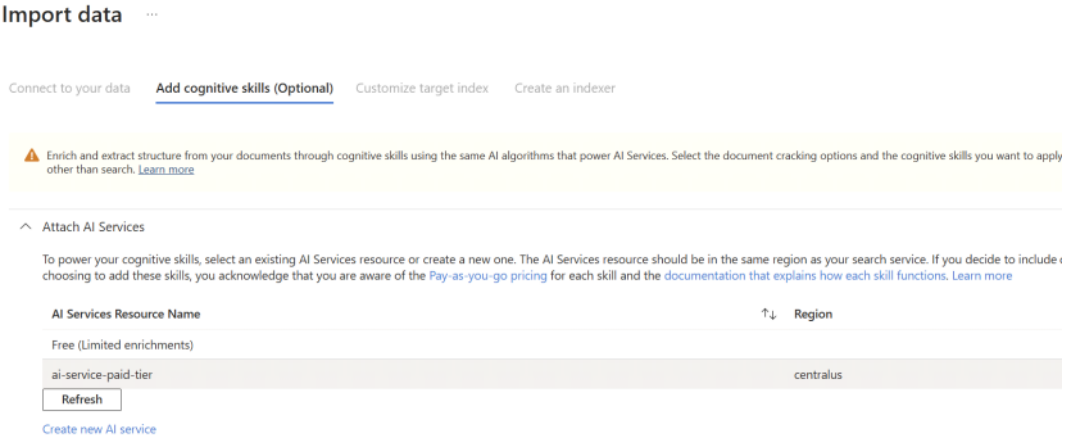


- Ran steps same as Index the Documents
- Had the same error of 20 documents limit - looks like AI search isn't limiting but AI services is limiting with the enrichment

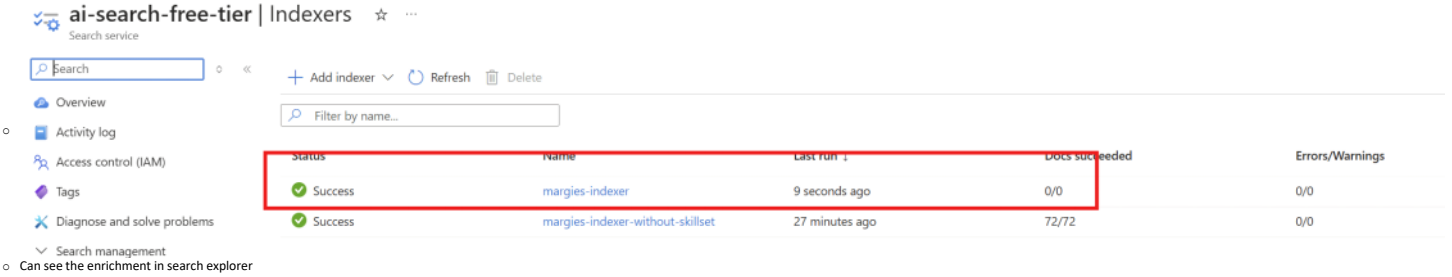


Trying with a new paid AI Service Resource

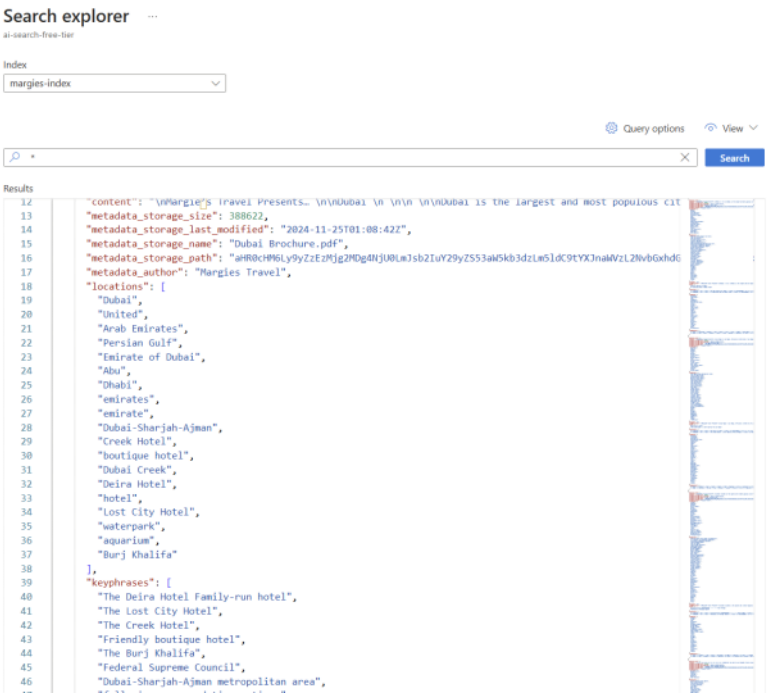
- From here - Create new AI service



- This time the indexer with skillset succeeded

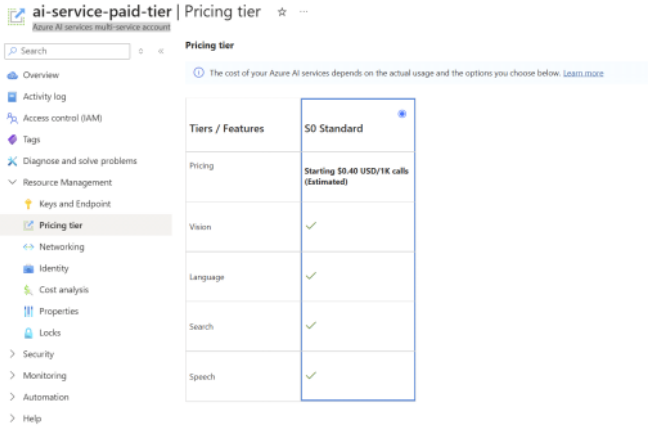


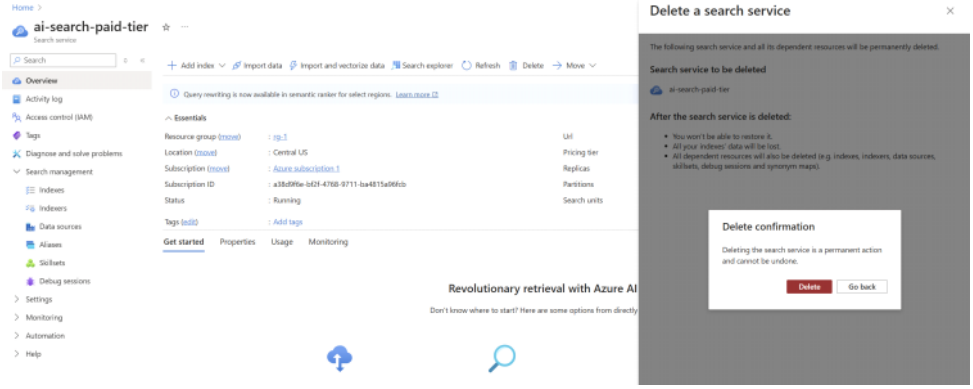
- Can see the enrichment in search explorer



Deleting the created PAID AI Search Service

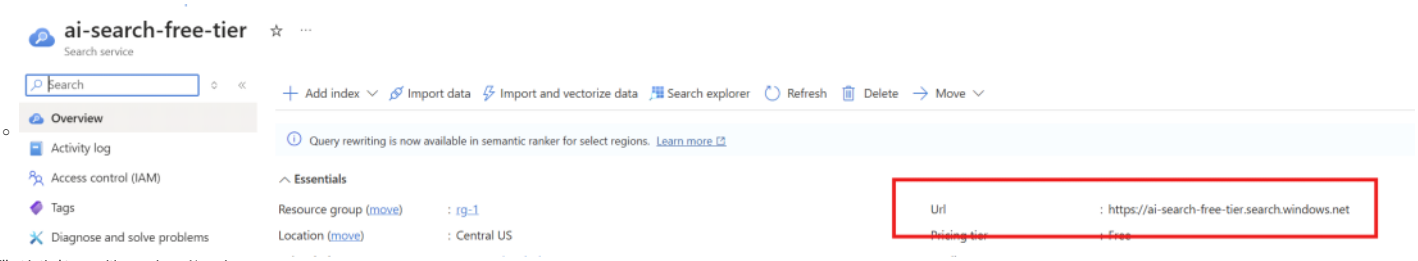
- AI search is not pay-as-you-go. It is a fixed hourly bill
- Not required as we can use margies data with free tier- Although we had to create a paid ai service
  - https://portal.azure.com/#@deril12341outlook.onmicrosoft.com/resource/subscriptions/a38d9f6e-bf2f-4768-9711-ba4815a96fcb/resourceGroups/rg-1/providers/Microsoft.CognitiveServices/accounts/ai-service-paid-tier/overview
  - Pricing of above





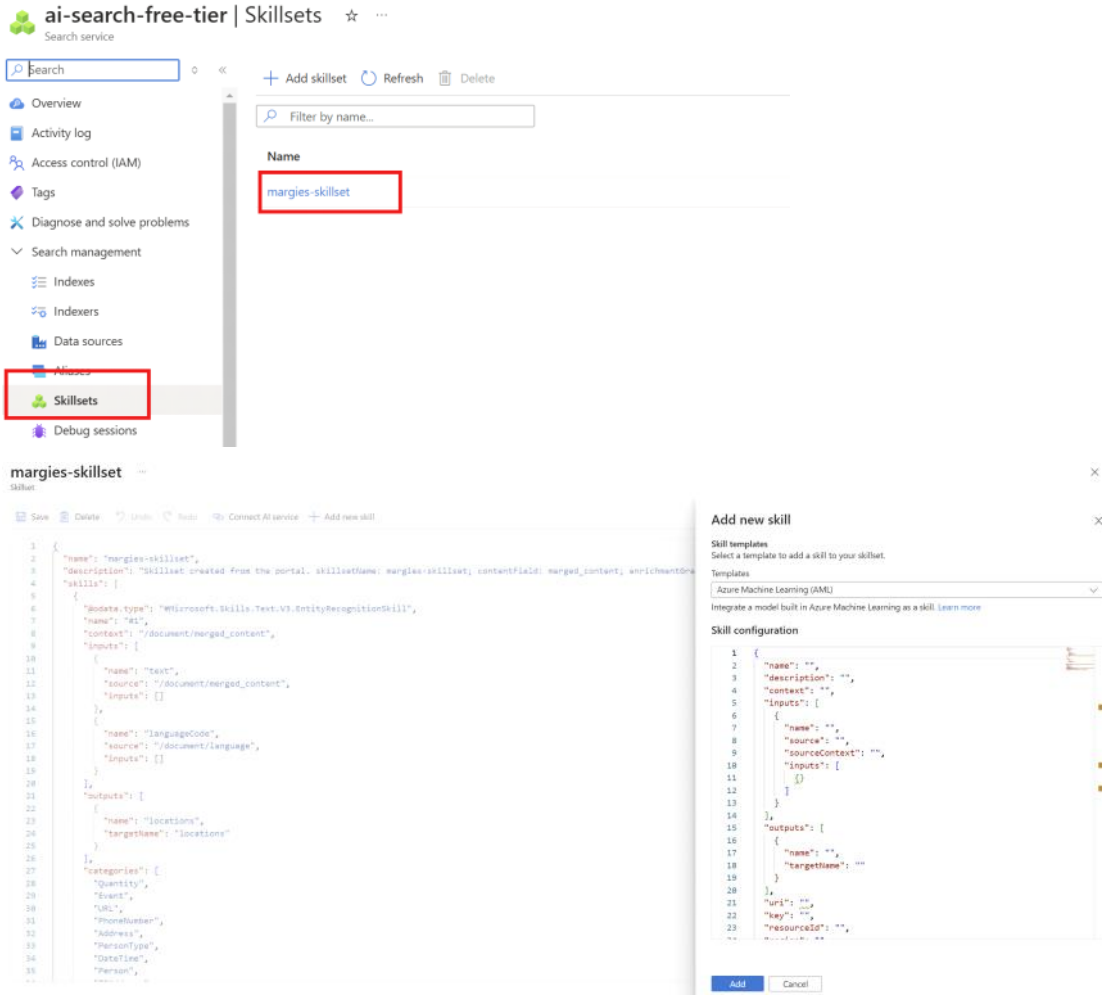
Modifying definitions using REST

- Everything is defined as JSON
- AI Search URL

- 

- Skillset is tied to cognitive services - AI service

Add a new skill

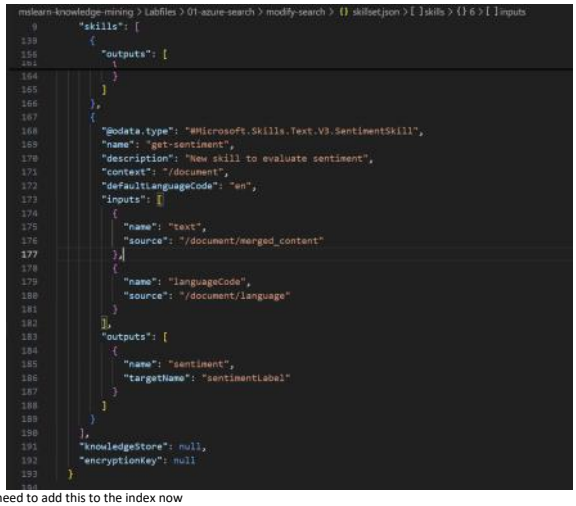
- 

We can do the same using REST (reusable)

- Create a skillset.json copying the existing one
- Add a new key-value pair for attaching cognitive services
- The key used here is of AI service and NOT AI Search key

```
{
    "name": "margies-skillset",
    "description": "Skillset for document extraction",
    "cognitiveServices": {
        "@odata.type": "#Microsoft.Azure.Search.CognitiveServicesByKey",
        "description": "Azure AI services",
        "key": "EOdV5nTHraQnFzycMpwrmNSw4Iuy6pQ45baQqgyWVPDE8UiG2q73JQQJ99AKAC1i4TkXJ3w3AAAEACOGHQR3"
    },
    "skills" : [
      {
        "@odata.type": "#Microsoft.Skills.Text.V3.EntityRecognitionSkill",
        "name":"#1",
        "description": null,
        "context": "/document/merged_content",
        "categories": [
          "Person",
          "Quantity",
          "Organization",
          "URL",
          "Email",
          "Location",
```

- Add a new skillset

- 

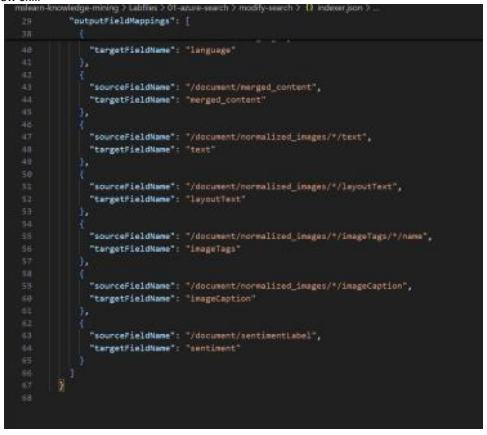
- We need to add this to the index now

Modify the Index

- We can get the JSON from portal and modify on top of it
- Added 2, one for new skillset and one for unencoded URL; indexer will define the mappings
  - 

Modifying the Indexer

- Get the indexer json from portal
- In indexer
  - fieldMappings - maps fields extracted from document content and metadata
  - outputFieldMappings - values extracted by skills in the skillset
- By default `metadata_storage_path` has the URL but encoded
- Save URL in non encoded format from the above field before it gets encoded
  - Encoding is good for efficient indexing
  - 

- For new skill
  - 

Run REST calls

mslearn-knowledge-mining > Labfiles > 01-azure-search > modify-search > modify-search.cmd

```
@echo off

rem Set values for your Search service
set url=https://ai-search-free-tier.search.windows.net
set admin_key=trPviONdFsaiEaPSygBwLuUr2SK1fnPCITjSA0r2aaAzSeAYcnlb

echo -----
echo Updating the skillset...
call curl -X PUT %url%/skillsets/margies-skillset?api-version=2020-06-30 -H "Content-Type: application/json" -H "api-key: %admin_key%" -d @skillset.json

rem wait
timeout /t 2 /nobreak

echo -----
echo Updating the index...
call curl -X PUT %url%/indexes/margies-index?api-version=2020-06-30 -H "Content-Type: application/json" -H "api-key: %admin_key%" -d @index.json

rem wait
timeout /t 2 /nobreak

echo -----
echo Updating the indexer...
call curl -X PUT %url%/indexers/margies-indexer?api-version=2020-06-30 -H "Content-Type: application/json" -H "api-key: %admin_key%" -d @indexer.json

echo -----
echo Resetting the indexer
call curl -X POST %url%/indexers/margies-indexer/reset?api-version=2020-06-30 -H "Content-Type: application/json" -H "Content-Length: 0" -H "api-key: %admin_key%"

rem wait
timeout /t 5 /nobreak

echo -----
echo Rerunning the indexer
call curl -X POST %url%/indexers/margies-indexer/run?api-version=2020-06-30 -H "Content-Type: application/json" -H "Content-Length: 0" -H "api-key: %admin_key%"
```

Home >

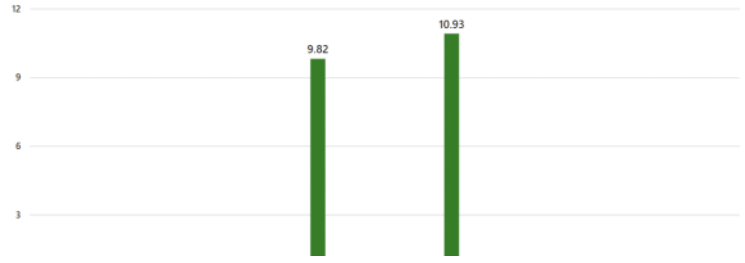
margies-indexer

Indexer

Run | Reset | Save | Refresh | Edit JSON | Delete

Execution history | Settings

Number of recent runs to show: 10



Status	Last run	Duration	Docs succeeded	Errors or warnings
Success	11/24/2024, 9:4...	9 s	72	0/0
Reset	11/24/2024, 9:3...	100 ms	0	0/0
Success	11/24/2024, 9:0...	118 ms	0	0/0
Success	11/24/2024, 9:0...	10 s	72	0/0

New Fields

Search explorer

ai-search-free-tier

Index: margies-index

Query options | View

* | Search

Results

```
      "cloud",
      "ship",
      "water",
      "building",
      "lake",
      "vehicle",
      "large",
      "boat",
      "architecture"
    ],
    "imageCaption": [
      "{\"tags\":[\"building\",\"outdoor\",\"sky\",\"tower\",\"tall\",\"city\",\"night\",\"skyscraper\",\"roof
      "{\"tags\":[\"sky\",\"boat\",\"outdoor\",\"water\",\"ship\",\"docked\"],\"captions\":[{\"text\":\"a larg
    ],
    "sentiment": "neutral",
    "url": "https://rg13286088654.blob.core.windows.net/margies/collateral/Dubai%20Brochure.pdf"
  },
  {
    "@search.score": 1,
    "content": "\nMargie's Travel Presents… \n\nLas Vegas \n Las Vegas, officially \n\nthe City of Las Vegas \r
    "metadata_storage_size": 569917,
```

Using Azure Search SDK to query/search data

- For python we need package `azure-search-documents`

```python
# Wrapper function for request to search index
def search_query(search_text, filter_by=None, sort_order=None):
    try:

        # Create a search client
        azure_credential = AzureKeyCredential(search_key)
        search_client = SearchClient(search_endpoint, search_index, azure_credential)


        # Submit search query
        results =  search_client.search(search_text,
                                        search_mode="all",
                                        include_total_count=True,
                                        filter=filter_by,
                                        order_by=sort_order,
                                        facets=['metadata_author'],
                                        highlight_fields='merged_content-3,imageCaption-3',
                                        select = "url,metadata_storage_name,metadata_author,metadata_storage_size,metadata_storage_last_modified,language,sentiment,m
        return results



    except Exception as ex:
        raise ex
```