

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+AST880W1g==

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

Home >

ai-search-free-tier

Search service

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Search management

Indexes

Indexers

+

Add index

Import data

Import and vectorize data

Se

Query rewriting is now available in semantic ranker for select regions. [Learn](#)

Essentials

Resource group (move) : rg-1

Location (move) : Central US

Subscription (move) : Azure subscription 1

Subscription ID : a38d9f6e-bf2f-4768-9711-ba4815a96fcb

Status : Running

Tags (edit) : Add tags

Home > ai-search-free-tier > Import data >

Import data

Connect to your data

Add cognitive skills (Optional)

Customize target index

Create an indexer

Create and load a search index using data from an external data source. Azure AI Search crawls the data structure you provide, extracts searchable i

Data Source

Azure Blob Storage

Data source name *

margies-data

Data to extract

Content and metadata

Parsing mode

Default

Subscription

Azure subscription 1

Connection string *

DefaultEndpointsProtocol=https;AccountName=rg1328608865-

Choose an existing connection

Managed identity authentication

None

System-assigned

User-assigned

Container name *

margies

Blob folder

your/folder/here

Description

(optional)

Next: Add cognitive skills (Optional)

Import data

Connect to your data

Add cognitive skills (Optional)

Customize target index

Create an indexer

Enrich and extract structure from your documents through cognitive skills using the same AI algorithms that power AI Services. Select the document cracking options and the cognitive skills you want to apply to your documents. Optionally, save scenarios other than search. [Learn more](#)

Attach AI Services

To power your cognitive skills, select an existing AI Services resource or create a new one. The AI Services resource should be in the same region as your search service. If you decide to include cognitive skills in Azure AI Search, choosing to add these skills, you acknowledge that you are aware of the [Pay-as-you-go pricing](#) for each skill and the [documentation](#) that explains how each skill functions. [Learn more](#)

AI Services Resource Name

Free (Limited enrichments)

Refresh

Create new AI service

REGION in the table below specifies the region only for included regional services. This does not specify a region for included non-regional services (e.g. Translator Text service). See [service status page](#) for more details.

Import data

Skillset name *

margies-skillset

☒ Enable OCR and merge all text into merged_content field

Source data field *

merged_content

Enrichment granularity level

Source field (default)

☐ Enable incremental enrichment

Checked items below require a field name.

☒ Text Cognitive Skills

Parameter

Field name

☐ Extract people names

people

☐ Extract organization names

organizations

☒ Extract location names

locations

☒ Extract key phrases

keyphrases

☒ Detect language

language

☐ Translate text

Target Language English

translated_text

☐ Extract personally identifiable information

pii_entities

☒ Image Cognitive Skills

Field name

☒ Generate tags from images

imageTags

☒ Generate captions from images

imageCaption

Import data

Index name *

margies-index

Key *

metadata_storage_path

Suggester name

Search mode

analyzingInfixMatching

+ Add field

+ Add subfield

Configure vector field

Delete

Field name	Type	Retrievable	Filterable	Sortable	Facetable	Searchable	Analyzer	Suggester
content	Edm.String	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
metadata_storage_content_type	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_storage_size	Edm.Int64	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_storage_last_modified	Edm.DateTime...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_storage_content_md5	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_storage_name	Edm.String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
metadata_storage_path	Edm.String	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_storage_file_extension	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_content_type	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_language	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_author	Edm.String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
metadata_title	Edm.String	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
metadata_creation_date	Edm.DateTime...	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		...
locations	CollectionE...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
keyphrases	CollectionE...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
language	Edm.String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
merged_content	Edm.String	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
text	CollectionE...	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
layoutText	CollectionE...	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
imageTags	CollectionE...	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...
imageCaption	CollectionE...	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Standard - Luce...	...

Previous: Add cognitive skills (Optional)

Next: Create an indexer

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

margies-indexer

Execution

Download as CSV

Execution result

Status

Timed out

Error message

Indexing was stopped because the free skillset execution quota has been reached. When no cognitive services subscription is attached to a skillset, only 20 documents can be enriched at a time. To index more documents, attach a cognitive services subscription to your skillset. See https://aka.ms/cog-arch-attach for more information.

Datasource

margies-data

Target index

margies-index

Start

11/24/2024, 8:29:21 PM

End

11/24/2024, 8:29:29 PM

Docs succeeded

20/20

Errors or warnings

0/0

Search by Document Key, Operation, Messa...

filter by severity : All filter by key : All

Message

- 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->

Trying the above with a Paid tier AI Search Service

- Estimated cost per month = \$250

Create a search service

BasicsScaleNetworkingTagsReview + create

Project details

Subscription *

Azure subscription 1

Resource Group *

rg-1

Create new

Instance Details

Service name * ⓘ

ai-search-paid-tier

Location *

Central US

Pricing tier * ⓘ

Standard

160 GB/Partition, max 12 replicas, max 12 partitions, max 36 search units

Change Pricing Tier

- 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

Connect to your dataAdd cognitive skills (Optional)Customize target indexCreate an indexer

⚠️ Enrich and extract structure from your documents through cognitive skills using the same AI algorithms that power AI Services. Select the document cracking options and the cognitive skills you want to apply other than search. [Learn more](#)

Attach AI Services

To power your cognitive skills, select an existing AI Services resource or create a new one. The AI Services resource should be in the same region as your search service. If you decide to include these skills, you acknowledge that you are aware of the [Pay-as-you-go pricing](#) for each skill and the [documentation that explains how each skill functions](#). [Learn more](#)

AI Services Resource Name

Free (Limited enrichments)

Refresh

Create new AI service

Region

Trying with a new paid AI Service Resource

- From here - [Create new AI service](#)

Import data

Connect to your dataAdd cognitive skills (Optional)Customize target indexCreate an indexer

⚠️ Enrich and extract structure from your documents through cognitive skills using the same AI algorithms that power AI Services. Select the document cracking options and the cognitive skills you want to apply other than search. [Learn more](#)

Attach AI Services

To power your cognitive skills, select an existing AI Services resource or create a new one. The AI Services resource should be in the same region as your search service. If you decide to include these skills, you acknowledge that you are aware of the [Pay-as-you-go pricing](#) for each skill and the [documentation that explains how each skill functions](#). [Learn more](#)

AI Services Resource Name

Free (Limited enrichments)

ai-service-paid-tier

Refresh

Create new AI service

Region

centralus

- This time the indexer with skillset succeeded

ai-search-free-tier | Indexers

Search service

Search

+ Add indexer Refresh Delete

Filter by name...

Status	Name	Last run	Docs succeeded	Errors/Warnings
Success	margies-indexer	9 seconds ago	0/0	0/0
Success	margies-indexer-without-skillset	27 minutes ago	72/72	0/0

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Search management

Can see the enrichment in search explorer

ai-search-free-tier

margies-index

View

Search

```

12 "content": "UNARGIS Travel Presents. UNDubai is the largest and most populous cit
13 "metadata_storage_size": 388622,
14 "metadata_storage_last_modified": "2024-11-25T01:08:42Z",
15 "metadata_storage_name": "Dubai Brochure.pdf",
16 "metadata_storage_path": "aHR0cDp6YyZzE7Pjg2HDg4UjU0UWU5b2ZuY29yZS53aW5kb3dzLW51dC9tYXJnaWw6L2NhbG6hdG
17 "metadata_author": "Margies Travel",
18 "locations": [
19     "Dubai",
20     "United",
21     "Arab Emirates",
22     "Persian Gulf",
23     "Emirate of Dubai",
24     "Abu",
25     "Dhabi",
26     "emirates",
27     "emirate",
28     "Dubai-Sharjah-Ajman",
29     "Creek Hotel",
30     "boutique hotel",
31     "Dubai Creek",
32     "Deira Hotel",
33     "hotel",
34     "Lost City Hotel",
35     "waterpark",
36     "aquarium",
37     "Burj Khalifa"
38 ],
39 "keyphrases": [
40     "The Deira Hotel Family-run hotel",
41     "The Lost City Hotel",
42     "The Creek Hotel",
43     "Friendly boutique hotel",
44     "The Burj Khalifa",
45     "Federal Supreme Council",
46     "Dubai-Sharjah-Ajman metropolitan area",

```

- 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

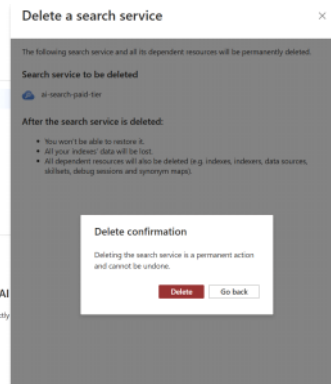
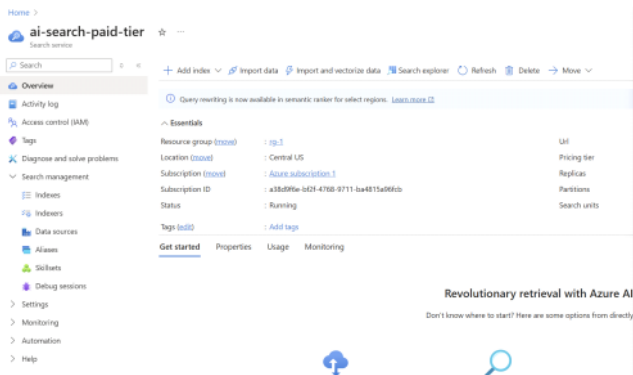
- Pricing of above



① The con-

Page 10 of 10

Tiers / Features	SO Standard
Pricing	Starting \$0.40 USD/1K calls (Estimated)
Vision	✓
Language	✓
Search	✓
Speech	✓



- Everything is defined as JSON

- AI Search URL

- Skillset is tied to cognitive services - AI service

Add a new skill

margies-skillset

```
1 {
2   "name": "margies-skillset",
3   "description": "Skillset created from the portal. skillsetName: margies-skillset; contentField: merged_content; enricherName: ",
4   "skills": [
5     {
6       "@odata.type": "Microsoft.Skills.Text.V3.EntityRecognitionSkill",
7       "name": "#1",
8       "context": "/document/merged_content",
9       "inputs": {
10        "name": "Text",
11        "source": "/document/merged_content",
12        "inputs": {}
13      },
14      "outputs": {
15        "name": "LanguageCode",
16        "source": "/document/language",
17        "inputs": {}
18      }
19    },
20    {
21      "name": "Locations",
22      "targetName": "Locations"
23    }
24  ],
25  "categories": [
26    "Quantity",
27    "Event",
28    "URL",
29    "PhoneNumber",
30    "Address",
31    "PersonType",
32    "Date/Time",
33    "Person"
34  ]
35 }
```

Add new skill

Skill templates
Select a template to add a skill to your skillset.

Templates

Azure Machine Learning (AML)
Integrate a model built in Azure Machine Learning as a skill. [Learn more](#)

Skill configuration

```
1 {
2   "name": "",
3   "description": "",
4   "context": "",
5   "inputs": {
6     "name": "",
7     "source": "",
8     "sourceContext": "",
9     "inputs": {
10      "name": "",
11      "targetName": ""
12    }
13   },
14   "outputs": {
15     "name": "",
16     "targetName": ""
17   }
18 },
19 {
20   "name": "",
21   "uri": "",
22   "key": "",
23   "resourceId": ""
24 }
```

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": "E0dV5nTHraQnFzycMpwrmNSw4Iuy6pQ45baQqgyWPDE8U1G2q73JQQJ99AKAC114TKXJ3w3AAAEACOGHQ3"
  },
  "skills": [
    {
      "@odata.type": "Microsoft.Skills.Text.V3.EntityRecognitionSkill",
      "name": "#1",
      "description": null,
      "context": "/document/merged_content",
      "categories": [
        "Person",
        "Quantity",
        "Organization",
        "URL",
        "Email",
        "Location",
        "Date/Time",
        "Person"
      ]
    }
  ]
}
```

- Add a new skillset

```

mslearn-knowledge-mining > Labfiles > 01-azure-search > modify-search > {} skillset.json > {} skills > {} 6 > {} inputs
159 {
160   "skills": [
161     {
162       "outputs": [
163         {
164           {
165             }
166           },
167         {
168           "metadata.type": "Microsoft.Skills.Text.V3.SentimentSkill",
169           "name": "get-sentiment",
170           "description": "New skill to evaluate sentiment",
171           "context": "/document",
172           "defaultLanguageCode": "en",
173           "inputs": [
174             {
175               "name": "text",
176               "source": "/document/merged_content"
177             },
178             {
179               "name": "languageCode",
180               "source": "/document/language"
181             }
182           ],
183           "outputs": [
184             {
185               "name": "sentiment",
186               "targetName": "sentimentLabel"
187             }
188           ]
189         },
190       ],
191       "knowledgeStore": null,
192       "encryptionKey": null
193     }
194   ]
195 }

```

- 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

```

{
  "name": "sentiment",
  "type": "Edm.String",
  "facetable": false,
  "filterable": true,
  "retrievable": true,
  "sortable": true
},
{
  "name": "url",
  "type": "Edm.String",
  "facetable": false,
  "filterable": true,
  "retrievable": true,
  "searchable": false,
  "sortable": false
}
]
}

```

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

```

{
  "fieldMappings": [
    {
      "sourceFieldName": "metadata_storage_path",
      "targetFieldName": "metadata_storage_path",
      "mappingFunction": {
        "name": "base64Encode"
      }
    },
    {
      "sourceFieldName": "metadata_storage_path",
      "targetFieldName": "url"
    }
  ]
}

```

- For new skill

```

mslearn-knowledge-mining > Labfiles > 01-azure-search > modify-search > {} indexer.json > {}
29 {
30   "outputFieldMappings": [
31     {
32       "targetFieldName": "language"
33     },
34     {
35       "sourceFieldName": "/document/merged_content",
36       "targetFieldName": "merged_content"
37     },
38     {
39       "sourceFieldName": "/document/normalized_images/*text",
40       "targetFieldName": "text"
41     },
42     {
43       "sourceFieldName": "/document/normalized_images/*layoutText",
44       "targetFieldName": "layoutText"
45     },
46     {
47       "sourceFieldName": "/document/normalized_images/*imageTags/*name",
48       "targetFieldName": "imageTags"
49     },
50     {
51       "sourceFieldName": "/document/normalized_images/*imageCaption",
52       "targetFieldName": "imageCaption"
53     },
54     {
55       "sourceFieldName": "/document/sentimentLabel",
56       "targetFieldName": "sentiment"
57     }
58   ]
59 }

```

Run REST calls

```
mslearn-knowledge-mining > Labfiles > 01-azure-search > modify-search > modify-search.cmd
1 @echo off
2
3 rem Set values for your Search service
4 set url=https://ai-search-free-tier.search.windows.net
5 set admin_key=trPvIONdFsaIEaPSyg9wLUr2SKIrnPCITjSA0r2aaAzSeAYcn1b
6
7 echo -----
8 echo Updating the skillset...
9 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
10
11 rem wait
12 timeout /t 2 /nobreak
13
14 echo -----
15 echo Updating the index...
16 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
17
18 rem wait
19 timeout /t 2 /nobreak
20
21 echo -----
22 echo Updating the indexer...
23 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
24
25 echo -----
26 echo Resetting the indexer
27 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%"
28
29 rem wait
30 timeout /t 5 /nobreak
31
32 echo -----
33 echo Rerunning the indexer
34 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

12

9

6

3

9.82

10.93

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

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

"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 large

],

"sentiment": "neutral",

"url": "https://rg13286088654.blob.core.windows.net/margies/collateral/Dubai%208rochure.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'

```

# 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

```