TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTRONIKY A INFORMATIKY

Klaudové služby pre získavanie informácii

2021 Michal Stromko

Úvod

Cieľom mojej práce bolo zistenie fungovania klaudových služieb pre umelú inteligenciu a zistenie fungovania webových vyhľadávačov. V mojej práci som sa hľavne zameral na fungovanie webových vyhľadávačov.

V dnešnej dobe vo väčšine webových stránok je vytvorený vyhľadávač obsahu na stránke. Je to impelementované, napríklad pre lepšie vyhľadávania informácií na stránke. Vzniklo to kvôli tomu, pretože veľké množstvo webových stránok na internete má obrovské množstovo informácií.Vyhľadávanie jednej informácie by mohlo tvrať aj desať minút. Takémuto zdĺhavému hľadaniu informácie sa predišlo vytvorením vyhľadávacieho okna na stránke.

Boli by sme si pomysleli, že vytvorenie takéhoto vyhľadávača je jednoduché. Žiaľ, to nie je pravda. Za vytvorením takéhoto vyhľadávača môžeme nájsť množnstovo strávených hodín programovania. Treba si aj uvedomiť to, že takéto vyhľadávače fungujú na umelej inteligencii. Umelá inteligencia dokáže rozoznať minimálne jednu informáciu, napr. do vyhľadávača zadáme *"Mobilný telefón"*. Niekdedy bolo vyhľadanie jednej informácie bežné. Technológie v dnešnej dobe postupujú obrovskou rýchlosťou a stáva sa štandardov vahľadávania napr. výraz: *Koľko stojí telefón Xiaomi Mi 11 ?*. Vyhľadávač na stránke nám dokáže odpovedať na takúto otázku a zároveň nám aj ponúkne vložiť tovar do košíka.

V mojej práci som sa pokúsil o vytvorenie jednoduchého vyhľadávača na stránke ZP Wiki.

Ciele práce

Mojou hlavou úlohou pri riešení tejto práce bolo, porozumieť fungovaniu vyhľadávania informácií na stránkach. Popri študovaní ako to funguje som sa pokúsil v prostredí **Microsoft Azure** vytvoriť vyhľadávanie pre stránku **ZP Wiki**.

Vytvorenie vyhľadávanie zahŕňa:

- vytvorenie nasledujúcich aplikácii:
    - Azure Corgnitive Search
    - Dakabázu, napr.:
        - SQL databases
        - Azure Blob storage
    - Zdroja informácii (Resouce group)
- vytvorenie indexu pre skránku
- vytvorenie kontajnera pre ZP Wiki
- Vytvorenie tutoriálov pre lepšie vytváranie vyhľadávania

Fungovanie vyhľadávania na stránke

Po zadaní do textového poľa pre vyhľadávanie, nasleduje na stránke množstvo oprérácii. Pre správne vyhľadávanie je dôležité, aby vyhľadávač bol schopný v reálnom čase prehľadať stránku a vytvoriť si dátovú štruktúru inak povedané *index*. Vyhľadávače, ktoré majú názov **fulltext** pri vyhľadávaní používajú kľúčové slová, ktoré vyhľadajú v indexe.

Princíp vyhľadávania

Principiálne vyhľadávače používajú len prvé tri kroky:

1. Crawlovanie
    - pojem, ktorý zahŕňa vyhľadanie alebo zber informácii (dát), ktoré sa uložia do databázy
2. Indexácia
3. Výsledky vyhľadávania
4. **Crawler**

Crawler

Je to jeden z najdôležitejších nástrojov pre prechádzanie súborov webových stránok. Je označovaný za program, ktorý si ukladá dáta, napr. obsah stánok, metadáta (sú to informácie o danej stránke, ako príklad hashe dokumentu, dátumy stiahnutia dokumentu a podobne.) Primárnou úlohou Crawlera je ukladanie *Hypertextových odkazov*, ktoré sa nachádzajú na stránkach. Pred uložením takéto odkazu ho otvorí a vyhľadá ďalšie informácie spolu s ďalšími odkazmi. Robí to preto, aby získal čo najviac pravdivých informácii. Firmy ako Microsoft, Google, Apple a ďalšie majú svoje stránky uložené na minimálne tisickach GB, keby sme pustili crawler na takéto stránky, tak by spotreboval obrovské množstvo úložiska na uloženie dát o stránkach. Netreba zabúdať ani na fakt, že vyhľadávanie informácii v takejto databáze by trvalo príliš dlho, možno v desiatkach minút. Keby nastala takáto situácia, úžívateľ prestane používať danú stránku. Preto sa do tohto nástroja zadefinovalo overovanie informácii spolu s vyhodnocovaním či dané dáta majú byť zapísané do databázy. Veľakrát nastáva aj situácia, že veľké množstov stránok využíva rovnakú cestu do súboru. Vtedy sa takáto duplikovaná cesta uloží do pamäte iba raz a druhá adresa dostane len informáciu, kde sa nachádza zvyšok cesty do súboru.

Vytváranie indexu

Pri vytváraní indexu sa do pamäte zapisujú len najdôležitejšie informácie, ktoré následne slúžia pre rozhodovanie, ktoré stránky budú užívateľovi zobrazené na obrazovke. Takéto informácie sa triedia podľa relevantnosti.

Všeobecne sú to napr. tieto typy dát:

- typ stránky
- jazyk stránky
- informácie o doméne (napr. či táto stránka je bezpečná)
- spätné odkazy
- holý text (obsahuje slová ktoré sú uložené)

Aktualizácia indexu

Je dôležitá pre správne fungovanie vyhľadávania, aby boli v databáze uložené aktuálne informácie na danej stránke. Poznáme 2 typy aktualizácii:

1. Prírastková aktualizácia
    - pri aktualizácii indexu sa nové dáta z databázy vyhľadávača pridajú do súčastného indexu. Vzniká tým len problém toho, že je potrebné dáta zoradiť na správne miesto v indexe.
2. Hromadná aktualizácia
    - pri tejto metóde sa kontroluje, ktorá nová stránka pribudla v databáze.
    - z takýchto dát sa vytvorí nový index, ktorý bude mať menej parametrov.
    - k spojenu dvoch indexov teda nového a starého dochádza až počas samostatného vyhľadávania

Microsoft Azure

Pri vypracovávaní môjho zadania som pracoval v prostredí Microsft aplikácie. Táto aplíkácia funguje vo webovom rozhraní. V tejto aplikácii sa dajú vytvárať SQL databázy, Maria DB, Posgre SQL databázy, virtuálne stroje a mnoho dalších aplikácii. Veľkú časť tutoriálov som vytvoril na prácu vo webovom rozhraní. Existuje aj pripojenie na takúto aplikáciu pomocou terminálu v OS založenom na UNIX, aj túto metódu som využíval.

Na prihlásenie do tohto portálu som vytvoril tutoriál, ktorý nájdete na nasledujúcom odkaze. Turoriál na vytvorenie konta na azure portály.

Vytvorenie SQL databázy

Základom vytvorenia vyhľadávania je vytvorená databáza, ktorá bude udržiavať informácie o indexe. Takúto databázu je možné vytvoriť dvoma spôsobmi. Prvá možnosť vytvorenia databázy je priamo na portály **Microsoft Azure**. Druhá možnosť je pomocou terminálu.

Vytvorenie databázy priamo na portály Microsoft Azure

Vytvorenie takejto databázy nie je moc náročné. Stačí mať len zbehlosť v správnom vypĺňaní formulárov. Pri vypracovávaní projektu som využil aj takýto spôsob vytvorenia databázy. Pre jednoduchšie vytvorenie takejto databázy som vytvoril tutoriál. Tento tutoriál nájdete na nasledujúcom odkaze.

Návod na vytvorenie databázy

Vytvorenie databázy pomocou terminálu

Pre vytvorenie databázy pomocou terminálu je potrebná inštalácia programu, ktorý bude fungovať v terminály. Tento program má názov **Azure CLI**, pripájam link na nainštalovanie datého programu, a následné prihlásenie užívateľa pomocou príkazu do potrálu **Microsoft Azure**

Inštalácia Azure CLI

Po nainštalovaní a prihásení som si vytvoril skript s príponou *sh*, do ktorého som napísal príkazy pre vytvorenie databázy. Tento súbor nájdete v prílohe s názvom **sql_database.sh**. Následne som tento skript spustil pomocou príkazu **sh sql_database.sh**

Po vykonaní tohto príkazu sa mi ako výstup príkazu zobrazily informácie o vytvorení databázy. Tieto informácie sú napr. Názov sql databázy, adresa servera, kde je databáza spustená, meno užívateľa, ktorý sa môže do nej prihlásiť a ďalšie informácie. Tieto infromácie som si zapísal do súboru **sql_database_out.txt**, ktorý je súčaťou prílohy.

Vytvorenie tabuľky v databáze a pridanie hodnôt do tabuľky

Pre vytvorenie tabuľky v databáze som si nainštaloval program **Azure Data Studio**, stránku na stiahnutie a naištalovanie programu nájdete na nasledujúcom linku.

Inštalácia Azure Data Studio

Po následnej inštalácii sa program spustil a prihlásil som sa do databázy. Neskôr som si vytvoril tabuľku nasledovným príkazom. Tento príkaz nájdete v prílohe **create_table.sql**

```
    CREATE TABLE students
(
    StudentId   INT             NOT NULL    PRIMARY KEY,
    Name        [NVARCHAR](50)  NOT NULL,
    Surname     [NVARCHAR](50)  NOT NULL,
    Email       [NVARCHAR](50)  NOT NULL,
    StartStudy  INT             NOT NULL,
    SubjectName [NVARCHAR](15)  NOT NULL
)
```

Následne som do tejto tabuľky pridal 15 záznamoch o študentoch. Tento skript nájdete v prílohe pod názvom **Insert_table.sql**

Neskôr som už len vytvoril select, ktorým som si vyskúšal či sa dané dáta nachádzajú v tabuľke.

```
SELECT * FROM studenti;
```

Vytvorenie Azure Cognitive Search

Azure Cognitive search je kladudová vyhľadávacia služba, ktorá poskytuje vývojárom API (Aplication Programming Interface) nástroj na jednoduché vytvorenie vyhľadávania na stránke. Rozhranie API a architektúra kognitívneho vyhľadávania zjednodušuje úlohu pri pridávaní sofistikovaného vyhľadávania informácii.

Vytvorenie ACS som realizoval pomocou portálu **Microsoft Azure**. Vypĺňanie formulárov je celkom jednoduché, ale má jednu chybu. Ide o chybu spojenú s firmou Microsoft, ktorá pri študenských vytvára preddefinovanú databázu hotelov. Firme ide o to, aby používateľ pri vytvorení databázy dokázal aj vyhľadávať údaje v danej databáze. Na vytvorenie ACS som vytvoril jednoduchý tutoriál, ktorý nájdete v nasledujúcom linku.

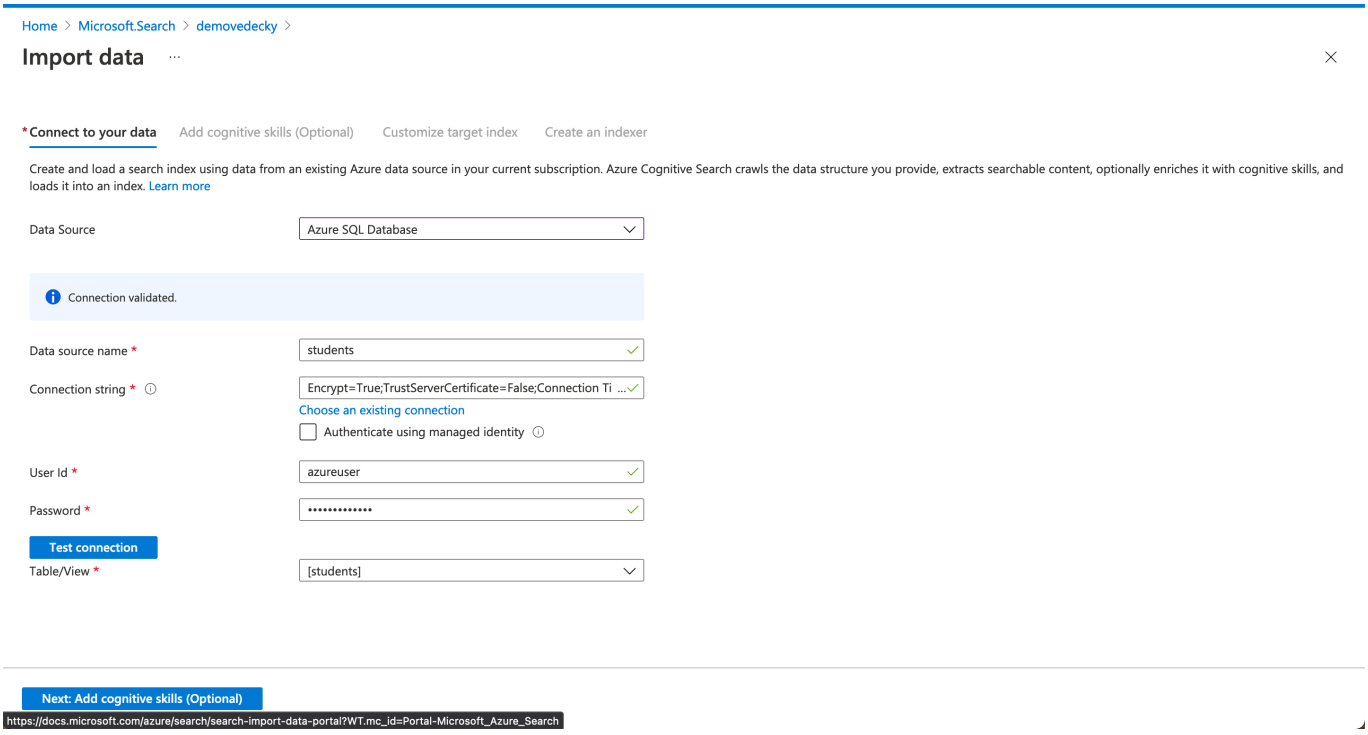
Vytvorenie ACS

Vytorennie indexu v ACS

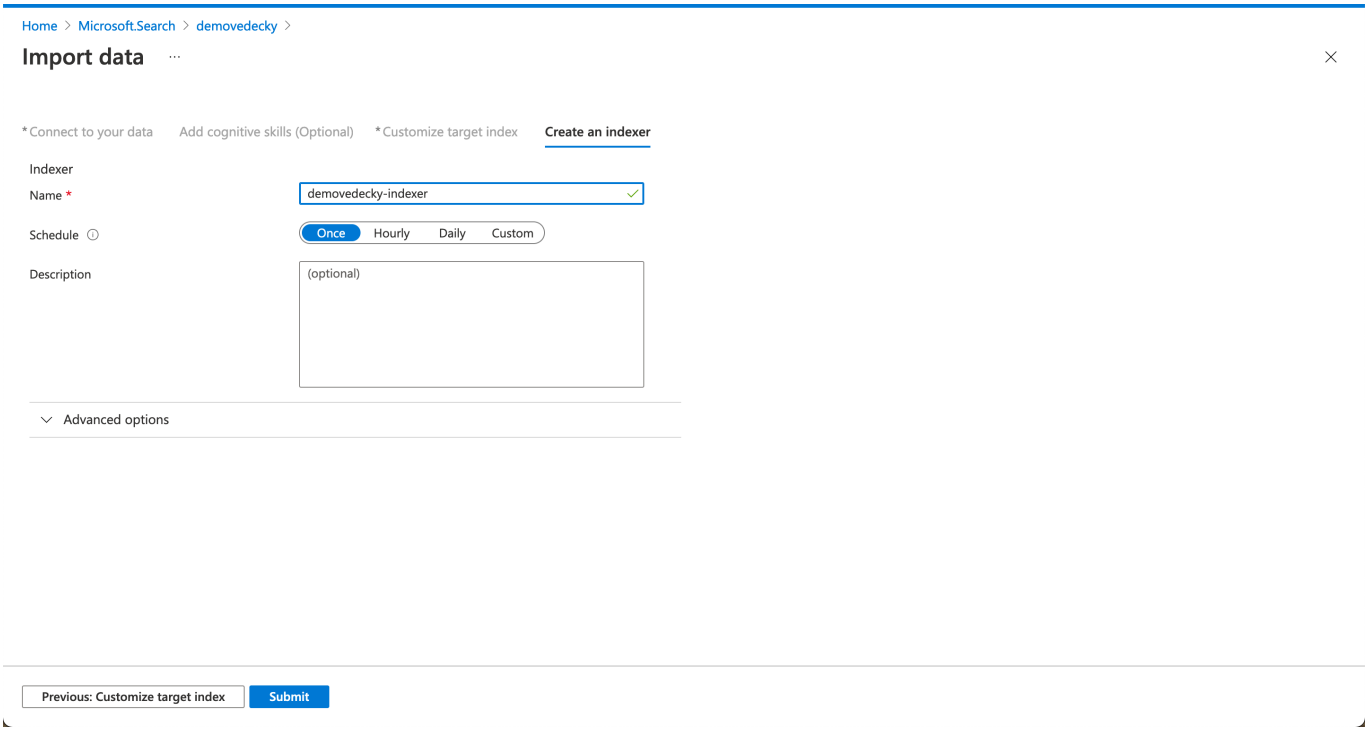
Index tvorí základnú časť pre vyhľadávanie v ACS. Pri vytváraní indexu, je potrebné mať vytvorenú databázu, do ktorej sa index zapíše. Pokiaľ nemáte takto vytvorenú databázu, tak dokážete vytvoriť index iba na hotely, ktoré sú predefinované pri vytvorení. Vyhľadávanie v indexe a aj práca s ním, napr. napísanie skriptu vyžaduje znalosť používania nástroja *JSON* Vytvorenie indexu som realizoval na portáli **Microsoft Azure** a vytvoril som aj k tomu tutoriál, ktorý nájdete po kliknutí na nasledujúci link. Vytvorenie indexu

Pre vytvorenie indexu, ktorý je prepojený s databázou je potreba dobre vyplňiť formulár. Moje vyplenie fromulára môžete vidieť na nasledujúcich obrázkoch.

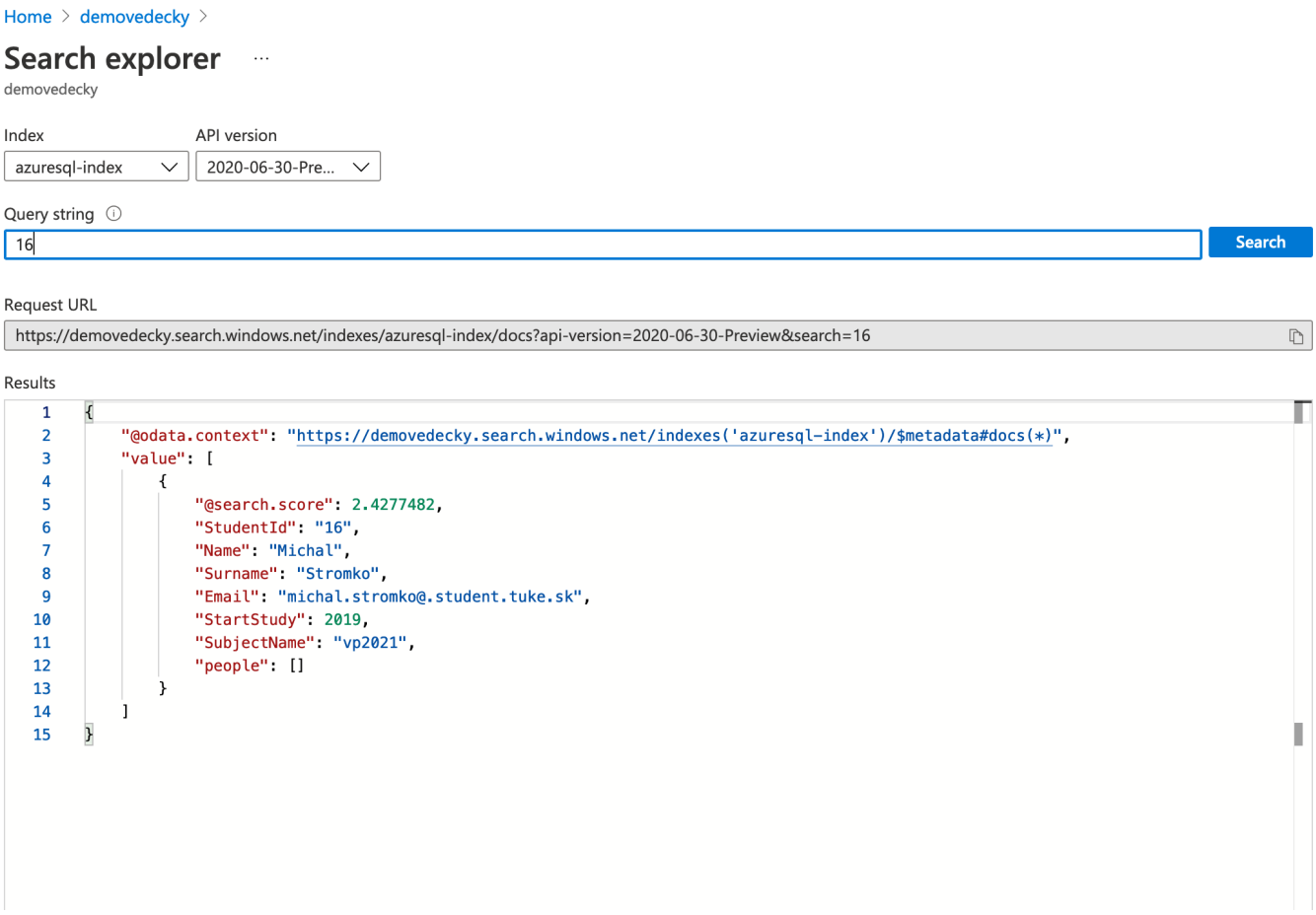
Obrázok na vytvorenie indexu, ktorý bude prepojený s databázou.



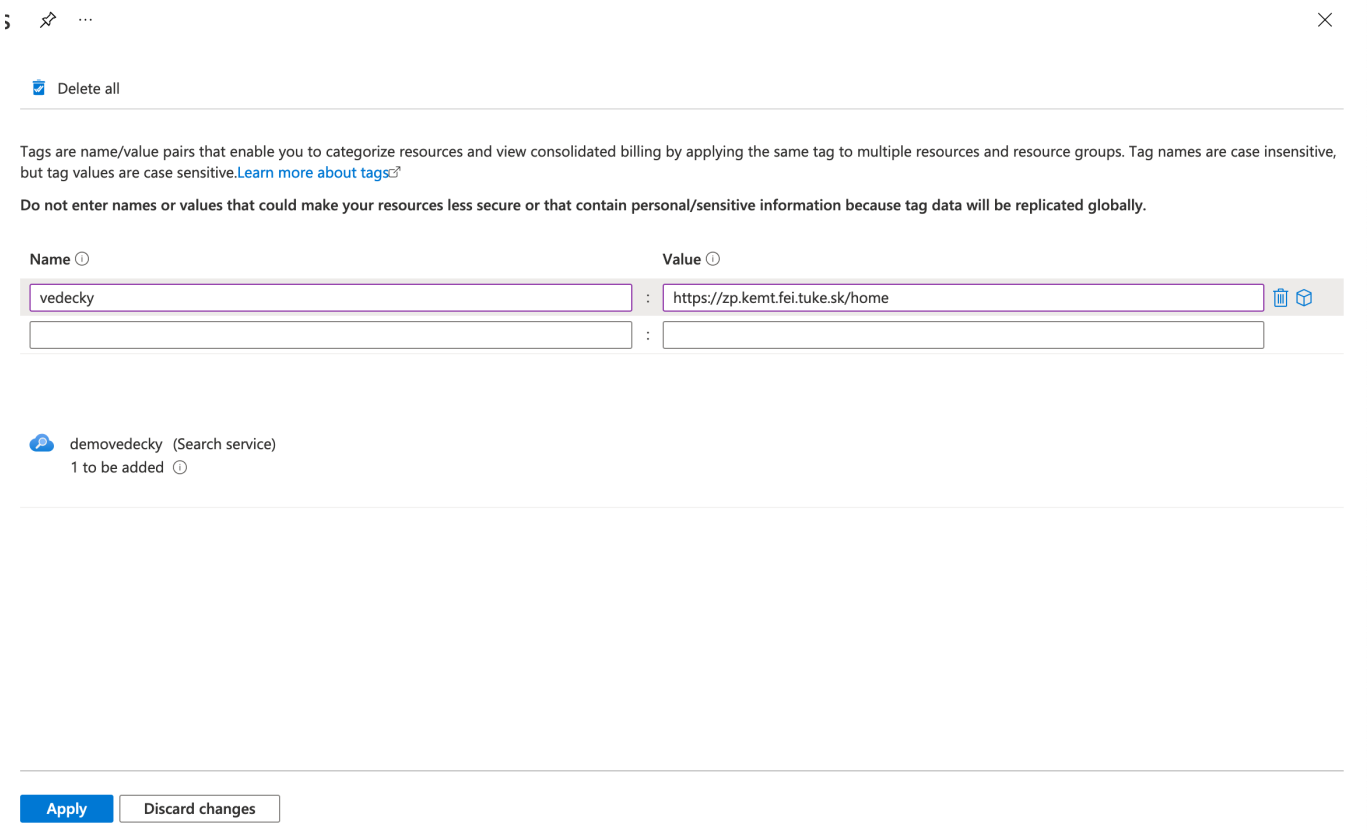
Nasleduje obrázok s vytvorením indexera.



Následne som sa pokúsil vyhľadať pomocou ACS pomocou mojeho ID informácie o mne v databáze. Dá sa to aj realizovať pomocou iných operácii. Dokážem si v databáze zobratiť všetkých študetov, ktorý napríklad začali štúpdium v roku 2018.



Ako poslednú úlohu v ACS som pridal spojenie stránky ZP Wiki s ASC. Vytvorenie takéhoto spojenia môžete vidieť na obrázku.



Záver

Pri vypracovávaní zadania na predmet Vedecký projekt som sa strtol s veľkým množstvom prekážok. Najväčšou prekážkou pre mňa bolo pochopenie fungovania Porálu **Microsft Azure**. Musel som pochopiť to, aký treba zvoliť postup pre vytvorenie takéhoto vyhľadávania. Veľa krát sa mi stala situácia že postup vytvorenia napríklad ASC sa menil každý týždeň. To znamená že pri takto rýchlej zmene vytvorenie ASC vznikali chyby. Takéto chyby som zaznamenal medzi návami aplikácii. Chcel som prepojiť Azure Blob s ACS a nefugovalo mi to zd dôvodu že Azure Blob musle mať v názve jedno písmenon veľké a ACS takýto názov s veľkým písmenon nepodporuje. Z toho dôvodu som zvolil postup vytvorenia SQL databázy, ktorá takýto problém nemá.

Momentálny stav vypracovania tohto projektu je v stave kedy mám vytvorenú funkčnú databázu s dátami o študentoch a dokážem pomocou ACS v tejto databáze vyhľadávať. Viem vyhľadať informácie o jednom študentovi, ale aj o viacerých naraz. Výhodou vyhľadávania na portály je to že stačí napísať reťazec znakov a systém automaticky napíše json skript, ktorý spustí.

Bol by som rád keby som mohol pokračovať vo vypracovávaní takéhoto zaujímavého projektu aj ďalej. Je tu množstvo vyľadovania a zisťovania informácii ako by sa dal implementovať ACS nástoroj alebo len jeho časť na stránku ZP Wiki. Určite sa pokúsim dotiahnuť tento projekt do úspešného konca.