

TECHNICKÁ UNIVERZITA V KOŠICIACH FAKULTA ELEKTRONIKY A INFORMATIKY

Klaudové služby pre získavanie informácií

Úvod

Cieľom mojej práce bolo zistenie fungovania kľúčových služieb pre umelú inteligenciu a zistenie fungovania webových vyhľadávačov. V mojej práci som sa hlavne zamerlal na fungovanie webových vyhľadávačov.

V dnešnej dobe vo väčšine webových stránok je vytvorený vyhľadávač obsahu na stránke. Je to implementované, napríklad pre lepšie vyhľadávania informácií na stránke. Vzniklo to kvôli tomu, pretože veľké množstvo webových stránok na internete má obrovské množstvo informácií. Vyhľadávanie jednej informácie by mohlo trvať 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 množstvo 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*". Niekedy bolo vyhľadanie jednej informácie bežné. Technológie v dnešnej dobe postupujú obrovskou rýchlosťou a stáva sa štandardom vyhľadávania napr. výraz: *Kolko 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 hlavnou 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 vyhľadávania zahŕňa:

- vytvorenie nasledujúcich aplikácií:
 - Azure Cognitive Search
 - Databázu, napr.:
 - SQL databases
 - Azure Blob storage
 - Zdroja informácií (Resource group)
- vytvorenie indexu pre stránku
- 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 pooužívajú 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 stránok, 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 takého odkazu ho otvorí a vyhľadá ďalšie informácie spolu s ďalšími odkazmi. Robí to preto, aby získal čo najviac pravdivých informácií. Firmy ako Microsoft, Google, Apple a ďalšie majú svoje stránky uložené na minimálne tisíckach 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 informácií v takejto databáze by trvalo príliš dlho, možno v desiatkach minút. Keby nastala takáto situácia, užívateľ prestane používať danú stránku. Preto sa do tohto nástroja zadefinovalo overovanie informácií spolu s vyhodnocovaním či dané dáta majú byť zapísané do databázy. Veľakrát nastáva aj situácia, že veľké množstvo 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 pridávajú do súčasného 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 spojeniu dvoch indexov teda nového a starého dochádza až počas samostatného vyhľadávania

Microsoft Azure

Pri vypracovávaní môjho zadania som pracoval v prostredí Microsoft aplikácie. Táto aplikácia funguje vo webovom rozhraní. V tejto aplikácii sa dajú vytvárať SQL databázy, Maria DB, Posgre SQL databázy, virtuálne stroje a mnoho ďalších aplikácií. 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. [Tutoriál 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 zbežnosť v správnom vyplňaní 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 portálu **Microsoft Azure**

[Inštalácia Azure CLI](#)

Po nainštalovaní a prihlásení 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 informácie som si zapísal do súboru **sql_database_out.txt**, ktorý je súčasťou 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 nainštalovanie 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 (Application Programming Interface) nástroj na jednoduché vytvorenie vyhľadávania na stránke. Rozhranie API a architektúra kognitívneho vyhľadávania zjednodušuje úlohu pri pridávaní sofistikovaného vyhľadávania informácií.

Vytvorenie ACS som realizoval pomocou portálu **Microsoft Azure**. Vyplňanie 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](#)

Vytvorenie 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 vyplniť formulár. Moje vyplnenie formulára môžete vidieť na nasledujúcich obrázkoch.

Obrázok na vytvorenie indexu, ktorý bude prepojený s databázou.

The screenshot shows the 'Import data' page in the Microsoft Azure Search portal. The breadcrumb navigation at the top reads 'Home > Microsoft Search > demovedecky >'. The page title is 'Import data'. Below the title, there are four tabs: '*Connect to your data' (selected), 'Add cognitive skills (Optional)', 'Customize target index', and 'Create an indexer'. A descriptive text states: 'Create and load a search index using data from an existing Azure data source in your current subscription. Azure Cognitive Search crawls the data structure you provide, extracts searchable content, optionally enriches it with cognitive skills, and loads it into an index. [Learn more](#)'. The 'Data Source' dropdown is set to 'Azure SQL Database'. A blue notification bar indicates 'Connection validated.'. The form fields are as follows: 'Data source name' is 'students'; 'Connection string' is 'Encrypt=True;TrustServerCertificate=False;Connection Ti ...' with a 'Choose an existing connection' link and an 'Authenticate using managed identity' checkbox; 'User Id' is 'azureuser'; 'Password' is masked with '*****'; a 'Test connection' button is present; and 'Table/View' is '[students]'. At the bottom, a blue button says 'Next: Add cognitive skills (Optional)'. The footer contains the URL 'https://docs.microsoft.com/azure/search/search-import-data-portal?WT.mc_id=Portal-Microsoft_Azure_Search'.

Nasleduje obrázok s vytvorením indexera.

Home > Microsoft Search > demovedecky >

Import data

×

*Connect to your data Add cognitive skills (Optional) *Customize target index **Create an indexer**

Indexer

Name *

demovedecky-indexer ✓

Schedule ⓘ

Once Hourly Daily Custom

Description

(optional)

Advanced options

Previous: Customize target index **Submit**

Následne som sa pokúsil vyhľadať pomocou ACS pomocou mojeho ID informácie o mne v databáze. Dá sa to aj realizovať pomocou iných operácií. Dokážem si v databáze zobrátiť všetkých študentov, ktorý napríklad začali štúdiium v roku 2018.



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

;

✕

...

×

☒ Delete all

Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. Tag names are case insensitive, but tag values are case sensitive.[Learn more about tags](#)

Do not enter names or values that could make your resources less secure or that contain personal/sensitive information because tag data will be replicated globally.

Name ⓘ	Value ⓘ
vedecky	https://zp.kemt.fei.tuke.sk/home

demovedecky (Search service)

1 to be added ⓘ














Apply Discard changes

Azure Blob

Pre porovnanie fungovania Azure Cognitive Search so zdrojom SQL databáza som vytvoril vyhľadávanie pomocou Blob storage. V prvom kroku som si vytvoril Blob storage. Následne som vytvoril **Container** do ktorého som nahral všetky *README* súbory zo stránky.

Blob storage

Name	Last modified	Public access level	Lease state
<input type="checkbox"/> blobvedevky	5/1/2021, 2:07:54 PM	Private	Available

Name	Modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>  README 10.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	7.07 KiB	Available
<input type="checkbox"/>  README 11.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	29 B	Available
<input type="checkbox"/>  README 12.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	8.12 KiB	Available
<input type="checkbox"/>  README 13.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	10.12 KiB	Available
<input type="checkbox"/>  README 14.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	176 B	Available
<input type="checkbox"/>  README 15.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	12.06 KiB	Available
<input type="checkbox"/>  README 16.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	7.43 KiB	Available
<input type="checkbox"/>  README 17.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	11.94 KiB	Available
<input type="checkbox"/>  README 18.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	5.12 KiB	Available
<input type="checkbox"/>  README 19.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	8.81 KiB	Available
<input type="checkbox"/>  README 2.md	5/1/2021, 2:08:09 PM	Hot (Inferred)	Block blob	4.05 KiB	Available
<input type="checkbox"/>  README 20.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	19.82 KiB	Available
<input type="checkbox"/>  README 21.md	5/1/2021, 2:08:23 PM	Hot (Inferred)	Block blob	8.1 KiB	Available

Index API version
 azureblob-index 2020-06-30-Pre...

Query string ⓘ

Search

Request URL

Results

```

1  {
2    "@odata.context": "https://vedeckydemoblob.search.windows.net/indexes('azureblob-index')/$metadata#docs(*)",
3    "value": [
4      {
5        "@search.score": 1.7794347,
6        "content": "----\ntitle: Michal Stromko\npublished: true\ntaxonomy:\n  category: [vp2021]\n  tag: [ir,cloud,demo,\n  nlp]\n  author: Daniel Hladek\n---\n\nrok začiatku štúdia: 2019\n\n# Vedecký projekt 2021\n\nKlaudové služby pre získavanie informácií\n\nCieľom projektu je zistiť ako fungujú klaudové služby pre umelú inteligenciu a ako fungujú webové vyhľadávače.\n\nÚlohy:\n\n- Zistite čo je to získavanie informácií.\n- Oboznámte sa s Azure Cognitive Search a získajte prístup k službe. Pre prihlásenie môžete použiť Vášu študentské prihlasovacie údaje.\n- Vypracujte minimálne jeden tutoriál pre prácu s Azure Cognitive Search.\n- Vypracujte krátky report na 2 strany kde napíšete čo ste robili a čo ste sa dozvedeli.\n\nZásobník úloh:\n\n- Vytvorte index a vyhľadávanie na ZP Wiki.\n- Napíšte tutoriál o tom ako ste to dokázali.\n\nStretnutie 9.4:\n\nStav:\n- Vytvorená sada tutoriálov o Azure a Azure Cognitive Search. Dobré na ZKT. Tutorály sú na [Github](https://github.com/michal552703/Vedecky-projekt).\n- Zistili sme, že ACS indexuje iba z Azure Storage.\n\nÚlohy:\n\n- Vytvoríť pokusný Azure Storage zdroj. Môže to byť Azure Blob alebo Azure Table alebo Azure SQL.\n- Použiť indexer na indexovanie obsahu zdroja.\n- Zapísať postup do súboru MD.\n\nStretnutie 19.3:\n\nStav:\n\n- Vytvorený prístup na Azure Portal\n\nÚlohy:\n\n- Vypracujte minimálne jeden tutoriál pre prácu s Azure Cognitive Search.\n\n\n\n\n",
7        "metadata_storage_content_type": "application/octet-stream",
8        "metadata_storage_size": 1439,
9        "metadata_storage_last_modified": "2021-05-01T12:08:23Z",
10       "metadata_storage_content_md5": "L31yeLn4Gr7dF56x3SmXSg==",
11       "metadata_storage_name": "README 31.md",
12       "metadata_storage_path":
"aHR0cHM6Ly9jc2IxdMDAzmjAwMDYyOGM5M2QyLmJsb2IuY29yZS53aW5kb3dzLm5ldC9ibG9idmVkZXZreS9SRUFETUULmJAZmM5tZA2",

```

V druhom kroku som si vytvoril nový ACS. V tomto ACS som vytvoril index, ktorý som spojil s Blob storage. Po tomto vytvorení som začal vyhľadávanie.

Po zadaní rôzneho parametra som zistil že vyhľadávanie podľa Blob storage je jednoduchšie, pretože nemusím vytvárať žiadnu databázu ani tabuľku. Vyhľadávanie funguje na princípe hľadania reťazcov v súboroch README.

Záver

Pri vypracovávaní zadania na predmet Vedecký projekt som sa strtol s veľkým množstvom prekážok. Najväčšou prekážkou pre mňa bolo pochopenie fungovania Porálu **Microsoft Azure**. Musel som pochopiť to, aký treba zvoliť postup pre vytvorenie takéhoto vyhľadávania. Veľa krát sa mi stala situácia že postup vytvorenia napríklad ACS sa menil každý týždeň. To znamená že pri takto rýchlej zmene vytvorenia ACS vznikali chyby. Takéto chyby som zaznamenal aj medzi aplikáciami. 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. Ďalej som vytvoril Blob storage do ktorého som nahral README súbory zo stránky. Vyhľadávanie pomocou Blob storage funguje bez problémov a konštatujem že je efektívnejšie.

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