

Semestrální práce KIV/EITM

Týmový projekt

Marek Lovčí A19N0093P
lovcim@students.zcu.cz

Jakub Vaněk
A19N0097P
vanekjak@kgm.zcu.cz

Lucie Tauchenová
A20N0055P
tauchenl@students.zcu.cz

9. 1. 2021

Obsah

1	Zadání	2
2	Technologie	3
2.1	Heroku	3
2.2	GitHub	4
2.3	Flask	4
3	Vypracování	6
3.1	UI	6
3.2	Nasazení	8
3.3	Web Crawler	9
4	Závěr	10

1 Zadání

Web Data Extractor

Zadání Web Data Extractor je zaměřeno na simulaci robotů webových vyhledávačů a indexaci webových stránek. Projekt primárně ověří možnosti těžby nestrukturovaných dat, integraci řešení napříč platformami (Github + Heroku) a napříč datovou a prezentační vrstvou.

WOW Effect: možnost indexace stránek vyjmutých z Google indexu, napodobení funkce vyhledávacího enginu s mnohem menší programovou náročností.

Sekundární efekt: Napodobení služby <https://web.archive.org/> tím, že bude uložen obsah zadaných webových stránek a potenciálně bude tento obsah časově referencovatelný. Změna oproti zmíněné službě je ta, že naše řešení bude podstatně méně datově náročné, protože nebudeme ukládat jejich kompletní obraz, nýbrž pouze obsah.

- Web Data Extraction - První část řešení je zaměřena na extrakci nestrukturovaných dat z webových stránek. Uživatel v rozhraní (prezentační vrstvě) zadá webovou stránku kterou by měl robot načíst. Data budou uložena do databáze.
- Prezentační vrstva má 2 funkce - určení výchozího bodu indexace a prezentace výsledků

Pro prezentaci výsledků bude vytvořeno webové rozhraní podobné rozhraní služby Google. To umožní vyhledávání klíčových výrazů z načtených stránek a jejich přehledné zobrazení.

2 Technologie

2.1 Heroku



Heroku je cloudová platforma podporující více programovacích jazyků. Jako jedna z první cloudových platforem byla vyvíjena od června roku 2007, kdy podporovala pouze programovací jazyk Ruby. Nyní je podpora rozšířena na programovací jazyky Java, Node.js, Scala, Clojure, Python, PHP a Go.

Aplikace běžící na heroku mají standartně unikátní doménu použitou k routování HTTP požadavků správnému kontejneru. Tyto kontejnery Heroku nazývá *dyno*. Všechny tyto kontejnery jsou rozprostřeny v síti těchto kontejnerů, které jsou uloženy na několik serverech. Git server obstarává push požadavky od povolených uživatelů. Všechny služby poskytované Heroku běží na EC2 cloud-computing platformě společnosti Amazon.

Společnost Heroku také poskytuje několik dalších služeb. Mezi nimi můžeme najít např. Heroku Postgres, poskytující cloudové řešení PostgreSQL databáze, nebo také Heroku Teams, což je nástroj pro team management.

Ceny služeb společnosti Heroku se pohybují od 0\$ pro studenty, až po několik 1000\$ měsíčně za jednotlivé kontejnery pro aplikace s vysokou náročností.

2.2 GitHub



GitHub je webová služba podporující vývoj software za pomoci verzovacího nástroje Git. Mezi vývojáři je velmi oblíbený, což dokazuje přes 100 milionů repozitářů. Poskytuje funkce sociálních sítí – notifikace o změnách, diskuze nad kódem, návrhy změn, či zasílání vlastních řešení (pull-requesty). Poskytuje také možnost psaní vlastní wiki, systém pro issue tracking, historii verzování a mnoho dalšího.

GitHub také nabízí službu *Github Actions*. Tato služba automatizuje celý proces nasazení software. Kód je sestaven, otestován a nasazen přímo z githubu. Poskytuje tak celý systém pro Continuous Integration a Continuous Delivery

2.3 Flask



Flask je mikro webový framework napsaný v programovacím jazyce Python. Je klasifikován jako mikro webový framework, protože nevyžaduje konkrétní nástroje ani další vnitřní knihovny. Nemá žádnou vrstvu abstrakce databáze, ověřování formulářů ani žádné jiné komponenty třetích stran poskytující běžné funkce. Výhodou je však fakt, že je o všechny tyto funkce jednoduše rozšiřitelný.

Hlavní filosofií tohoto frameworku je udržet jádro webové aplikace jednoduché, ale rozšiřitelné.

3 Vypracování

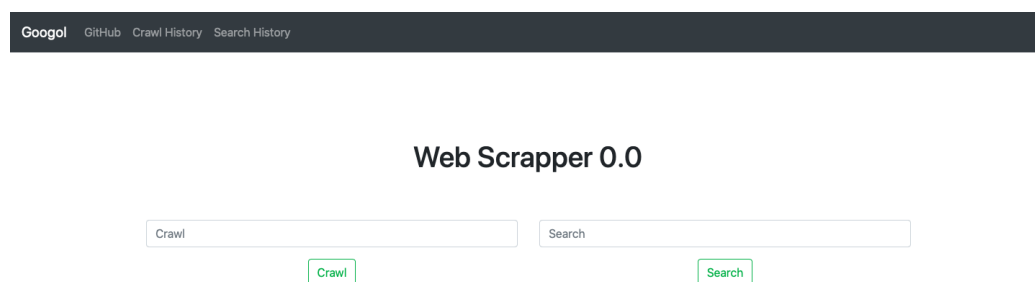
Součástí této semestrální práce byl vytvořen tzv. Web Crawler s webovým vyhledávačem. Pomocí služby GitHub Actions je tato webová aplikace sestavena a následně automaticky otestována a nasazena na cloudovou platformu Heroku.

3.1 UI

Webová aplikace se skládá z několika pohledů. Na jednotlivé pohledy se můžeme dostat z horní lišty webové aplikace. Těmito pohledy jsou:

- Hlavní stránka
- Github
- Crawl History
- Search History

Hlavní stránka obsahuje dvě textová pole. Do prvního uživatel zadá, které stránky chce prohledávat. Stránku potvrdí tlačítkem Crawl. Toto tlačítko spustí Web Crawler, které danou stránku načte do databáze. Více o funkcionalitě web crawleru je popsáno v samostatné sekci v kapitole vypracování.



Obrázek 3.1: Hlavní stránka

Druhé textové pole slouží k zadání klíčových slov. Ty je možné vyhledávat na stránkách, které byly nastaveny k prohledání. Po jeho kliknutí je otevřen seznam výskytů klíčových slov na prohledaných stránkách včetně náhledů.

Searched word: "galerie" - Number of results: 3 (0.0 s)

Praha – Wikipedie

... V. Milunice a F. Gehryho Praha je kulturní metropolí celé České republiky, Evropské město kultury 2000 . [80] muzeí, galerií , divadel , kin Národní **galerie** v Praze Veletržního paláce Mucha , Picasso , Monet Van Gogh . Každoročně se zde koná nejnavštěvovanější festival v Česku, světelný Signal Festival

<https://cs.wikipedia.org/wiki/Praha>

Západočeská univerzita v Plzni – Wikipedie

...kulta právnická Fakulta filozofická Fakulta pedagogická Fakulta zdravotnických studií Univerzita třetího věku Ústav jazykové přípravy Menza Kollárova **Galerie** Ladislava Sutnara Univerzitní knihovna Kávárna Družba Kulturka Výzkumné centrum NTC Cheb [editovat editovat zdroj] Fakulta ekonomická v Chebu Letní...


https://cs.wikipedia.org/wiki/Západočeská_univerzita_v_Plzni

Plzeň – Wikipedie

...1 Střední školství 6.1.2 Vyšší odborné školství 6.1.3 Vysoké školství 6.2 Kultura 6.2.1 Divadla 6.2.2 Kina 6.2.3 Rozhlas a televize 6.2.4 Muzea 6.2.5 **Galerie** 6.2.6 Knihovny 6.2.7 Významné kulturní akce 6.3 Sport 7 Pamětihodnosti 8 Osobnosti 8.1 Rodáci 8.2 Studenti 8.3 Ostatní osobnosti Plzeň 9 Partnerská m...

<https://cs.wikipedia.org/wiki/Plzeň>

Go to Page



Tančící dům od V. Milunice a F. Gehryho

Praha je kulturní metropolí celé České republiky, Evropské město kultury 2000.^[80] Působí zde desítky muzeí, galerií, divadel, kin a nejvýznamnějších kulturních institucí. Národní galerie v Praze spravuje největší sbírku výtvarného umění v Česku. Ve stálé expozici Veletržního paláce jsou díla světových umělců jako např. Mucha, Picasso, Monet nebo Van Gogh. Každoročně se zde koná nejnavštěvovanější festival v Česku, světelný Signal Festival.

Pražský magistrát vnakládá na

Obrázek 3.2: Výsledky vyhledávání

Následně záložky Crawl History a Search History slouží k nahlédnutí do historie vyhledávání na stránce. Crawl History obsahuje seznam stránek přidáných k prohledání. Po kliknutí na jeden z odkazů je otevřen pohled, na kterém je vidět, kdy byla stránka prohledávána. Obdobně v záložce Search History je seznam vyhledávaných slov s počtem výsledků a s možností dostat se na stránku.

Crawl History

Plzeň – Wikipedie Plzeň Pilsen) je statutární město Čech Plzeňského kraje . Leží na soutoku řek Mže , Radbuza , Úhlava Úslava , z nichž vzniká řeka Berounka . Žije zde přibližně 175 tisíc [2] čtvrtým největším městem v Česku . V Plzeňské metropolitní oblasti žije podle... https://cs.wikipedia.org/wiki/Plzeň	13:31, 26.04.2021
Západočeská univerzita v Plzni – Wikipedie Západočeská univerzita v Plzni univerzitní vysoká škola Plzni , která byla založena roku 1991 Vysoké školy strojní a elektrotechnické Pedagogické fakulty v Plzni . Obsah 1 Obecné informace 1.1 Historie 1.1.1 Seznam rektorů 1.2 Mezinárodní spolupráce... https://cs.wikipedia.org/wiki/Západočeská_univerzita_v_Plzni	13:25, 26.04.2021
Praha – Wikipedie Praha německy Prag ; v jiných jazycích často Prague Praga) je hlavní město Česka , zároveň je 14. největším Evropské unie . Leží mírně na sever od středu Čech Vltavě , uvnitř Středočeského kraje , jehož je správním centrem, ale není jeho... https://cs.wikipedia.org/wiki/Praha	13:10, 26.04.2021

Previous **1** Next

Obrázek 3.3: Historie prohledaných stránek

Search History

galerie	13:32, 26.04.2021
Pomník	13:31, 26.04.2021
děkan	13:31, 26.04.2021
absolvent	13:31, 26.04.2021
doktor	13:31, 26.04.2021
rektor	13:31, 26.04.2021

Previous **1** 2 ... 5 Next

Obrázek 3.4: Historie hledání na stránkách

3.2 Nasazení

K nasazení webové aplikace na platformu Heroku bylo využito služby Github Actions. Konfigurace je popsána v souboru ve formátu yml. Na začátku je

nastavena akce, při které bude SW nasazen. Nastaveno na akci *push. on*:

push

. Poté je nastaven OS, na kterém aplikace poběží (*ubuntu-latest*) a nastavené 3 verze pythonu, pro které aplikaci testujeme. Těmito verzemi jsou verze 3.6, 3.7 a 3.8. V poslední části jsou nastaveny kroky pro správné nasazení aplikace. Nejdříve je nastavení a linkování verze pythonu. Následně je python nainstalován. Dále je načten soubor *requirements.txt*, ve kterém jsou popsány verze jednotlivých nástrojů, které jsou v naší aplikaci použity (Flask==1.1.2 apod.). Následně je vytvořena databáze, která je popsána v souboru *models.py*.

Toto nastavení zařídí nasazení otestované aplikace na server, nebo upozornění na chybu v opačném případě.

3.3 Web Crawler

Web Crawler je internetový bot, který automaticky prochází webové stránky. Začíná se seznamem URL adres k návštěvě. Z těchto stránek uloží do databáze jejich obsah. Během tohoto načítání identifikuje i veškeré odkazy (obsah atributů HTML atributů *src* a *href*) a ty přidá do seznamu URL k pozdějšímu prohledávání. Webový vyhledávač pak na dotaz uživatele může díky databázi odpovědět, na kterých stránkách jsou hledaná slova k nalezení.

4 Závěr

V semestrální práci byla vypracována webová aplikace simulující web crawler s následnou možností vyhledávání klíčových slov na prohledávaných stránkách. Aplikace je pomocí služby GitHub Actions automaticky sestavena, otestována a nasazena na server platformy Heroku.