



Seghegfa search engine

Dokumentácia

Obsah

Užívateľská dokumentácia	2
Programátorská dokumentácia	3
Opis programu.....	3
Cieľ programu	3
Získanie	3
Inštalácia programu	3
Spustenie programu.....	3
Popis architektúry.....	3
Popis základného použitia	5

Užívateľská dokumentácia

Aplikácia Seghegfa je nástroj na vyhľadávanie slov v textových dokumentoch uložených na v pamäti počítača.

Na začiatku užívateľ načíta pomocou tlačidla *Vybrať súbory* textové súbory, v ktorých chce svoj hľadaný výraz vyhľadávať. Následne potvrdí svoju voľbu súborov pomocou tlačidla *Nahraj*.

Potom je užívateľ presmerovaný na stránku, kde nájde:

1. Vyhľadávacie pole, do ktorého môže zadať slovo, ktoré chce vyhľadávať.
2. Pole na zadanie parametru šírky vyhľadávacieho okna.
3. Zoznam nahratých súborov s ich názvami.
4. Tlačidlo *Hľadať*, pomocou ktorého spustí vyhľadávanie.

Po zadaní hľadaného slova do vyhľadávacieho poľa môže užívateľ stlačiť tlačidlo *Hľadať*. Následne je presmerovaný na stránku s výsledkami svojho hľadania. Jednotlivé výsledky sú prezentované v oddelených sekciách a hľadané slovo je v nich farbene zvýraznené. V prípade, že užívateľa zaujíma aj širší kontext hľadaného slova, môže nájdený výsledok rozkliknúť pomocou myši. Novo otvorené okno môže následne zavrieť pomocou kliknutia na malý sivý krížik v rohu okna alebo kliknutím mimo okna.

V prípade, že chceme zmeniť vyhľadávaný výraz vrátíme sa zo stránky s výsledkami na stránku s vyhľadávacím polom pomocou tlačidla *Návrat*, ktoré je zvyčajne umiestnené v ľavom hornom rohu prehliadača. Pokiaľ užívateľ potrebuje znovu nahrať nové či iné súbory, môže kliknúť na logo Seghegfa, ktoré ho po kliknutí presmeruje na úvodnú stránku.

Programátorská dokumentácia

Opis programu

Aplikácia Seghegfa je pythonovský program využívajúci knižnicu flask, ktorý slúži na vyhľadávanie zadaného slova v nahratých textových dokumentoch. Jednotlivé dokumenty sú na strane servera nahraté a rozdelené na tokeny. Spracované dáta sú uložené do session. Následne sa testujú všetky texty na prítomnosť hľadaného slova, pričom v prípade nájdania slova je užívateľovi v podobe html stránky prezentovaný výsledok – slovo a jeho okolie s nastaviteľnou šírkou. Program nájde všetky výskyty slova v dokumente. Jednotlivé výskyty sú klikateľné, a po kliknutí sa užívateľovi ukáže nájdené slovo v širšom kontexte dokumentu.

Cieľ programu

Cieľom programu je nájsť používateľom zadané slovo v nahratých textových dokumentoch a odprezentovať výsledky v prehľadnej forme webovej stránky s ukážkou najbližšieho okolia nájdeného slova.

Získanie

Celý program je voľne dostupný na: <https://github.com/JakubCiesko/seghegfa>

Inštalácia programu

Pred spustením programu musíme mať nainštalovaný python a knižnicu flask.

Spustenie programu

Program sa dá spustiť pomocou príkazového riadku, v ktorom máme otvorený adresár so všetkými súbormi z <https://github.com/JakubCiesko/seghegfa>. Na spustenie serveru použijeme príkaz `py server.py` a v prehliadači otvoríme `localhost:5000` alebo `127.0.0.1:5000`, resp. adresu zobrazenú v príkazovom riadku.

Popis architektúry

Na začiatku je volaná funkcia `index`, ktorá navracia html template úvodnej stránky `index.html`. Následne sa po HTTP post metóde pomocou funkcie `upload` načítajú nahraté dáta. Pred samotným načítaním prebehne v rámci funkcie `upload` kontrola veľkosti a formátu každého súboru. Prednastavená maximálna veľkosť súboru je 3,28 MB a prípona súboru `txt`. Dáta sa

v rámci funkcie upload spracujú pomocou funkcie process_data. Vo funkcii process_data sa načítava obsah súboru postupne po riadkoch, ktoré sú následne zbavené interpunkcie a pomocou metódy split rozdelené na tokeny. Výstupom funkcie je potom dictionary s názvami súborov ako kľúčmi a spracovanými tokenami ako hodnotami. Takto spracované dáta sa uložia vo funkcii upload do flask session. Funkcia upload potom navracia html template search_page.html, do ktorého pomocou jinja2 zapisuje názvy spracovaných súborov.

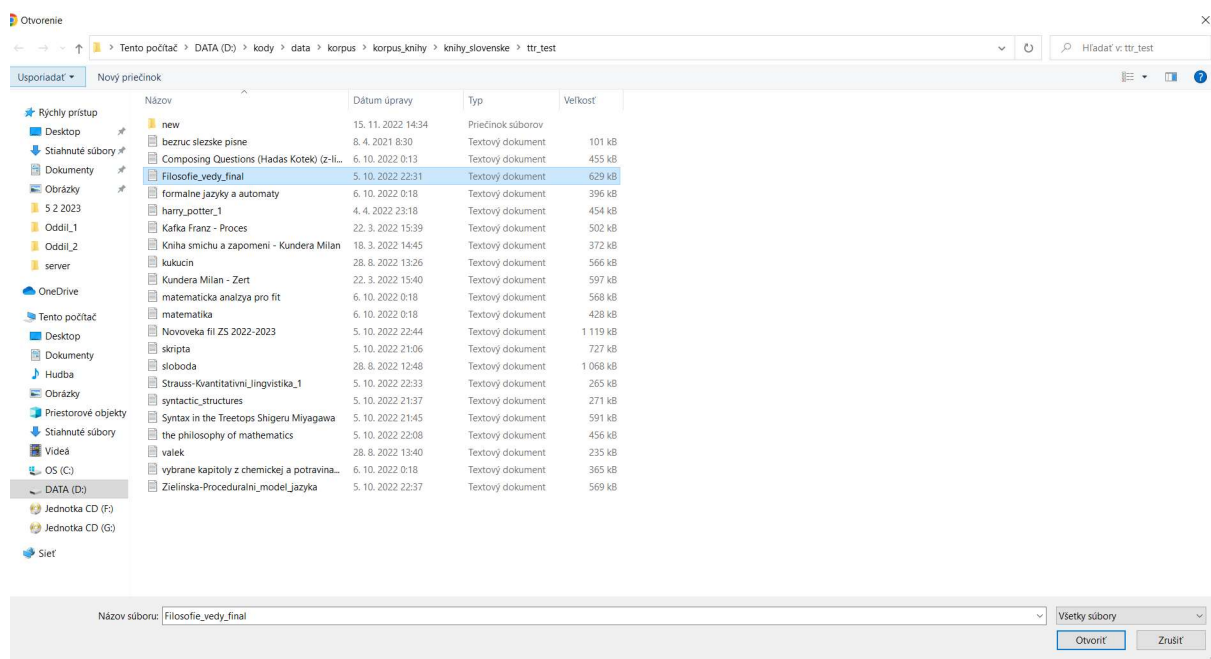
Užívateľ je po spracovaní dát presmerovaný na stránku s výsledkami /search. Po zadaní vyhľadávaného slova do poľa pomocou HTTP metódy get získavame hodnotu premennej q_word. Tá vstupuje vo vnútri funkcie search, spolu s window_size (tiež nastaviteľné užívateľom), a spracovanými dátami ako argument do funkcie search_texts. Funkcia search_texts prechádza jednotlivé spracované texty a hľadá v nich všetky výskyty stringu q_word. V prípade úspechu zapíše do premenných (dict) snippets a texts užší a širší kontext nájdeného slova, tieto záznamy sa ukladajú ako hodnota pre kľúč s názvom súboru. Funkcia search_texts navracia užší a širší kontext hľadaného slova – dictionary snippets a texts. Výstup tejto funkcie je následne zobrazený užívateľovi pomocou html template search_results.html, do ktorého sú dynamicky pridané výsledky hľadania.

Popis základného použitia

Webová aplikácia Seghegfa slúži na vyhľadávanie zadaného výrazu v nahratých textových súboroch. Na obrázku č. 1. vidíme vstupnú stránku aplikácie. Po stlačení tlačidla *Vybrať súbory* sa otvorí okno (obrázok č. 2), v ktorom zvolíme textové súbory, ktoré chceme nahráť a v ktorých chceme vyhľadať hľadané heslo. Voľba súborov sa potvrdzuje pomocou tlačidla *Otvoriť*, a následne pomocou tlačidla *Nahraj*.




Obrázok č. 1: Vstupná stránka webovej aplikácie Seghegfa slúžiaca na nahranie súborov.



Obrázok č. 2: Dialógové okno slúžiace pre výber a nahranie textových súborov na disku počítača.

Po stlačení tlačidla *Nahraj* server spracuje požiadavku užívateľa a presmeruje ho na novú stránku. Na tejto stránke (obrázok č. 3) nájdeme informácie o nahratých súboroch (ich názvy) a vyhľadávacie pole, do ktorého môžeme zadať nami vyhľadávané slovo. Taktiež môžeme zvoliť hodnotu šírky výberu pomocou poľa vedľa nápisu šírka výberu. Šírka výberu je prednastavená na hodnotu 3, čo znamená, že sa nám budú zobrazovať výsledky s nájdeným slovom a jeho okolím 3 slová doľava a 3 slová doprava.



Seghegfa

Šírka výberu:


Hľadaj

Nahrané súbory:

Syntax in the Treetops Shigeru Miyagawa.txt
the philosophy of mathematics.txt
vybrane kapitoly z chemickej a potravinárskej chemie.txt

Vytvoril: Jakub Čieško

Obrázok č. 3: Stránka s informáciami o nahratých súboroch, vyhľadávacím polom a nastaviteľným parametrom šírky výberu.



Seghegfa

Šírka výberu:

Hľadaj

syntax

Nahrané súbory:

Syntax in the Treetops Shigeru Miyagawa.txt
the philosophy of mathematics.txt
vybrane kapitoly z chemickej a potravinárskej chemie.txt

Vytvoril: Jakub Čieško

Obrázok č. 4: Ukážkové vyplnenie poľa s vyhľadávaným slovom.

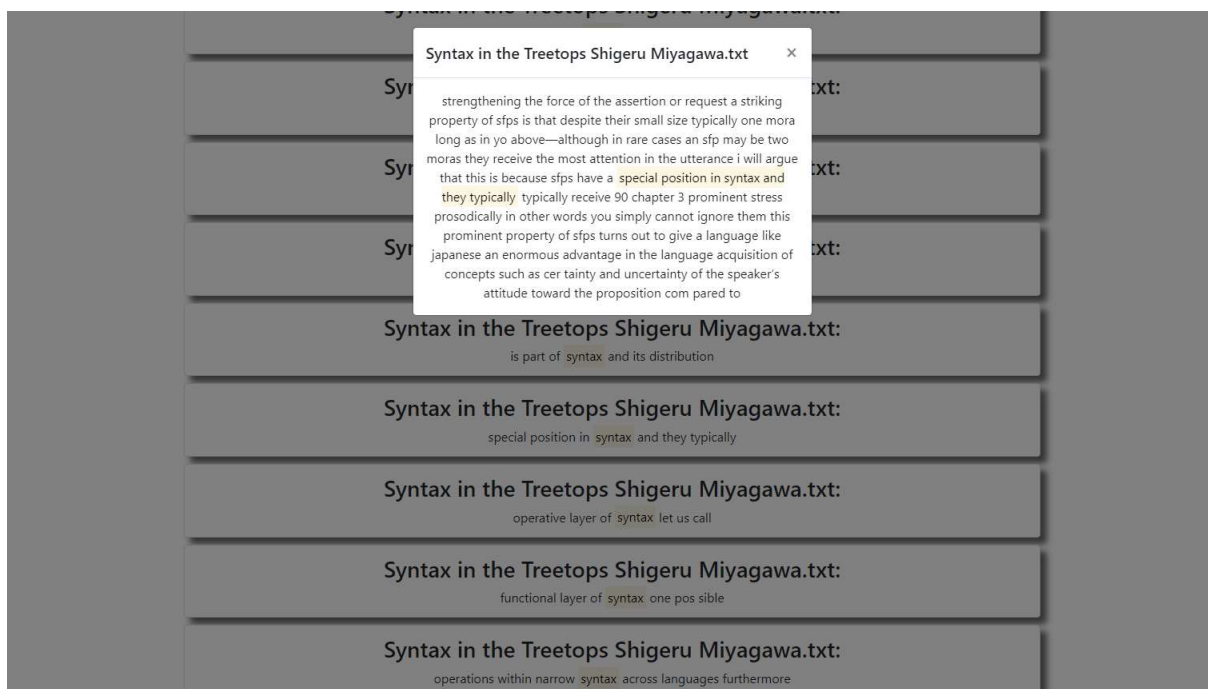
Na obrázku číslo 4 vidíme príklad zadania vyhľadávaného slova (*syntax*) a šírky výberu 3. Po zadání všetkých potrebných parametrov (vyhľadávaného slova a prípadne šírky výberu) a stlačení tlačidla *Hľadať* sa dostaneme na stránku s výsledkami nášho hľadania.

V jednotlivých sekciách môžeme vidieť (viď obrázok číslo 5) všetky výskyty nami zadaného slova v nahratých dokumentoch. Hľadané slovo je farebne odlišené a každý výskyt je označený názvom súboru, v ktorom bolo dané slovo nájdené. Každá sekcia s výskytom je interaktívna – po kliknutí na ňu sa otvorí okno (viď obrázok číslo 6), v ktorom je širšie okolie hľadaného slova v danom dokumente. Po otvorení okna je pôvodné okolie slova znova farebne odlišené. Okno môžeme zavrieť pomocou kliknutia na malý sivý krížik vnútri okna alebo kliknutím mimo plochu okna.

V prípade, že chceme zmeniť vyhľadávaný výraz, môžeme sa vrátiť na predchádzajúcu stránku viditeľnú na obrázkoch číslo 3 a 4 pomocou tlačidla *Návrat* umiestneného zvyčajne v ľavom hornom rohu prehliadača. Ak chceme celý proces nahratia súborov zopakovať od začiatku, stačí kliknúť na logo pandy Seghegfa. Po stlačení sa dostaneme na vstupnú stránku vyobrazenú na obrázku číslo 1.



Obrázok č. 5: Stránka s výsledkami vyhľadávania slova *Syntax*. Jednotlivé výskyty slova sú vyznačené.



Obrázok č. 6: Zatváratel'né okno otvorené kliknutím na výsledok hľadania. Obsah tohto okna tvorí úsek nahratého textu s vyznačeným nájdeným slovom a jeho okolím.