vi2022

Matúš Krajčovič

Téma projektu

W6 - Anchor texts a štatistika k anchor textom. Document frequency, collection frequency.

Popis problému

Podľa wikipédie [https://en.wikipedia.org/wiki/Anchor_text] sa pod názvom "anchor text" skrýva viditeľný popisok každého hypertextového linku. V HTML kóde sa tento popisok skrýva v atribúte tagu `<a>`:

```
<a href="link">Anchor text</a>
```

Takýchto linkov sa na internete nachádza veľmi veľa. V našom projekte budeme analyzovať tieto popisky nachádzajúce sa v článkoch wikipédie. Použijeme dostupný dataset anglickej wikipédie [https://dumps.wikimedia.org/enwiki/latest/] s údajmi v XML štruktúre.

Je všeobecne známe, že alternatívne popisky pri HTTP linkoch sú veľmi dôležité nie len pre dobré SSO a skóre vo vyhľadávaní ale aj z hľadiska používateľského zážitku, teda použiteľnosti a prístupnosti. Popisok linku by mal byť dostatočne opisný aby bolo zrejmé, kam sa po kliknutí naň dostaneme. Podľa (Nakayama, 2008) [https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.3046&rep=rep1&type=pdf] sú pri extrakcii dát z wikipédie linky vhodné napríklad na hľadanie synoným (link adresa a anchor text) alebo hľadanie významov slov.

Konzultácia č. 1: 27. 9. 2022 - vybraná téma W6 - Anchor texts a štatistika k anchor textom. Document frequency, collection frequency (čo sa týka štatistík porovnáva sa to s anchor textami ako aj s celými textami) (Python)

Psudokód riešenia problému na testovacích dátach

Pri riešení problému na skúšobných dátach budem postupovať nasledovne:

- stiahnem si menšiu časť z celého datasetu wikipédie
- využijem jazyk python na načítanie údajov (knižnica xml.etree) - postupne prejdem cez všetky stránky `<page>` v súbore a aj za pomoci regulárnych výrazov si pri každom anchor texte uložím:
    - text linku
    - názov článku
- tieto linky si uložím
- zo získaných dát vytvorím štatistiku:
    - document frequency
    - collection frequency
- výsledok je vo forme jednoduchého programu do ktorého zadáme slovo a vypíše nám štatistiky ku danému slovu alebo slovnému spojeniu
- po úspešnej implementácii vyskúšame tento program aj na celom datasete wikipédie

Konzultácia č. 2: Pseudokód OK, nasledujúca konzultácia vzorový kód.

Konzultácia Predvedenie: 11. 10. 2022 - Kód na sub-datasete OK. Používa regex.

Súčasné riešenia

Wikipédia je tu s nami už veľmi dlho, existuje teda veľa riešení, ktoré z nej dolujú dáta. (Milne, 2013) [https://www.sciencedirect.com/science/article/pii/S000437021200077X] vytvorili open-source projekt, ktorý dovoľuje výskumníkom

ale aj developerom získať kontrolu nad sémentikou wikipédie. Všetky články či kategórie sú reprezentované ako triedy, no ponúkajú aj paralelizované spracovanie wikipedia dumpov, o čo sa budeme snažiť aj my v našom projekte.

Podľa (Zesch, 2007) [https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.411.2044&rep=rep1&type=pdf] wikipédia obsahuje nesmierne množstvo informácii, hlavne tých sémantických, ktoré vieme dolovať aj zo spomínaných linkov - každý link hovorí o vzťahu dvoch stránok, resp pojmov. Pomocou linkov vieme zistiť podobnosť pojmov, pričom opisok nemusí byť nutne rovnaký ako samotný link, čo nám dáva ďalšiu informáciu navyše (napr. synonymá). Podľa (Zhu, 2019) [https://link.springer.com/article/10.1007/s10489-019-01452-1] sú práve odkazy najlepšou metrikou vzťahu dvoch pojmov. Tieto sémantické informácie využili na vylepšenie modelov podobnosti slov. V (Hong, 2017) [https://journals.sagepub.com/doi/abs/10.1177/0165551517693497] tiež využívali sémentickú informáciu odkazov na odporúčanie tagov.

Trošku mimo témy, ale veľmi zaujímavý výskum ponúka aj (Dimitrov, 2016) [https://dl.acm.org/doi/abs/10.1145/2872518.2889388?casa_token=BnlFD_wJuDsAAAAA:igLOwkLZuVBKabDqmi0BNAxbScVXlQEafVh0KgFN9kRU3xneRQJlb8hL8hIGpI8oaxieGdzos39BL3U], ktorý hovorí aj o dôležitosti pozícii linkov. Používatelia majú preferenciu klikať na odkazy v ľavej časti obrazovky.

Niektoré riešenia a ich zdrojové sú dostupné aj na githube, či dokonca aj vo forme npm nbalíkov. Sú to napr. wikiextractor [https://github.com/attardi/wikiextractor] alebo mapwiki [https://www.mattmoocar.me/blog/MapWiki/].

Konzultácia č. 3: 26. 10. 2022 - Kód na väčšom sub-datasete. Vlastné indexovanie.

Popis riešenia

Extrakcia dát z datasetu

Pri extrakcii dát som použil nástroj spark s knižnicou pre spracovanie XML súborov. Postupne som si rozparsoval celý súbor a počas toho som extrahoval informácie z jednotlivých stránok. Zameral som sa na tagy `<title>`, z ktorého som získal názov článku a `<text>`, ktorý obsahuje samotný text článku.

Pomocou regulárneho výrazu som extrahoval všetky linky v texte v tvare `\[\[link\]\]`, ktoré nemajú prexif (napr. Category). Linky som vyčistil od anchorov a vybral som ich alternatívny popisok - anchor text. Následne som ku každému linku priradil všetky články, v ktorých sa nachádza aj s počtom výskytov. Z nich som vypočital collection a document frekvencie, ktoré boli spolu s názvom linku vo výstupnom súbore.

Vyhľadávanie

Z výstupného súboru som som pomocou knižnice PyLucene vytvoril index. Použil som tri stĺpce, jeden s fulltextovým vyhľadávaním pre názov linku a zvyšené dva pre document a collection frekvenciu. Vytvoril som tiež python skript, v ktorom viem zadávať queries a na výstupe mi zobrazí nájdené linky spolu s ich frekvenciami na wikipédii = zoradené podľa collection frequency.

Testovanie

Samotné parsovanie som testoval ručne - z veľkého datasetu som si vybral niekoľko wiki stránok v xml štruktúre a vo výsledku som hľadal, či obsahuje všetky linky, ktoré sa na daných stránkach nachádzali a či sa zhodujú aj dokument a collection frekvencie (niektoré linky som aj pomenil či pridal nech sa to dá otestovať).

Použité dáta

Pri práci na našom projekte sme použili dostupný dataset anglickej wikipédie [https://dumps.wikimedia.org/enwiki/latest/] s údajmi v XML štruktúre:

```
<siteinfo/>
<page/>
  <title/>
  <id/>
  ...iné tagy...
  <text/>
    infobox
```

```
    samotný text článku
<page/>
<page/>
...[ďalšie stránky]...
```

Tag `<title>` obsahuje názov článku a tag `<text>` zas samotný obsah. V texte sa nachádza množstvo štruktúr, z toho asi najvýznamnejšie sú linky:

```
[[link|alt]]
```

Tieto linky môžu mať rôzne prefixy, ako napríklad Category, Wiki, User, alebo anchory kategórií.

Použitie softvéru

V priečinku so súborom parser.py a vstupným wikipedia dumpom sa spustí skript parser.sh:

```
./parser.sh wiki_dump_file output_file
```

V tomto skripte sa pripravia súbory a spustí spark-submit príkaz.

Pri indexovaní zas využijem skript index.sh a za pomoci index.py skriptu a výstupu z parsovania vytvorím index:

```
./index.sh parser_output_file index_name
```

Pri vyhľadávaní cez search.sh skript si za pomoci search.py súboru a priečinku s indexom viem spustiť vyhľadávanie:

```
./search.sh index_name
```

Teraz môžem zadávať požiadavky pre vyhľadávanie.

Konzultácia č. 4: 22. 11. 2022 - Pracuje s celým datasetom. Paralelné spracovanie robil na svojom notebooku. Celé riešenie OK. Dopracuje UX.

Konzultácia č. 5: 29. 11. 2022 - Celý projekt OK.
