Pospešitev asistenta z uporabo razpršitvene tabele

Anže Marinko, mentor: Gregor Graselli
Inštitut Jožef Stefan - Ljubljana
februar 2019

I. Uvod

V tem prispevku bomo predstavili doprinos uporabe razpršitvene tabele (RT) k pospešitvi e-asistentovega iskanja odgovora na vprašanje.

Če asistent primerja dobljeno vprašanje z vsakim izmed vzorcev v bazi, porabimo veliko časa za iskanje, zato bi si želeli najti množico vzorcev, ki jih je sploh relavantno pregledovati.

II. Metoda

Vsaki besedi v vprašanju želimo pripisati seznam indeksov vzorcev iz baze. Lahko bi iz črk v besedi izračunali neko vrednost, ki bi ji bil pripisan tak seznam, a zdi se, da je dovolj dobro, če ne celo bolje, da je ta funkcija injektivna in celo bolje identiteta, torej da je rezultat funkcije na besedi kar beseda sama. V primeru asistenta se je identiteta odlično izkazala.

Za vsak koren, ki nastopa v vzorcih, sestavimo seznam indeksov vzorcev, v katerih nastopa, in jih zložimo v razpršitveno tabelo. Ko dobimo vprašanje, za vsako besedo v njem privzamemo za kandidate za korene kar vse rezine od začetka besede. Za vsakega kandidata razberemo seznam indeksov in unija vseh takih seznamov vseh kandidatov za korene iz celotnega vprašanja je kar seznam indeksov, ki so kandidati za odgovor.

III. Rezultati

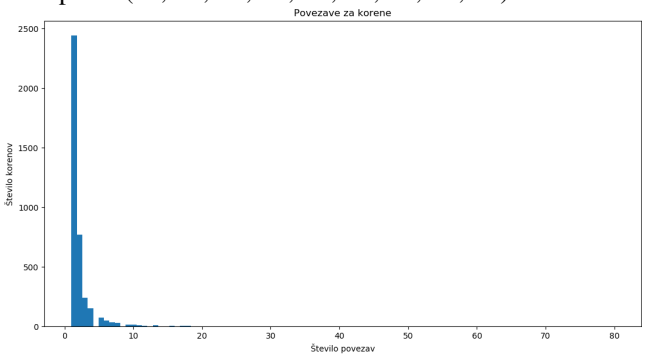
Oglejmo si rezultate dobljene ob testiranju na občinskem e-asistentu za Ljubljano.

A. Porazdelitev korenov v bazi

Tabela povezav vsebuje 3906 korenov, povprečno število povezav na koren pa je 2.112.

Porazdelitev števila povezav na koren:

- 1. kvartil: 2
- mediana: 1
- 3. kvartil: 1
- Top 10: (80, 53, 53, 43, 38, 33, 30, 30, 28)



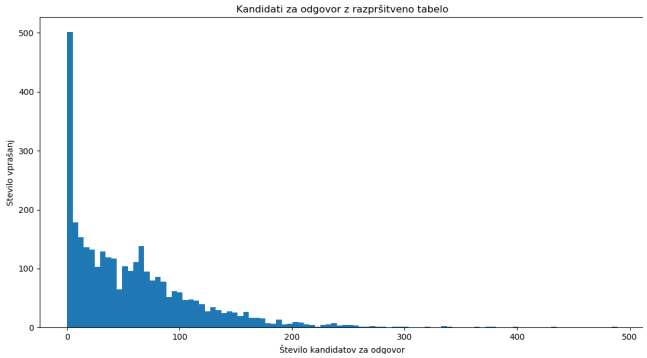
B. Število kandidatov za odgovor

Testirano je bilo na dejanskih vprašanjih iz prakse.

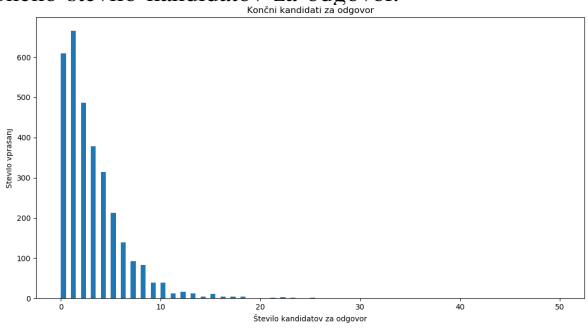
Baza vprašanj vsebuje 3148 vprašanj, število vzorcev vprašanj v bazi pa je 2105. Povprečno število kandidatov na vprašanje je 57.815.

Nekaj statistik o številu kandidatov na vprašanje:

- 1. kvartil: 84
- mediana: 45
- 3. kvartil: 13
- Top 10: (489, 435, 398, 379, 375, 364, 339, 335, 333)



Na koncu asistent dobi le še nekaj vzorcev med katerimi naključno izbere odgovor, vsi ized teh vzorcev pa nastopajo tudi med kandidati za odgovor ob uporabi RT. Oglejmo si končno število kandidatov za odgovor.

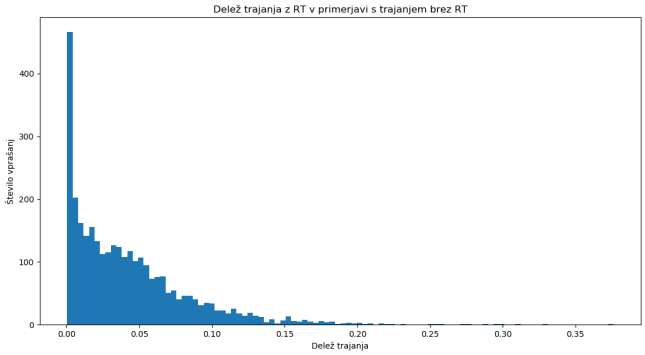


C. Doprinos uporabe RT

Oglejmo si porazdelitev časa iskanja odgovora z in brez RT:

Relativni čas iskanja odgovora definirajmo kot (čas z uporabo RT / čas brez uporabe RT). Povprečen relativen čas iskanja odgovora na vprašanje je 0.04293.

- 1. kvartil: 0.06158
- mediana: 0.03277
- 3. kvartil: 0.01061
- Max 10: (0.37621, 0.32926, 0.30867, 0.29997, 0.29354, 0.28684, 0.27597, 0.27431, 0.25793)



IV. Razprava

Opazimo, da je v empirično najslabšem primeru iskanje z uporabo RT 3-krat hitrejše kot iskanje brez RT. V povprečju pa smo z RT 23,3-krat hitrejši.

To so rezultati dosedanjega dela, ki pa še ni optimalno. Verjetno bi bilo dobro poizkusiti tudi drugačne podatkovne strukture od navadnega python-ovega slovarja.