Poročilo drugega projekta pri IEPS

Jan Novak Domijan, Urša Kumelj, Manca Drašček
UL FRI

POVZETEK

Projekt zajema ekstrakcijo receptov s Kulinarika.net, njihovo čiščenje in shranjevanje v vektorsko podatkovno bazo za lažje iskanje po vsebini. Najboljše rezultate pri iskanju so omogočile vektorske vložitve, izračunane iz naslovov, oznak in časa priprave z modelom LaBSE. Sistem je zato omejen na odgovarjanje na vprašanja, ki so povezana s temi tremi elementi.

1 UVOD

Cilj tega projekta je izluščiti uporabne informacije iz izbrane spletne strani https://www.kulinarika.net/recepti/sladice in jih shraniti v vektorsko podatkovno bazo za odgovarjanje na vprašanja iz specifične domene. V okviru tega projekta smo izolirali ključne podatke (kot so recepti, sestavine in postopki priprave), jih standardizirali in optimizirali za iskanje podobnosti v vektorskem prostoru. Končni cilj je omogočiti hitro in natančno odgovarjanje na uporabniška vprašanja v okviru naše domene.

2 ČIŠČENJE SPLETNIH STRANI

Spletne strani so sestavljene iz različnih HTML značk. Med analizo smo opazili, da vsaka stran vsebuje nezaželene reklame na levem, desnem in spodnjem delu, ki pa niso relevantne za naš projekt. Glavna vsebina, ki nas zanima, se nahaja znotraj značke `<section>` z ID-jem `recepti`. Na podlagi te ugotovitve smo razvili funkcijo, ki v bazo podatkov shranjuje le prečiščeno vsebino.

2.1 Segmentiranje prečiščene vsebine

Iz analizirane vsebine smo izluščili ključne informacije:

- Naslov recepta
- Podnaslov
- Sestavine
- Postopek priprave
- Komentarji
- Oznake (tags).

Za uspešno ekstrakcijo teh podatkov smo uporabili knjižnici `lxml` in `re`.

2.1.1 Opis. Opis smo strukturirali v koherenten stavek za lažje izračunavanje vektorskih vložitev: 'Opis recepta je DESCRIPTION. Za recept porabimo TIME. Recept je DIFFICULTY.', kjer je DESCRIPTION podnaslov posameznega recepta, TIME potreben čas za pripravo recepta (ekstrahiran z regularnimi izrazi) ter DIFFICULTY slovar prirejen za ubesedenje zahtevnosti. Ta slovar je oblikovan kot

- 1: zelo lahek
- 2: lahek
- 3: srednje težek
- 4: težek
- 5: zelo težek.

Za določitev stopnje zahtevnosti smo prešteli število oranžno obarvanih ikon kuharske kapice (od skupno petih), kar smo implementirali z uporabo XPath izrazov, podobno kot pri ekstrakciji opisa.

2.1.2 Postopek priprave. Za ekstrakcijo postopka priprave smo uporabili XPath, kar je predstavljalo najbolj učinkovito rešitev. Natančneje, izvlekli smo vsebino iz značke `<div>` z ID-jem `postopek`, ki vsebuje celoten opis kulinaričnega postopka.

2.1.3 Sestavine. Ekstrakcija sestavin je bila najbolj zapleten del procesa, saj recepti niso dosledno strukturirani. V idealnem primeru so sestavine organizirane v značkah `<p>`, z ločenimi sekcijami za posamezne komponente (kot so testo, preliv ali nadev), ki jih označuje razred `class='cf poglavje'`. Vendar pogosto sestavine padejo pod splošnejši razred `class='cf'`. Za zagotovitev doslednosti in lažje računanje vektorskih vložitev smo razvili posebno funkcijo, ki pretvori vse merske enote v besedne oblike. Končno se vse sestavine združijo v en sam niz, ohranjajoč logično strukturo recepta, kar kasneje omogoča lažjo analizo podatkov.

2.1.4 Oznake (tags). Oznake (tags) smo pridobivali iz značk `<a>`, ki so bile organizirane v hierarhični drevesni strukturi, kar je ekstrakcijo naredilo precej enostavno. Kljub temu da so bile oznake vgnezdene v več nivojih, smo lahko z uporabo XPath zanesljivo identificirali vse relevantne oznake za vsak posamezen recept. Te oznake vključujejo podatke o vrsti jedi (glavna jed, predjed ali sladica), kulinarični tradiciji (mediteranska, azijska kuhinja), posebnih prehrambenih zahtevah (vegetarijansko, vegansko, brez glutena) ter priporočenih letnih časih za pripravo.

2.1.5 Komentarji. Pridobivanje komentarjev je bilo najlažje izvesti z regularnimi izrazi. Vsak komentar je sestavljen iz dveh ključnih delov: avtorja, ki se nahaja znotraj značke `<a>` z razredom `class='avtorMnenja'`, in vsebine komentarja v znački `<div>` z razredom `class='msgbody'`. S pomočjo regex vzorcev smo uspešno izluščili obe komponenti, ki smo ju nato združili v en sam smiseln stavek.

Za posamezen segment smo ustvarili funkcije za vstavljanje le teh v bazo. Tekom računanja vektorskih vložitev in testiranja poizvedb se je ohranilo izluščenje zgoraj opisanih informacij z nekaj modifikacijami kot so oblikovanje le teh v stavek in kombiniranje različnih kombinacij med seboj.

3 IMPLEMENTACIJA VEKTORSKIH VLOŽITEV

Vektorsko vložitev smo računali na podlagi zgrajenih segmentov. Najprej smo za vsako zgoraj opisani segment izračunali ločeno vektorsko vložitev. Ta pristop se je izkazal za neučinkovitega, saj smo dobivali nepravilne rezultate. Npr. za primer poizvedbe 'Kakšen je postopek za pripravo rolade?', smo kot najbližji segement dobili sestavine za rogljičke. Poleg tega smo preizkusili vektorsko vložitev izračunati na podlagi celotnega recepta, kjer smo zgornje segmente združili v celoto. Tudi ta pristop se je izkazal za neučinkovitega, saj že pri enostavnih poizvedbah kot npr. 'Potratna rolada' nismo dobili željenega recepta. Računanje vložitve smo zato poenostavili na računanje iz naslovov receptov, trajanja priprave in oznak.

Vse zgoraj omenjene pristope smo izračunali na modelih LaBSE in SloBERTa, distilbert-base-uncased in CroSloEngualBERT. Vsi modeli so se glede segmentacije odzvali podobno in so imeli pri prvih dveh pristopih več napak. LaBSE se je na koncu izkazal za najboljši model. Vse modele smo za računanje vekorskih vložitev, na podlagi naslovov receptov, testirali na 25 poizvedbah 'jabolčni kompot', 'Recept za jabolčni kompot'.... V vektorski prostor smo dodali 50 prvih receptov iz baze. Model SloBERTa je kot prvi najbližji recept pravilno vrnil 7/25 tetsov, CroSloEngularBERT 9/25, distilbert 9/25 in LaBSE 22/25. Kot najboljši model se je izkazal LaBSE, zato smo se pri končni implementaciji odločili zanj.

4 METRIKA PODOBNOSTI

Na vseh testiranih modelih smo za poizvedbe testirali vse metrike podobnosti kosinusna razdalja, L1 in notranji produkt. Vendar pa so bili rezultati ne glede na metriko enaki, zato smo se na koncu odločili uporabiti metriko notranjega produkta, saj so bili izračuni najhitrejši.

5 PRIMERI POIZVEDB IN NJIHOVA USPEŠNOST

Model smo testirali na rezličnih poizvedbah, s katerimi smo želeli pokriti različne tipe vprašanj. Npr. iskanje specifičnih receptov, iskanje receptov za posebne priložnosti ali praznike, iskanje receptov težavnosti, časovni zahtevnosti in iskanje receptov po glavni sestavini. Spodaj so našetete poizvedbe in njeni rezultati, prav tako pa so opisane omejitve za katere smo se odločili na podlagi končnih rezultatov:

- 'Podaj mi recepte, ki porabijo 15 min '
  - Cookiji: Opis recepta: slastni piškoti s koščki čokolade. Čas priprave: 15 min. Težavnost...
  - Muffini: Opis recepta: muffini s koščki čokolade. Čas priprave: 15 min. Težavnost...
  - Preprosti piškoti: Opis recepta: preprosti in zelo, zelo dobri. Čas priprave: 15 min. Težavnost...
- 'Podaj mi recepte z malinami'
  - Moj malinov kolač: Opis recepta: zelo dober in sočen...
  - Malinovi muffini: Opis recepta: čokoladni muffini z malinami in mandlji...
  - Poletni mufini z malinami: Opis recepta: odlični sočni mufini z malinami...
- 'Podaj mi božične recepte'
  - Božični keksi: Opis recepta: enostavni božični keksi. Čas priprave: 30 min...
  - Božični piškoti: Opis recepta: piškoti z brusničnim džemom. Čas priprave: 45 min...
  - Rožičevo pecivo: Težavnost: srednje težek. Postopek: Iz rumenjakov in sladkorja...
- 'Podaj mi recept za čokoladne ježke'
  - Čokoladni ježki za krasitev tort: Težavnost: zelo lahek...
  - Čokoladni cookiji: Težavnost: zelo lahek. Postopek: Čokolado sesekljamo...
  - Čokoladni medenjaki: Težavnost: srednje težek. Postopek: Med segrejemo...

5.1 Omejitve

Sprva smo se lotili računanja vektorskih vložitev iz celotnega recepta, kjer smo testirali poizvedbe iz drugih odsekov npr. 'Kateri recepti so zelo lahki?', 'Podaj mi recepte z jajci.', 'Kakšne so sestavine za potratno rolado?', 'Ali lahko v roladi uporabim brezglutensko moko?', 'Kakšen je postopek za pripravo biskvita za rolado?', vendar pa rezultati niso bili ustrezni.

Zaradi tovrstnih neuspešnih rezultatov smo se odločili vektorsko vložitev računati samo iz naslovov receptov, oznak in trajanja, tako so naše poizvedbe omejene na vprašanja v povezavi teh naštetih odsekov. Vseeno pa so se našli primeri znotraj teh tematik, ki niso vračali želenih rezultatov:

- 'Podaj mi postopek za arašidovčke'
  - Cookiji z arašidi: Opis recepta: super piškotki s čokolado in arašidi...
  - Arašidovi blondiji: Opis recepta: arašidovo pecivo prelito s čokolado...
  - Oreo riž: Opis recepta: mlečni riž z Oreo piškoti. Čas priprave: 45 min. Težavnost
- 'Jabolčni kompot'
  - Kruh z jabolki: Težavnost: zelo lahek. Postopek: Kruh narežemo ali nalomimo...
  - Jabolčni džem: Težavnost: zelo lahek. Postopek: Jabolka operemo, olupimo...
  - Zvrnjena jabolčna krema: Težavnost: srednje težek. Postopek...
- 'Vrni mi recept za sacher torto'
  - Skutina torta - Cheesecake. Težavnost: lahek. Postopek: Piškote zdrobimo z valjarjem...
  - Sacher muffini: Opis recepta: čokoladni muffini z marmelado. Čas priprave: 1 ura 30 min...
  - Cheesecake pecivo iz pirine moke: Opis recepta: klasični Cheesecake malo drugače...