Klasifikacija recepta po državi
Projektni prijedlog

Stjepan Požgaj, Mihovil Stručić, Luka Tomić

23. Travnja 2020.

Uvodni opis problema

Problem koji ćemo pokušavati riješiti u ovom projektu je klasifikacija recepta po državi pomoću liste sastojaka koji se koriste u njemu. Automatska klasifikacija kuhinje iz koje recept dolazi omogućuje stranicama sa receptima da automatski klasificiraju novo dodani recept i dodaju određenu vrstu mjerenja sličnosti između recepata pa tako mogu nekom korisniku predlagati neke druge recepte. Također aplikacije za dostavu i narudžbu hrane omogućuju korisniku da filtrira restorane po tipu kuhinje (talijanska, meksička, domaća…). Restorani se mogu sami deklarirati da imaju jela koje pripadaju nekoj kuhinji, ali sustav preporuka aplikacije može i sam procjeniti koliko je jelovnik nekog restorana "meksički" ili "talijanski" i prema tome ga rangirati među ostalim restoranima. Skup podataka kojeg ćemo koristiti objavljen je u sklopu Kaggle natjecanja *What's Cooking?*[1] (podaci za to natjecanje su preuzeti s Yummly[2]) koje je održano prije nekoliko godina u kojem je sudjelovalo preko 1000 timova stoga će biti zanimljivo vidjeti koliko ćemo mi biti uspješni u usporedbi s njima. Skup podataka se sastoji od 39 774 recepata iz 20 država i u receptima su 6 532 različita sastojaka(iako neki sastojci su jako slični, stoga je ovaj broj možda i prevelik).

Cilj i hipoteze istraživanja problema

Iskoristiti ćemo klasifikacijske tehnike strojnog učenja kako bi istražili koje kombinacije sastojaka dovode do različitih kuhinja i pokušati predvidjeti iz koje države dolazi recept pomoću liste njegovih sastojaka sa što većom točnošću. Konkretno, sastojke ćemo koristiti kao značajke (feature), recepte koje treba klasificirati, a države kao moguće labele klasifikacije. Pri tome će jedan recept moći pripadati samo jednoj državi. Očekujemo da će klasifikatori imati više problema ispravno klasificirati recepte iz kuhinja koje su možda slične s nekom drugom kuhinjom, a manje problem kod recepata iz kuhinja koje su značajno različite od svih ostalih kuhinja. Pokušati ćemo interpretirati modele koje neki od algoritama nauče kako bi vidjeli koji su to točno sastojci koji sačinjavaju neku kuhinju. Naravno, ultimativni cilj je biti što bolji na Kaggleovom leaderboardu.

Pregled dosadašnjih istraživanja

Postoji nekoliko istraživanja koje se bave klasifikacijom recepta po državama koristeći sastojke. [3] istražuje povezanost između države i sastojaka koristeći klasifikacijske tehnike poput asocijativne klasifikacije i SVM koristeći podatke sa stranice food.com i tako pokušava otkriti koji su esencijalni sastojci recepata neke kuhinje i automatski klasificira kuhinju iz koje recept dolazi. Sličnost između kuhinja je promatrana pomoću matrice konfuzije koju su generirali korišteni klasfikatori. Sličan put prati i [4] koristeći pritom Naivni Bayes, Multinomialnu logističku regresiju, Random Forest i SVM.

[5] koristi iste podatke kao što ćemo i mi korisiti te ne stavljaju naglasak na same algoritme i modele strojnog učenja koji su korišteni nego na utjecaj preprocesiranja podataka na točnost klasifikacije. Pokušavaju što više očistiti podatke, pronaći sastojke koji su po samom tekstu kako su navedeni različiti, a zapravo su jednaki. Tako smanjuju broj sastojaka sa početnih 6714 na 4793. Za klasifikacijske algoritme koriste Naivni Bayes, neuralnu mrežu i SVM te ostvaruju točnost od 0.809 pomoću SVM algoritma. Taj rad će nama služiti za usporedbu koliko smo uspješni u rješavanju problema.

Sva prethodno navedena istraživanja podatke o sastojcima za neki recept pretvaraju u format prihvatljiv algoritmima za strojno učenje koristeći bag-of-words pristup ili TF-IDF vektorizaciju teksta sastojaka.

Materijali, metodologija i plan istraživanja

Prilikom učitavanja skupa podataka smo iz recepata maknuli interpunkcijske znakove, brojeve i ostatak suvišnog teksta da bismo identificirali iste sastojke zapisane na drugačiji način. Možemo se poslužiti metodom analognoj onoj za stvaranje TF-IDF (term frequency - inverse document frequency) vektora za svaki dokument. U našem slučaju će dokumenti biti zamijenjeni receptima, a termini sastojcima. Svaki se svaki sastojak u svakom receptu spominje najviše jednom, no njihov značaj možemo skalirati ovisno ukupnom broju pojavljivanja u receptima. Osim ovih podataka svakom receptu možemo pridružiti i primjerice broj sastojaka.
Tako dobivene podatke ćemo transformirati metodama za smanjenje dimezionalnosti poput PCA (principal component analysis) i t-SNE (t-stochastic neighbor embedding).

Metode i pristupi koje ćemo isprobati:

- k-means - podjela recepata u grupe temeljena na odabranoj mjeri udaljenost i pridruživanje svake grupe nečijoj kuhinji tijekom faze treniranja, te određivanje najbliže grupe svakom od recepata u test fazi
- Naivi Bayesov klasifikator za kategoričke varijable
- RandomForrest
- Stroj potpornih vektora
- XGBoost -imlementirana u istoimenom Python libraryju

Sve navedeno pisat ćemo u Pythonu koristeći Jupyter i po potrebi Colab bilježnice te biblioteke poput sklearn[5], matplotlib[6] i scipy[7].

Kvalitetu modela na skupu za treniranje određivat ćemo unakrsnom validacijom računajući prosječnu greška svih modela, gdje točnost modela, naravno, udio recepata čija

je kuhinja ispravno predviđena. Velika je prednost što se radi o problemu s Kaggleovog natjecanja pa ćemo jednostavno moći usporediti kvalitetu našeg rješenja s drugim rješenjima.

Očekivani rezultati predloženog projekta

Dakle, kao što smo napomenuli, očekujemo da će naši rezultati biti sumjerljivi sa dobrim rezultatima s Kaggle natjecanja. Također nam je cilj što bolje upoznati metode koje smo spomenuli i što bolje ih primijeniti na ovaj problem.

Literatura

[1] https://www.kaggle.com/c/whats-cooking

[2] https://www.yummly.com/

[3] Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang. 2014. Automatic recipe cuisine classification by ingredients. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (*UbiComp '14 Adjunct*). Association for Computing Machinery, New York, NY, USA, 565–570. DOI: https://doi.org/10.1145/2638728.2641335

[4] S. Jayaraman, T. Choudhury and P. Kumar, "Analysis of classification models based on cuisine prediction using machine learning," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, 2017, pp. 1485-1490, doi: 10.1109/SmartTechCon.2017.8358611.

[5] https://scikit-learn.org/

[6] https://matplotlib.org/

[7] https://www.scipy.org/

[8] https://web.math.pmf.unizg.hr/nastava/su