Razpoznavanje avtorstva na Enron korpusu

Simon Janežič

Univerza v Ljubljani,
Fakulteta za računalništvo in informatiko,
Večna pot 113, Slovenija
sj8495@student.uni-lj.si

Povzetek V članku predstavimo nalogo razpoznavanja avtorstva ter podamo primer uporabe. Nato povzamemo zgodovino tega področja ter nekatere metode razpoznavanja avtorstva, ki temeljijo na n-gramih in strojnem učenju. Opisane metode tudi preizkusimo na podmnožici Enron korpusa, ki je bila uporabljena pri nalogi razpoznavanja avtorstva PAN leta 2011. Svoje rezultate tudi primerjamo z rezultati omenjene naloge.

1 Uvod

Z razpoznavanjem avtorstva želimo napovedati avtorja nekega besedila z neznanim avtorjem ali pa analizirati besedila z vprašljivim avtorstvem. Razpoznavanje avtorstva se v pravnih vodah uporablja kot del lingvistične forenzike. Primer take uporabe so vprašanja kot so ali je nekdo, ki naj bi storil sammor res napisal svoje poslovilno pismo ali pa identifikacija avtorja grozilnega sporočila[1]. Uporabnost pa sega tudi v literarne vode. Primer bi bilo dobro poznano vprašljivo avtorstvo Williama Shakespeara[3].

V tem članku se lotimo naloge iz razpoznavanja avtorstva, ki jo je v letu 2011 razpisal PAN[1]. Naloga je razpoznavanje avtorstva v Enron korpusu[2]. Ta je sestavljen iz elektronskih pošt zaposlenih pri Enron. Podatkovna množica naloge vsebuje podmnožico dejanskega korpusa. Množica je razdeljena na učne, validacijske in testne podatke (elektronske pošte). Rezultati nad testnimi podatki so bili uporabljeni za razvrstitev udeležencev te naloge, torej jih bomo za primerjavo uporabili tudi mi. Cilj je seveda napovedati avtorje elektronskih pošt na podlagi označenih pošt iz učne in validacijske množice. Pri tem možnih omemb osebnih imen v samih elektronskih poštah ne moremo upoštevati, saj so bila ta s strani PAN v predobdelavi odstranjena. Problema se lotimo kot klasifikacijski problem strojnega učenja, kjer vsak možen avtor predstavlja svoj razred. Z besedil pozkušamo razbrati značilke, ki so pomembne za določitev stila pisanja avtorja. Problem je zanimiv, saj se ukvarjamo z elektronskimi poštami, torej neformalnimi besedili z možnimi slovničnimi napakami in podobno, kar bi nam pri določitvi avtorstva lahko tudi pomagalo.

[1] http://pan.webis.de
[2] https://www.cs.cmu.edu/~./enron/

2 Obstoječo delo

Prvi poizkusi kvantificiranja stila pisanja segajo v 19. stoletje s študijo Shakespearovih iger[6]. V tej študiji je Mendenhall razpoznavanje avtorstva osnoval na zelo preprostih značilkah, kot so štetje dolžin povedi in besed. Naslednjo prelomnico predstavlja študija političnih esejev z imenom 'The Federalist Papers'[7]. Metoda Mostellera in Wallace-a je v tej študiji osnovana na Bayesovi statistični analizi ferkvenc majhne podmnožice pogostih besed (npr. 'and', 'the', itd.). Dosegla sta opazno diskriminacijo med različnimi avtorji teh besedil. To je predstavljalo začetek netradicionalnih metod razpoznavanja avtorstva, ki ne temeljijo na človeškem strokovnjaku. Od takrat do poznih devetdesetih let so poizkušali definirati različne značilke, ki bi znale kvantificirati slog pisanja. Temu področju raziskovanja rečemo stilometrija. Primeri takih značilk so dolžine besed, ferkvence besed, ferkvence znakov in funkcije bogastva besednjaka. Primer zadnje je število besed, ki se v besedilu pojavljajo zgolj enkrat[4]. Pri tej in podobnih funckijah imamo problem z različnimi dolžinami besedil in nelinarno rastjo besednjaka z večanjem dolžine besedila, kar pomeni, da so te funkcije same po sebi nezanesljive.

Dosegljivost ogromne količine besedil na spletu je v zadnjih 15 letih vodila do razvoja učinkovitih tehnik ekstrakcije informacij, strojnega učenje ter obdelave naravnega jezika. Razpoznavanje avtorstva je z vidika strojnega učenja definirano kot večrazredni tekstovni klasifikacijski problem. Po zgledu klasifikacije teme v tekstu se tudi na tem področju uporablja predstavitev teksta kot vektor pojavitve besed (vreča besed)[8]. Obstaja pa velika razlika med klasifikacijo teme in stila pisanja oziroma avtorstva. V primeru razpoznavanja avtorstva se ravno pojavitve funkcijskih besed (npr. 'and', 'the', itd.) izkažejo kot najkoristnejše značilke za določanje avtorstva[2]. Te besede so v klasifikaciji teme besedila ponavadi ignorirane, saj ne predstavljajo koristne semantične informacije. Obstaja več člankov na temo izbire teh besed. Preprosta in precej uspešna metoda je, da izberemo najpogostejše besede, ki se pojavljajo v korpusu. Eden izmed parametrov nato seveda postane koliko besed vzamemo. V starejših člankih so uporabili najpogostejših 100 besed, v novejših pa tudi več tisoč[9]. To so jim omogočali algoritmi strojnega učenja, ki lahko delajo z visokim številom dimenzij (npr. SVM). Problem predstavitve besedila z vrečo besed je seveda izgubljena kontekstna informacija. Zato je bila za značilke predlagana uporaba ferkvenc več zaporednih besed[5], kar pa se ni vedno bolje odrezalo kot uporaba posameznih besed. Z uporabo več zaporednih besed se namreč poveča tudi dimenzija problema. Še en pristop pa je uporaba ferkvenc znakov, oziroma več zaporednih znakov kot značilke[4].

Poleg omenjenih pristopov so bile razvite tudi metode, ki koristijo razne sintaktične in semantične lastonsti besedila. Mi jih tu ne bomo omenjali, saj v našem članku preizkušamo le zgoraj omenjene pristope. Dober pregled metod je podal Stamatatos v svojem članku[10].

3 Metode

Problema se lotujemo s strojnim učenjem. Bistveno je torej pridobivanje značilk in uporabljeni algoritmi strojnega učenja.

3.1 Pridobivanje značilk

Če hočemo rapoznavanje avtorstva elektronskih pošt obravnavati kot klasifikacijski problem strojnega učenja je ključno, da iz teksta pridobimo dobre značilke, ki opisujejo slog pisanja v danem mailu. Po zgledu metod opisanih v obstoječih delih smo besedilo opisali kot vektor pojavitve najpogostejših besed v celotnem korpusu. Prav tako smo poizkušali vključiti tudi vektorje pojavitve najpogostejših kombinacij besed različne dolžine (n-grami). Eksperimentirali smo z različnimi števili najpogostejših n-gramov, različnimi metodami preobdelve teksta in ostalimi možnimi parametri ter nakoncu izbrali dobre vrednosti na podlagi rezultatov nad validacijsko množico z uporabo logistične regresije.

Naj se najprej lotimo predobdelave teksta. V primerjavi s problemom klasifikacije teme teksta se je pri klasifikaciji avtorja v našem primeru izkazalo, da je manjša količina predobdelave boljša izbira, saj očitno s pretirano obdelavo hitro lahko zanemarimo določene karakteristike avtorja. Najboljše rezultate na validacijski množici smo dobili, če smo v tekstu ohranili vse znake ter besedila nismo pretvorili v manjšo začetnico. To je nekako smiselno, saj nam v neformalnem besedilu uporaba posebnih znakov in velikih začetnic lahko pove veliko o avtorju. Kot je že bilo omenjeno v obstoječih delih tudi funkcijskih besed nismo odstranili, saj so v tej domeni koristne. Edina predobdelava, ki nam je izboljšala rezultate pa je bilo korenjenje (angl. stemming).

Najboljše rezultate na validacijski množci je dosegla kombinacija n-gramov, kjer smo za opis maila uporabili 500 v korpusu najpogostejših 1-gramov, 500 v korpusu najpogostejših 2-gramov in 250 v korpusu najpogostejših 3-gramov. Naj tu omenimo še, da smo eksprementirali tudi z različnimi kombinacijami znakovnih n-gramov, kot tudi kombinacijo besednih in znakovnih a nam je uporaba teh poslabšala rezultate. Poleg omenjenega bi bilo zanimivo omeniti še, da smo vektorje pojavitve n-gramov poizkušali tudi normalizirati, torej narediti vektor invarianten na dolžino besedila, a nam je tudi to le poslabšalo rezultate. Očitno nam tudi ta varianca koristi pri klasifikaciji avtorja.

Po zgledu starejših pristopov [6] smo poizkušali uvesti tudi nekaj agregiranih atributov, kot so povprečna dolžina besede v tekstu in dolžina teksta, a nam te niso izboljšale rezultatov.

Kljub temu, da obstoječa dela poročajo o uspehih z uporabo najpogostejših n-gramov smo z bolj klasičnim pristopom za klasifikacijo teksta poizkušali v nabor izbranih n-gramov dodati še specifične besede določenih avtorjev. Tega smo se lotili tako, da smo maile najprej združili po avtorjih. Nad dobljenimi besedili smo nato izvedli izračun TF-IDF uteži besed. Ker smo pri izračunu za en dokument uporabili kar vse elektronske pošte enega avtorja bi temu lahko rekli tudi TF-IAF (term frequency-inverse author frequency). Nato smo za vsakega avtorja vzeli 100 najpomembnejših besed (število izbrano s pomočjo validacijske

množice). Ker tudi tu nismo filtrirali teksta je to vključevalo tudi možne posebne znake. Smo pa tu vseeno odstranili funkcijske besede in dobili boljše rezultate s pretvorbo teksta v male črke. Dobljen nabor besed smo dodali k atributom. Možne duplikate smo seveda izbrisali. Ta pristop je uspešno izboljšal rezultate nad validacijsko množico. Poizkusili smo tudi s specifičnimi 2-grami avtorjev, a so bili rezultati slabši.

3.2 Strojno učenje in evalvacija

S pridobljenimi učnimi primeri smo nato lahko uporabili algoritme strojnega učenja. Uporabili smo Naivnega Bayesa, Logistično regresijo in SVM. Za nastavljanje parametrov smo prav tako kot prej uporabili validacijsko množico. Prav tako, kot v dejanskem tekmovanju/nalogi smo tudi mi za metriko uspešnosti uporabili mikro in makro povprečeno F1-mero. Najbolje se je odrezala logistična regresija. Moramo omeniti, da nam zaradi časovne kompleksnosti za SVM algoritem najbrž ni uspelo najti dobrih parametrov.

4 Rezultati

Preostal nam je še končni test nad dano testno množico naloge. Model, ki smo ga testirali v tej fazi je tokrat pri učenju in pridobivanju značilk upošteval tudi podatke v validacijski množici. Zaradi primerljivosti rezultatov z rezultati tekmovanja smo tudi tu za metriko uspešnosti modela uporabili mikro in makro povprečeno F1-mero. Pridobljene rezultate in njihovo primerjavo z rezultati tekmovanja smo prikazali v tabeli 1. Naj omenimo, da smo tu reševali in primerjali le podnalogo tekmovanja, kjer so v testni množici le besedila, katerih avtorje poznamo z učne/validacijske množice. Torej ni razreda neznananega avtorja, kar je bila ločena podnaloga tekmovanja.

Tabela 1. Rezultati razpoznavanja avtorstva na Enron korpusu

model	mikro povp. F1-mera	makro povp. F1-mera
večinski klasifikator	0.07	0.002
mediana tekmovanja	0.46	0.35
naša logistična regresija	0.58	0.43
najboljši r. tekmovanja	0.66	0.52

Kot trivialni model smo v primerjavo vključili tudi večinski klasifikator. Ker imamo opravka z velikim številom razredov, opazimo zelo nizko točnost, ki jo ta doseže. Opazimo tudi, da smo dosegli boljše rezultate kot je bila mediana tekmovanja. Uporabili smo mediano in ne povprečja saj je bilo na dnu nekaj rezultatov, ki so bili zelo blizu ničle, zato teh nismo hoteli upoštevati. Ni pa nam uspelo doseči rezultata zmagovalca tekmovanja.

5 Zaključek

Predstavili smo nekatere metode za razpoznavanje avtorstva ter jih poizkusili na Enron korpusu. Dosegli smo precej dobre rezultate, vendar še vedno slabše od zmagovalca PAN tekovanja leta 2011. Da bi se približali rezultatom zmagovalca bi morali poizkusiti med pridobivanjem značilk vključiti tudi rezultate metod, ki nam povedo več o sintaktičnih in semantičnih lastnostih besedila (npr. oblikoslovno označevanje).

Literatura

1. Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.
2. Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 2005.
3. D Hugh Craig and Arthur F Kinney. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press, 2009.
4. Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.
5. Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.
6. Thomas Corwin Mendenhall. The characteristic curves of composition. *Science*, pages 237–249, 1887.
7. Frederick Mosteller and David Wallace. Inference and disputed authorship: The federalist. 1964.
8. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
9. Efstathios Stamatatos. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05):823–838, 2006.
10. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.