PODATKOVNO RUDARJENJE

SEMINARSKA NALOGA

Avtorji: Gal Mrvar, Gregor Novak, Avguštin Kastelic, Jure Mohar

Študijsko leto: 2018/2019

Predavatelj: doc. dr. Tomaž Curk

1. KAZALO VSEBINE

No table of contents entries found.

Vsebina:

Vir podatkov opis tega itd.

Način obdelave podatkov in orodja

Ugotovitve grafi itd.

Ocenitev klasifikatorja (test)

Zaključek

Klasifikator:

## Namen klasifikacije

S klasifikacijo smo nameravali analizirati podatke (twitte ter članke) in jih deliti na tri kategorije:

* Pozitivno
* Negativno
* Nevtralno

Te ugotovitve pa smo potem nameravali razporediti na graf v istem časovnem obdobju kot ceno oz. volumen kriptovalute BTC.

## Izbira klasifikatorja

V prvem poskusu smo izbrali knjižnico za procesiranje besedila textBlob, ki se kar pogosto uporablja pri sentimentalni analizi teksta. Seveda pa smo pri tem naleteli na kar nekaj težav: Uspešnost klasifikatorja je bila slaba, manjša kot večinski klasifikator, kar pomeni, da je bila klasifikacija praktično neuporabna.

Naslednja možnost je bila uporabiti naivni bayes klasifikator, prav tako od textBloba, ki uporablja nltk (natural language toolkit) in ima za učno množico že zbrane podatke ocen oz. mnenj filmov. Tukaj smo se odločili besedila razbiti na posamezne besede in izključiti ponovitve besed, linke ter vse za klasifkator »neuporabne« besede s tem smo upali, da bomo izboljšali klasifikator. Vendar kljub temu nismo uspeli dobiti boljše točnosti od večinskega klasifikatorja.

Nato smo poizkušali ustvariti svoje učne in testne podatke s katerimi bi učili klasifikator vendar je bilo to praktično nemogoče saj smo imeli preprosto preveč podatkov, da bi lahko vsem ročno določili kateri so pozitivni, negativni ali nevtralni. Zato smo idejo relativno hitro opustili.

Kot zadnjo alternativo smo ponovno izbrali prvi klasifikator (kateri je bil večinskemu še najbljižji) in odstraniti linke ter nekoristne informacije. Tako smo dobili klasifikator, ki je bil za 2 odstotka boljši od večinskega klasifikatorja.

## Testiranje klasifikatorja

Klasifikator smo testirali tako, da smo kreirali svoje testne podatke iz naključno izbranih twittov, teh je bilo 115. Te podatke smo klasificirali ročno ter nato pognali klasifikator in preverili rezultate.

Uspešnost:

* Večinski klasfikator: 52%
* Naš klasifikator: 54%

Ugotovitve:

V kolikor se poglobimo v rezultate se nam zdi, da je bil rezultat kar pričakovan. V twittih se uporablja ogromno besednih zvez katere je praktično nemogoče klasificirati prav tako je ogromno sarkastičnih povedi itd. Poleg tega smo delili na 3 kategorije kar pomeni da 54% točnost niti ni tako slaba.

Naivni Bayes s pomočjo nltk

Poskusili smo zgraditi tudi svoj Naivni Bayesov klasifikator. Tudi tokrat smo iz besedila najprej odstranili irelevantne besede in tekst, kot so URL povezave, imena uporabnikov ipd. Ker je podatkov, ki smo jih uporabljali preveč, bi, kot že zgoraj omenjeno, trajalo predolgo, da bi vse ročno sentimentalno ocenili (v tem primeru pa potem niti ne bi potrebovali klasifikatorja, saj bi že imeli ocenjene vse podatke) in nekaj podatkov uporabili kot učno množico, nekaj pa kot testno. Kljub temu smo ocenili nekaj tweetov z namenom, da jih uporabimo za testiranje klasifikatorja. Še vedno pa bi nam vzelo preveč časa, da bi ocenili dovolj podatkov in s tem pridobili smiselno veliko učno množico. Zato smo za učne podatke uporabili prosto dostopne podatke projekta Sentiment140, kjer so za učno množico uporabili 1.600.000 tweetov, ocenjene z 0 – negativno ali 4 – pozitivno.

Za gradnjo klasifikatorja smo uporabili knjižnico nltk, s katero je narejen tudi Naivni Bayes v knjižnici TextBlob. Izkazalo se je, da v primeru velike učne množice (npr. vseh 1.600.000 tweetov) traja učenje klasifikatorja predolgo za praktično uporabo, prav tako pa je pozneje količina časa, ki ga tak klasifikator potrebuje za klasifikacijo podatkov zalo velika. Zaradi tega smo bili primorani za učne podatke uporabiti le delež podatkov iz učne množice od Sentiment140.

Klasifikator smo zgradili z uporabo približno 4400 tweetov od vseh 1.600.000, že v tem primeru pa sta bili gradnja klasifikatorja in njegova uporaba nekoliko zamudni. Iz učnih podatkov smo ustvarili seznam vseh besed, ki se pojavijo v izbranih podatkih, nato pa se je za vsak tweet v učni množici določilo, katere besede od vseh prisotnih besed vsebuje. Te vrednosti vsakega tweeta, v kombinaciji s sentimentalno oceno posameznega tweeta, je klasifikator uporabil za učenje. Klasifikator pri klasifikaciji besedila v obliki odstotkov določi v kolikšni meri je besedilo sentimentalno pozitivno in v kolikšni meri je besedilo sentimentalno negativno. Ti dve vrednosti se seštejeta v 100% oz. 1. Nevtralno sentimentalno vrednost besedila pa lahko določimo tako, da privzamemo, da je sentimentalnost besedila nevtralna, če se vrednosti pozitivno/negativno razlikujeta za dovolj majhen delež, npr. 0,15 oz. 15%.

Vendar pa smo že z ročnim pregledom pridobljene analize brez testiranja točnosti uvideli, da po našem mnenju klasifikator ni dovolj kvaliteten za uporabo, mogoče celo slabši od prejšnjih poskusov ocenjevanja. Sklepali smo, da bi klasifikator lahko izboljšali z večjo učno množico, saj bi ta vsebovala več primerov različno ocenjenih tweetov. To pa ni bila sprejemljiva možnost, saj je bil že trenuten klasifikator dokaj počasen, več učnih podatkov pa bi to lastnost le še poslabšalo. Naše mnenje je tudi, da bi največjo uspešnost klasifikatorja dosegli z učnimi podatki, ki bi vsebovali tweete s področja kriptovalut, saj bi se tako lahko klasifikator specializiral za ocenjevanje besedila, ki vsebuje izraze in ostale vzorce, značilne za tweete o kriptovalutah. Takšno učno množico bi lahko pridobili, če bi ocenili dovolj veliko število tweetov iz naših podatkov, preostale tweete pa bi nato lahko ocenili z našim klasifikatorjem. To pa nam je, kot že opisano, zaradi časa predstavljalo preveliko oviro.