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

Klasifikatior:

## 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. volument kriptovalute BTC.

## Izbira klasifikatorja

V prvem poskusu smo izbrali knjižnjico 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.