POROČLO 1. NALOGE PRI PREDMETU INTELIGENTNI SISTEMI
TEXT MINING

UVOD

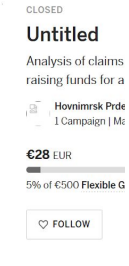
Pri tej nalogi se bom ukvarjal s tekstovnim rudarjenjem besedil. Tekstovno rudarjenje se uporablja za analizo besedil s pomočjo algoritmov umetne inteligence. Pomaga nam pri razumevanju in razvrščanju besedil v skupine z podobno vsebino, npr. pri spoznavanju tekstovnega programiranja sem ugotavljal avtorja besedila s pomočjo besedil ki so imeli znanega avtorja in gledal pod katerega avtorja mi razporedi besedilo brez avtorja.

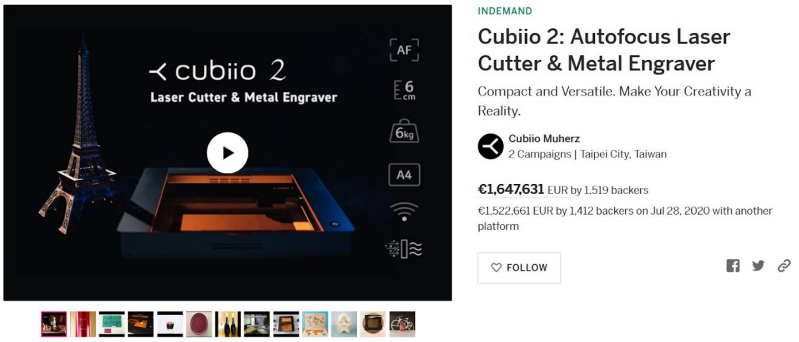
Za moj primer tekstovnega rudarjenja sem si izbral industrijo “Crowdfunding”, kjer podjetniki zbirajo zagonska sredstva za vzpostavitev podjetja z enim produktom za vsakdanjo rabo s pomočjo velikega števila ljudi(investitorjev). Na strani sem analiziral komentarje za nekaj uspešnih in nekaj neuspešnih izdelkov in poskušal izvleči uporabne informacije iz njih.

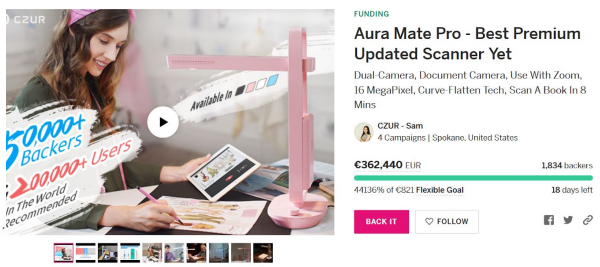
1. KORAK

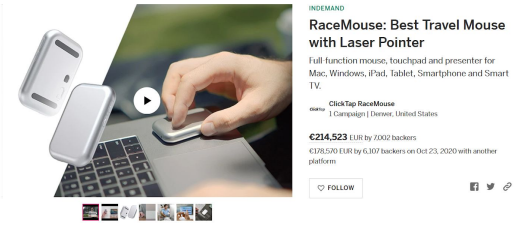
Using your favorite search engine, locate a web site or discussion forum on the Internet where people have posted complaints, criticisms or pleas for help regarding a company or an industry (e.g. airlines, utility companies, insurance companies, etc.).

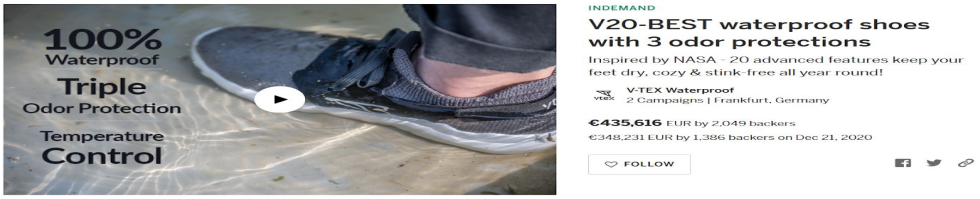
Platforma “crowdfunding” , ki sem jo izbral je bila Indiegogo, kjer sem si izbral 8 različnih produktov in jih primerjal med sabo, za nekatere sem vedel da se je izkazalo za prevare za druge pa nisem vedel nič. Ali bom lahko določil kateri izdelki so legitimni in kateri ne s pomočjo tekstovnega rudarjenja?











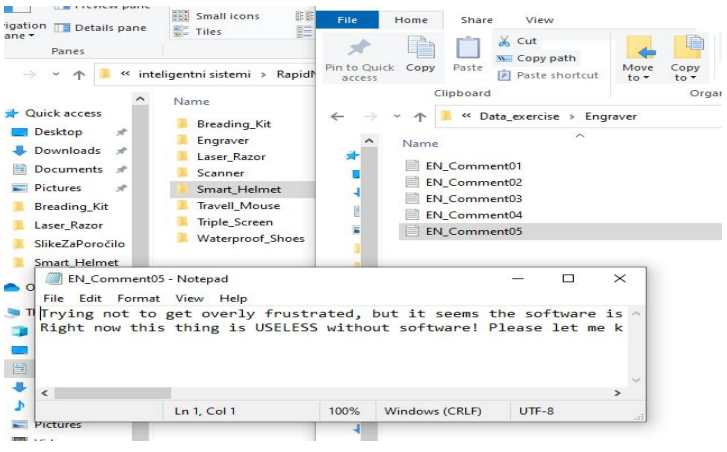
Produkti, ki sem jih izbral so bili:	Kodno ime
Laser engraver	EN
Laserska britev	LR
Prenosni scanner	SC
Pametna čelada	SH
Vodoodporni čevlji	WS
Prenosna miška	TM
Prenosni trojni zaslon	TS
Dihalka za potapljanje	BK

Za dihalko za potapljanje, lasersko britev in pametno čelado sem vedel da gre za prevaro ostalih pa nisem poznal.

2. KORAK

Copy and paste at least ten of these posts or comments into a text editor, saving each one as its own text document with a unique name.

Iz vsakege od zgoraj naštetih kampanih sem vzel 5 komentarjev, ki sem jih vključil v moj algoritem.
Izmed komentarjev sem izbral tiste, ki so bili daljši in posledično bolj razčlenjeni in argumentirani. Drugič moja izbira sestoji tudi iz novih nedavnih komentarjev in komentarjev iz samega začetka kampanije ni, kar lahko pripomore k pristranskim rezultatom, daj nimamo podatka iz mišljenja “investitorjev” na začetku.

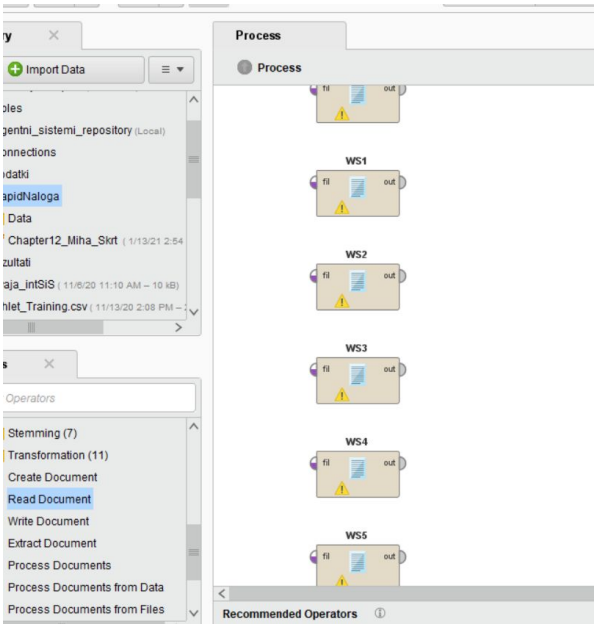


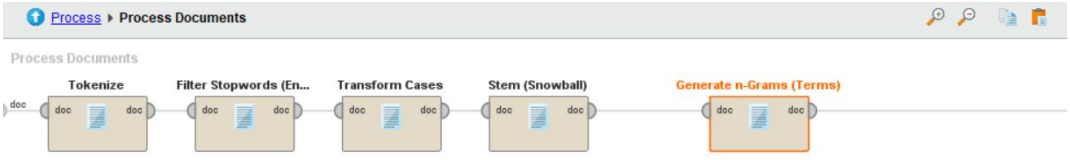
3. KORAK

Open a new, blank process in RapidMiner, and using the Read Documents operator, connect to each of your ten (or more) text documents containing the customer complaints you found.

Process these documents in RapidMiner. Be sure you tokenize and use other handlers in your sub-process as you deem appropriate/necessary. Experiment with grams and stems.

Komentarje sem nato vnesel v RapidMiner in jih obdelal. Za uspešno rudarjenje moramo komentarje pretvoriti v računalniku prijazen zapis, zato sem jih razbili na posamezne besede in analiziral kolikokrat se kakšna pojavi.
Za obdelavo besedila sem uporabil še operatorje stemming cse transform, ker pa internetni komentarji niso zapisani v pravilni angleščini in je veliko krajšank in tipkarskih napak, sem moral polek StopWord uporabiti še StopWord(Dictionary) in izločiti iz podatkov same črke in podobno in še zadnji: za n-gram operator sem določil naj mi poišče skupke besednih zvez in 2 besed, saj je bil jezik v komentarjih že tako enostaven. Na koncu sem opazil da so se besedne zveze pojavljale po 1x ali 2x in so imele same besede večjo težo.

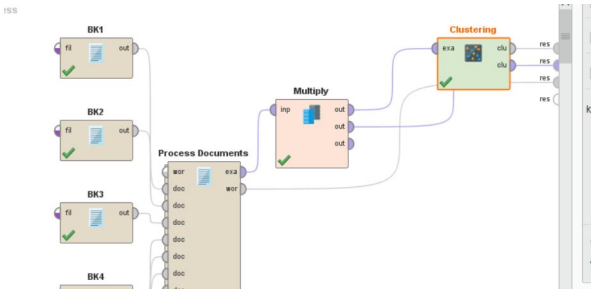




4. KORAK

Use a k-Means cluster to group your documents into two, three or more clusters. Output your word list as well.

Za obdelavo sem na koncu uporabil k-Means algoritem in skušal grupirati podatke v različne skupine ter poskušal ugotoviti ali dobim kaj pametne informacije iz vsega.



5. KORAK

Report the following:

a. Based on your word list, what seem to be the most common complaints or issues in your documents? Why do you think that is? What evidence can you give to support your claim?

Glede na besedno listo je veliko govora o naročilu produktov in o dostavi le teh ter denarju. besede kot so : *get, monei, product, work, delivery, order …* so med najbolj pogostimi in se pojavljajo v več komentarjih.
Vse kaže da projekti na strani pogosto končajo nerealizirani in stranke(investitorji) na suhem, ali pa stvar ne dela, ko jo dobijo.
Za dokaze bi se navezal kar na mojo začetno trdite da vem za produkte ki so se izkazali za prevare in so tudi navedeni zgoraj (v prilogah dodajam tudi link).

b. Based on your word list, are there some terms or phrases that show up in all, or at least most of your documents? Why do you think these are so common?

Polek zgoraj omenjenih stvari je bila pogosta tudi beseda Indigogo, kar je seveda smiselno saj je to ime platforme na kateri se vse dogaja, ostale skupne besede, ki se pojavijo v več komentarjih, imajo pa skupen pomen dostave obljubljenega izdelka kar je tudi glavni namen strani in projektov na strani, se pravi to je vsem skupno ne glede na use.

c. Based on your clusters, what groups did you get? What are the common themes in each of your clusters? Is this surprising? Why or why not?

Pri grupiranju komentarjev sem poskušal komentarje grupirati na več načinov.
Najprej sem se spraševal, če lahko ločim goljufive kampanije od pravih, zato sem uporabil **2 grči.** Presenečen sem bil ker se to ni zgodilo, res da so bili v eni skupini komentarji bolj razočaranih strank in strank s težavami in v drugem vse drugo od skeptičnih, zadovoljnih in komentarjev, ki so izražali slabo komunikacijo. Rečemo lahko da mi je algoritem filtriral

komentarje, ki so imeli resne in konkretne težave z produktom in ostale, Kar bi lahko uporabil za zaznavanje problematičnih kampanj.
Naslednjič sem grupiral podatke v **3 gruče** ampak še pred pričetkom grupiranja nisem vedel kaj pričakovati in na koncu tudi v podatkih nisem našel nobenega vzorca.

Ker sem uporabil 8 različnih kampanj v mojih podatkih in je bila vsak enako zastopana sem poskusil podatke grupirati še na **8 gruč** in pričakoval da bo algoritem ugotovil kateri komentarji pripadajo določenemu izdelku. Po izvedbi algoritma pa sem ugotovil da temu ni bilo tako, sprva v gručah nisem videl nobenih povezav, ob boljši preučitvi pa se mi je zdelo da se je algoritem še najbolje odrezal in grupiral podobne komentarje skupaj. Očitno je bilo v komentarjih veliko različnih tem(seveda vsak človek ima svoje mnenje in to se je tudi pokazalo.) Lahko bi upošteval tudi avtorje komentarjev in pogledal ali so bili grupirani skupaj vendar tega podatka nisem zabeleil in spremljal.

Za analizo vseh načinov grupiranja sem uporabil excel tabelo in si označil kaj je tema posameznega komentarja, in v katero gručo je bil dodeljen pri posameznem grupiranju. Kakšne vrste gruč sem dobil in kakšne informacije mi prinašajo pa v naslednji točki.

d. How might a customer service manager use your model to address the common concerns or issues you found?

SKUPNA TEMA POSAMEZNE SKUPINE	MOŽNOST UPORABE INFORMACIJE
SKEPTIKI PRODUKTA (Cluster 0)	Lahko bi uporabili za ustvarjanje boljšega opisa izdelka in bolje razložili delovanje. Je tudi prva zastavica da z izdelkom ni vse vredu.
UPORABNE INFORMACIJE (Cluster 1)	Ti komentarji so dobri za vključitev na začetno stran saj dobro infornirajo kupca(investitorja) o izdelku, lahko se uporabijo tudi za odgovore v sekciji FQA
PROBLEMI KOMUNIKACIJE (Cluster 2)	Ko se komentar znajde v tej gruči je lahko triger ki ponudniku izdelka pošlje mail naj komunicira z strankami. Je tudi močnejša zastavica da je produkt morda goljufiv.
FALSE ADVERTISING (Cluster 3)	Če se komentar znajde v tej kategoriji je zelo verjetno, da se stvar lažno oglašuje in je izdelek veliko slabši kot je bilo obljubljeno z veliko napakami. Akcija ki jo lahko ob tem sprožimo je sankcioniranje ponudnika za lažne oglase.
KAKO DELUJE? (Cluster 4)	Ti komentarji govorijo o strankah, ki ne vejo točno na kakšen princip izdelek deluje in sprašujejo o principu delovanja in o načinu uporabe. Dobra uporabna gruča za pisanje FAQ sekcije predvsem za nabiranje

	vprašanj.
IZDELEK NI DOSTAVLJEN, GOLJUFIJA (Cluster 5)	Izdelki pod to kategorijo so skoraj gotovo goljufije in jih je potrebno odstraniti iz strani, sploh če se pojavijo tudi pod gručo 2. Uporabno kot sito za moderatorja strani za odstranjevanje kampanj.
POZITIVEN ODZIV STRANK (Cluster 6)	Te komentarje, kjer so stranke zadovoljne in želijo še naročiti izdelek v drugič bi se lahko uporabilo kot dobre oglase na socijalnih omrežjih. Komentarje iz te gruče naj algoritem objavlja na socialna omrežja.
FAN BOYS (Cluster 7)	Tu so komentarji ljudi ki obožujejo izdelek neglede na vse in so dobra skupina za analizo kakšnim ljudem je dobro oglaševati tak ptodukt.

6. **KORAK (Challenge Step!)**

Using your knowledge from past chapters, removed the k-Means clustering operator, and try to apply a different data mining methodology such as association rules or decision trees to your text documents. Report your results.

PRILOGE:

VIRI KOMENTARJEV:

https://www.indiegogo.com/projects/the-skarp-laser-razor-21st-century-shaving#/comments
https://www.indiegogo.com/projects/untitled--150#/comments
https://www.indiegogo.com/projects/skully-ar-1-the-world-s-smartest-motorcycle-helmet#/comments
https://www.indiegogo.com/projects/v20-best-waterproof-shoes-with-3-odor-protections#/comments
https://www.indiegogo.com/projects/racemouse-best-travel-mouse-with-laser-pointer#/comments
https://www.indiegogo.com/projects/aura-mate-pro-best-premium-updated-scanner-yet#/comments
https://www.indiegogo.com/projects/cubiio-2-autofocus-laser-cutter-metal-engraver#/comments
https://www.indiegogo.com/projects/ofiyaa-portable-triple-screen-laptop-workstation#/comments

GITHUB REPOZITORY Z NALOGO:

https://github.com/wrsna/Inteligentni_sistemi_RapidMiner