Semanticka analiza kritike filmova

# Nesto o semantickoj analizi

Semanticka analiza je process odredjivanja znacenja dela teksta. Mi ljudi smo prilicno dobri u tome ali problem nastaje kada treba da naucimo masinu da radi to za nas. Kompanije sirom sveta su implementirale masinsko ucenje da radi to za njih automatski, ovo je veoma korisno za dobijanje povratne informacije o potrosackom misljenju jer kada razumeju kako se potrosac oseca mogu da indetifikuju sta im se svidja ili ne svidja I naprave stvari kao sto su sistemi preporuke I mnoge druge stvari za njih.

# O nasem program

Ovde je napravljena semanticka analiza kritike filmova u pajtonu koja indetifikuje da lije film dobar ili los na osnovu kritike u tekstualnom formatu. Nase podatke za testiranje I treniranje uzeli smo sa sajta https://www.kaggle.com. U nasem program koristimo module za ucenje koji sume biblioteke za masinsko ucenje u pajtonu:

* from sklearn.feature\_extraction.text import CountVectorizer
* from sklearn.ensemble import RandomForestClassifier

Takodje koristimo klasu koja na pomaze da ocistimo nase podatke u nasem slucaju tekst koji je kritika filma:

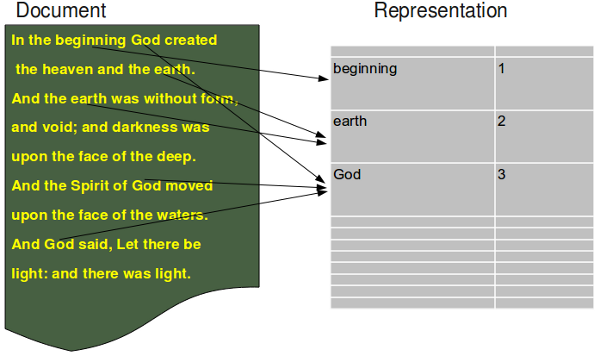
* from KaggleWord2VecUtility import KaggleWord2VecUtility

Koristimo I nltk za uklanjanje bespotrbnih reci iz podataka.

# Bag of words

Bag of words je jednostavna numericka reprezentacija dela teksta koju je lako klasifikovati. Samo izracunamo frekvenciju ponavljanja svake reci u delu teksta (slika 1) I napravimo recnik iz toga. Ovo se naziva tokenizacija u procesiranju prirodnog jezika. Koristimo CountVectorizer objekat iz paketa skleran kako bi to postigli.

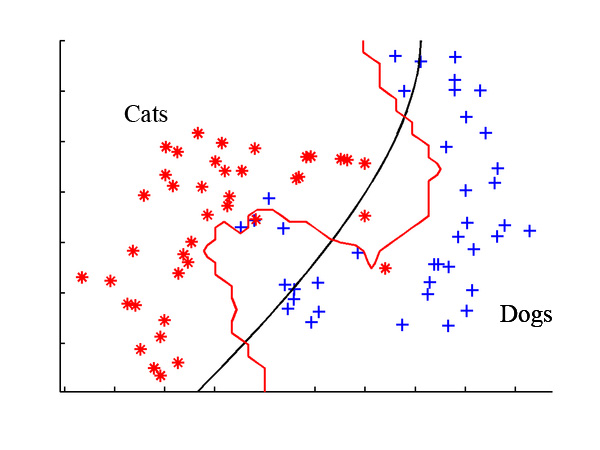
Slika 1



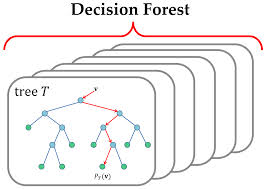
# Klasifikatori

Klasifikator je model masinskog ucenje koji ce se koristiti za klasifikovanje da li je deo teksta pozitivan ili negativan (slika 2). U nasem program klasifikator je RandomForest (slika 3) koja se sastoji od 100 stabala. RandomForest je set stabala odluke. Stabla odluke s utu kako bi modelovali moguce izlaze.

Slika 2



Slika



Ovo su u sustini komponente od kojih se sastoji nas program.