ANALIZA ISTINITOSTI TWITTER OBJAVA

Twtitter, kao i mnoge društvene mreže, postao je jedan od glavnih komunikacijskih kanala, te je podloga za objavljivanje zabavnih, informativnih sadržaja ali I npr. sadržaja o nesrećama.

Sveprisutnost mobilnih telefona omogućuje ljudima da obavijeste druge u slučaju nesreće u danom trenutku. No, nije uvijek sasvim jasno je li osoba izvještava o stvarnoj nesreći, ili je riječ o nečem drugom.

Pogledajmo primjer:

|  |  |
| --- | --- |
| tweet_screenshot.png | Autor ove objave eksplicitno koristi riječ “ablaze” što bi u doslovnom hrvatskom prijevodu značilo “u plamenu”.  U ovom slučaju, ključna riječ korištena je metaforički i svakoj osobi je jasno značenje ove objave.  No, prilikom analize podataka i ključnih riječi, računalu ne mora biti jasno što je točno autor htio reći. |

Naš zadatak je napraviti model koji analizira Twitter objave I može razlučiti koje su od njih stvarne, a koje ne.

Za pristup problemu, koristimo metodu obrade prirodnog jezika, potpodručje računarske znanosti, obrade informacija I umjetne inteligencije koje proučava interakciju između računala I ljudskih (prirodnih) jezika, posebno na načine kako procesirati I analizirati velike količine podataka iz prirodnog jezika pomoću računala.

Glavne datoteke koje će nam trebati su:

* Train.csv
* Test.csv
* Sample\_submission.csv

Svaki uzorak u train I test datoteci sadrži slijedeće informacije:

* text - sadržaj Twitter objave ili tweeta
* keyword - ključnu riječ (koja može biti prazna)
* location - lokaciju s koje je tweet poslan (također može biti prazna)
* id za svaku objavu
* target – 1 ako je stvarna nesreća, 0 ako nije

Ukoliko je analiza objave uspješna, tj. radi se o stvarnoj nesreći, program bi trebao vratiti 1, a u suprotnom 0.

Kao skup podataka koristimo dataset od 10,000 klasificiranih Twtitter objava.

Predviđeno je koristiti razne pythonove pakete kao npr

* numpy (Numerical Python) - linearnom algebrom,
* pandas (Python Data Analysis Library) - rad sa csv. datotetkama, analiza I organizacija podataka
* nltk (Natural Language Processing) - lingvistička analiza, kategorizacija teksta, morfološku analizu, čišćenje teksta od tzv. Stop riječi (primjer u engleskom su određeni I neodređeni članovi poput “the”, “a”, “an”, “in”…)
* re (Regular Expression) – niz znakova koji opisuje ruge nizove znakova u skladu s određenim sintaksnim pravilima
* scikit-learn (ekstenzija scipyja – Scientific Python) – machine learning biblioteka s algoritmima za klasifikaciju, regresiju (nadzirano učenje), grupiranje (ili clustering; nenadzirano učenje)

Ocjena uspješnosti rezultata projekta je ukupan broj pogođenih predikcija stvarnih nesreća.

Cilj nam je predstaviti dva algoritma “Cross-validation ili KFold “ I “random forrest” I usporediti njihove rezultate.

**Popis literature:**

<https://www.kaggle.com/c/nlp-getting-started/overview>