Facultatea de Matematică şi Informatică
Universitatea din Bucureşti

DOCUMENTATIE PROIECT INTELIGENTA ARTIFICALA

=CLASIFICARE TWEET-URI MISOGINE DIN LIMBA ITALIANA=

Nume: Secioreanu Ionel Stefanita
Domeniul: CTI
Grupa: 351
Semestrul 1 2020

Descrierea Cerintei

- **Scopul proiectului**

Se cere crearea unui program de clasificare text.Programul consta in procesarea de tweet-uri din limba italiana , punand label-urile de misogin, repsectiv nemisogin pe acestea.

- **Descrierea datelor**

Textul este impartit in 2 parti inegale , un fisier train.csv, ce contine 5000 de tweet-uri cu label-uri(1 pentru tweet misogin, 0 pentru nemisogin) si un fisier test.csv, cu 1000 de tweet-uri pe care se doreste a face procesarea de text. Predictiile vor fi facute sub forma unui fisier .csv format din 2 coloane , una cu id-ul tweet-ului, alta cu label-ul asociat, fisierul este apoi incarcat pe platforma Kaggle pentru a obtine un punctaj de acuratete.

Pasii Proiectului

Pentru o strategie locala mai buna datele din fisierul de antrenare ai fost imaprtite in 2 bucati inegale.75% din fisier a fost pus intr-un set “multime_de_antrenare_si_validare” iar restul de 25% in setul “multime_de_testare” prin functia “split” cu ajutorul ShuffleSplit din libraria sklearn (functie ce permuta la intamplare si imparte un fisier intr-un set de testare si unul de antrenare).Pentru o mai buna practica locala si pentru a evita fenomenul de overfitting (antrenam pe datele noastre foarte bine dar pe date noi algoritmul nu se descurca bine) a fost folosita si strategia de K-fold Cross-Validation cu K=10, in care datele din setul” multime_de_antrenare_si_validare” au fost impartite in 10 parti egale , 9 dintre ele folosind drept antrenare iar a 10 fiind partea de testare , fiecare parte din cele 9 fiind pe rand o parte de testare.Pentru aceste multimi au fost calculate pe rand scorul f1(media armonica dintre precizia si recalul calculat). Precizia fiind tweet-urile misogine din model / tweet-urile totale din model iar Recall-ul este reprezentat prin procentul de tweet-uri relevante selectate din tweet-urile totale.

Dupa care intr-un set sunt puse toate cuvintele din toate tweet-urile din setul “multime_de_antrenare_si_validare” prin intermediul functiei “get_corpus_vocabulary”, functie ce construieste un conter si il returneaza cu tweeturile tokenizate (prin intermediul TweetTokenizer din libraria nltk).

In continuare au fost create doua dictionare “wd2idx” si “idx2wd” prin functia “get_reprezentation” in care sunt puse cele mai uzuale N cuvinte .

Pentru o reprezentare numerica a datelor a fost folosit modelul Bag-Of-Words unde corpusul nostru de cuvinte tokenizate a fost transformat intr-un bow ,unde pentru fiecare text/tweet apare frecventa celor N cuvinte cele mai uzuale.

Dupa care se creaza un estimator pe care sunt date fit bag-of-words-ul creat si label-urile .se creaza un bag of words pentru textul din test.csv si se foloseste estimatorul pentru a da predict pe acest bow.Aceste predictii sunt salvate intr-o variabila si sunt scrise intr-un fisier "submisieProiectIA.csv".

Clasificatorii folositi

Clasificatorii folositi sunt KNN(K-Nearest-Neighbor) si Bernoulli Naive Bayes.Acestia au fost importati din libraria sklearn.neighbors respectiv sklearn.naive_bayes.

- **KNN**

Metrica folosita este cea default Minkowski. Iar hyper-parametrul ales in urma incercarilor manuale in functie de punctaje locale este 12(numarul de vecini).In acest algoritm se fac predictii , data este plasata si se calculeaza metrica pana la cei mai apropriati K vecini si in functie de label-urile vecinilor se face o predictie .Modelul lucreaza cu primele 100 de cuvinte cele mai frecvente.

- **Naïve Bayes Bernoulli**

Acest model este unul probabilistic , cu hyper-parametrul alfa default 1.0. Distributia folosita fiind Bernoulli.Ia feature-urile si calculeaza probabilitati diferite de a fi intr-un tweet misogin , la sfarsit returnand o probabilitate de a fi tweetul misogin sau nu .Numarul de cuvinte folosite(feature-uri) a fost de 1868.

Scoruri F1 pentru K-Cross Validation cu KNN(12) 100 de cuvinte

[0.83028721 0.85340314 0.81889764 0.82597403 0.83333333 0.80839895

0.85026738 0.8342246 0.83937824 0.83287671]

Timp de rulare :08.37 secunde

Scoruri F1 pentru Naïve Bayes Bernoull(1868 de cuvinte)

[0.86685552 0.87536232 0.86127168 0.85878963 0.85633803 0.86666667

0.87179487 0.86857143 0.84408602 0.84240688]

Timp de rulare :30.27 secunde

Matrice de confuzie Naïve Bayes Bernoull(1868 de cuvinte)

[[1519, 253], [237, 1741]]

Matrice de confuzie KNN K=12 si 100 de cuvinte

[[1580, 192], [443, 1535]]