Angelos Kamaris sdi1900070
AI 2 - project 1

November 28, 2022

Καμάρης Άγγελος sdi1900070

1 Πηγές

Αξιοποίησα τον κώδικα του φροντιστηρίου κατά κύριο λόγο, χρησιμοποιώντας επίσης εντολές από το sklearn. Αξιοποίησα τον οδηγό:
https://www.kirenz.com/post/2021-12-11-text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/
για NLTK preprocessing των δεδομένων μου.

2 Επεξήγηση Κώδικα

Χρησιμοποίησα τις στήλες rating και review σαν Υ και X αντίστοιχα, όπου επεξεργάστηκα τα rating, έτσι ώστε τα "κακά" (0-4) να έχουν την τιμή 0 και τα "καλά" (7-10) να έχουν την τιμή 1.

Επεξεργάστηκα τα δεδομένα από τα reviews κάνοντας τα κεφαλαία μικρά, βγάζοντας τα σημεία στίξης, και τις λέξεις που είχαν λιγότερα από 2 γράμματα, βρίσκοντας τις πιο πολυχρησιμοποιημένες λέξεις και αφαιρώντας τις πιο σπάνιες, και τέλος κάνοντας lematize και stematize .

Χώρισα τα δεδομένα μου έτσι ώστε 30% εξ αυτών να γίνονται αλιδατε και τα υπόλοιπα να χρησιμοποιούνται για το τραινινγ, χρησιμοποίησα tf-idfκαθώς μου έφερνε καλύτερα αποτελέσματα, από τους άλλους που αναφέρθηκαν στο φροντιστήριο.

Χρησιμοποιώ logistic regression για να κάνω train τα δεδομένα μου και εμφανίζω το την ευστοχία του με cross validation. Αξιοποιώ το learning curvetου φροντιστηρίου για να εκτυπώσω τις αποδόσεις του classifier μου με τον αριθμό των δεδομένων, σύμφωνα με το F1-score τους και το μέγεθος των δεδομένων.

Τέλος εκτυπώνω τα: F1-Score, Recall, Precisionγια τα trainingκαι testπου χώρισα πριν.

Αποφάσισα να βάλω max features στο tf-idf=1000, καθώς για κάτω από 500 δεν έχω εξίσου καλό accuracy (κάτω από 0.86) και για πάνω από 1000 έχω overfitting. Ο classifier έχει max iter=2000, μιας και δεν καθυστερεί με αυτό αλλά ούτε κάνει overfitting. Τέλος έδωσα 30% των δεδομένων μου στοtest, ώστε να έχω αρκετά δεδομένα για train.

Για την εισαγωγή ενός test, αρκεί να εισάγετε τα δεδομένα στην μεταβλητή test df , όπως το παράδειγμα στα σχόλια.

3 Παρατηρήσεις

Την μεγαλύτερη αλλαγή την παρατήρησα, στην χρήση preprocessing, όταν το μοντέλο δεν έκανε overfit ή underfit καθώς παρατηρούνται αλλαγές της τάξης $0.2 - 0.8$, αναλόγως και τα υπόλοιπα δεδομένα.

Η κύρια αιτία overfitting ήταν ο μεγάλος αριθμός feature στο vectorization ενώ για underfitting ήταν ο μικρός αριθμός δεδομένων εκπαίδευσης.

4 Αποτελέσματα

Τα αποτελέσματα που έχει το πρόγραμμά μου για τα δεδομένα που ανέφερα είναι:
F1 Score (train): 0.88153321015877 F1 Score (validation): 0.8702659145850121
Recall Score (train): 0.8921779918864098 Recall Score (validation): 0.8823529411764706
Precision Score (train): 0.8711394442037507 Precision Score (validation): 0.8585055643879174