1η ΕΡΓΑΣΙΑ

ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2017-2018

ΜΕΛΗ:

-ΓΙΑΝΝΗΣ ΓΙΑΝΝΑΚΙΔΗΣ 1115201500025

-ΚΩΝΣΤΑΝΤΙΝΟΣ ΠΑΣΧΟΠΟΥΛΟΣ 1115201500127

1. Wordcloud

Για το ερώτημα αυτό αρχικά θέτουμε τα stopwords με τα set που παρέχει το wordcloud και το επεκτείνουμε με το set του scikit.

Στην συνέχεια μετά από εκτελέσεις του προγράμματος διαπιστώσαμε αρκέτες παρόμοιες λέξεις στα wordcloud που δημιουργούνται.Για αυτό και κάνουμε set κάποιες ακόμα stopwords έτσι ώστε το τελικό αποτέλεσμα να είναι πιο αντιπροσωπευτικό για κάθε κατηγορία.

Για την δημιουργία των wordcloud αρχικά διαβάζουμε τα δεδομένα του train_set και τα αποθηκεύουμε ανά κατηγορία.Δημιουργούμε το wordcloud μέσω της αντίστοιχης συνάρτησης για κάθε κατηγορία δίνοντας ως τίτλο την αντίστοιχη κατηγορία καθώς και συγκεκριμμένες διαστάσεις για περισσότερο ευδιάκριτα αποτελέσματα στα .png αρχεία που δημιουργούνται.

2. Classification

Αρχικά διαβάζουμε τα δεδομένα από τα 2 .csv και αφαιρούμε την στήλη RowNum από τα δεδομένα μας με βάση την υπόδειξη της εκφώνησης.Ως stopwords χρησιμοποιούμε εκείνα του scikit και τα επεκτείνουμε με τα δικά μας που εντωπίσαμε στο wordcloud.

Για την επεξεργασία των δεδομένων μέσω της συνάρτησης TruncatedSVD() θέτουμε το μέγεθος των στοιχείων από τα οποία θα

αποτελούνται οι σειρές για τους πίνακες μας.Μέσω της preprocessing.LabelEncoder() θέτουμε τα labels για τις κατηγορίες μας και με την .fit() τα προσαρμόζουμε στο πλήθος των κατηγοριών που έχουμε στα δεδομένα μας.Με την συνάρτηση .transform() μετατρέπουμε τις κατηγορίες μας στα αντίστοιχα label.Για τα συγκεκριμμένα δεδομένα του train_set έχουμε 5 κατηγορίες άρα τα label είναι {0,1,2,3,4}.Στην συνέχεια μέσω της CountVectorizer μετατρέπουμε τα text μας ως έναν πίνακα απο ακεραίους για να είναι στην μορφή που θέλουμε και κάνουμε fit έτσι ώστε κάθε σειρά να αποτελείται από το ίδιο πλήθος στοιχείων.

Στην συνέχεια περνάμε στις μεθόδους κατηγοριοποίησης που ζητούνται.

-Support Vector Machines:

Μέσω της συνάρτησης .SVC() χρησιμοποιούμε την μέθοδο κατηγοριοποίησης.Έγιναν αρκετές δοκιμές για να καταλήξουμε στην συνάρτηση rbf για τον kernel καθώς είχε τα καλύτερα αποτελέσματα.Δώσαμε υψηλή παράμετρο C έτσι ώστε το μοντέλο να έχει περισσότερες επιλογές.

Στην συνέχεια μέσω της cross_val_score() υπολογίζουμε τις 4 μετρικές.

-Random Forests:

Μέσω της συνάρτησης RandomForestClassifier() χρησιμοποιούμε την μέθοδο κατηγοριοποίησης Random Forests.Όσο αναφορά τις παραμέτρους θέτουμε τον random number generator ίσο με ένα.Κάναμε αρκετές δοκιμές για τις υπόλοιπες παραμέτρους και τα αποτελέσματα ήταν παρόμοια εκτός από το μέγιστο βάθος του δέντρου.Εκεί παρατηρήσαμε πως αν δώσουμε κάποιον περιορισμό τότε υπάρχει μεγαλύτερη ακρίβεια στις μετρικές μας.

-Multinomial Naive Bayes:

Μέσω της συνάρτησης MNB() χρησιμοποιούμε αυτήν την μέθοδο

κατηγοροιοποίησης. Τα αποτελέσματα των μετρικών σε αυτήν την μέθοδο είναι αρκετά υψηλά και όποιες δοκιμές στις παραμέτρους είχαν ελάχιστες εώς και καθόλου βελτιώσεις καθώς και σε πολλές περιπτώσεις οι αποδόσεις παρουσίαζαν μείωση.Μία μικρή παρατήρηση που έγινε είναι πως για την παράμετρο alpha τιμές μεγαλύτερες του 0.0 και μικρότερες στο 1.0 πετύχαιναν αρκετές φορές καλύτερες αποδόσεις.Με τιμή του alpha=0.0 τα αποτελέσματα των μετρικών είχαν αισθητά μικρότερη απόδοση.

-K-NearestNeighbor:

Η υλοποίηση μας βρίσκεται στο αρχείο knn_functions.py

Αρχικά ελέγχουμε αν τα δεδομένα μας είναι στην σωστή μορφή για την σωστή λειτουργία του αλγόριθμου.Για να υπολογίσουμε την ομοιότητα μεταξύ 2 στιγμιοτύπων των δεδομένων χρησιμοποιείται η ευκλείδια απόσταση.Υπολογίζουμε την ομοιότητα ενός στιγμιοτύπου του test_set με ολόκληρο το train_set και κρατάμε τους k πιο όμοιους γείτονες.Στην συνέχεια με βάση τις συμβουλές της εκφώνησης χρησιμοποιούμε majority voting για να αποφασίσουμε.Έγιναν αρκετές δοκιμές για το πλήθος των γειτόνων(δηλαδή του k).Ένα εύρος τιμών μεταξύ [5,10] επιστρέφει υψηλές αποδόσεις.Αρκετά μικρές τιμές του k και αρκετά μεγάλες τιμές παρουσιάζουν αποκλίσεις στην απόδοση.

3. Beat the benchmark

Προκειμένου να ξεπεράσουμε την απόδοση μας αρχικά κάνουμε προεπεξεργασία στα δεδομένα μας.Το πρώτο στάδιο αυτής της προεπεξεργασίας αφορά στο να αφαιρέσουμε τα σημεία στίξης από τα δεδομένα μας.Αυτό επιτυγχάνεται μέσω της str.replace().Το δεύτερο στάδιο της προεπεξεργασίας μετατρέπει πιθανόν πολλαπλά space σε 1 space μέσω της ίδιας συνάρτησης.

Ως μέθοδο κατηγοριοποίησης χρησιμοποιούμε την Multinomial Naive Bayes καθώς είχαμε σταθερά τις υψηλότερες αποδόσεις μέσω αυτής.Χρησιμοποιούμε έναν αλγόριθμο που φτίαξαμε προκείμενουν

να βρούμε ποια παράμετρος alpha επιστρέφει τις υψηλότερες αποδόσεις.Κρατάμε αυτήν την παράμετρο και υπολογίζουμε όλες τις μετρικές μας.

Τέλος δημιουργούμε τα .csv αρχεία.