Τεχνικές Εξόρυξης Δεδομένων
1η Άσκηση

Βρουλιώτης Βασίλειος, Α.Μ. : 1115201300025
Κοκκάλης Γεώργιος, Α.Μ. : 1115201300069

WordCloud :

Αρχεία υλοποίησης : FootballCloud.py, FilmCloud,py, BusinessCloud.py, PoliticsCloud.py, Technology.py

Δημιουργούμε ένα αρχείο για καθε μία από τις κατηγορίες των άρθρων (πέντε), και σε κάθε αρχείο δίνουμε ως είσοδο στο wordcloud την αντίστοιχη κατηγορία του. Έτσι στο τέλος κάθε .py αρχείο αποθηκεύει σε ένα .png το αποτέλεσμα που παράγει.

Classification :

Αρχείο υλοποίησης : Classification.py

Για κάθε μια από τις μεθόδους classification χρησιμοποιούμε 10-fold Cross Validation με τις μετρικές Accuracy, ROC plot, για να αξιολογήσουμε και να καταγράψουμε την απόδοση τους.

Χρησιμοποιούμε μόνο το Content για την εκμάθηση των διαφόρων classifiers και για την πρόβλεψη των νέων δεδομένων. Ως προεπεξεργασία χρησιμοποιούμε την τεχνική LSI πάνω σε αυτό, με την επιθυμητή διάσταση να είναι 200.

Πρώτα υπολογίζουμε το accuracy, και στη συνέχεια φτιάχνουμε το αντίστοιχο διάγραμμα για καθένα fold. Για κάθε μία μέθοδο, φτιάχνουμε 10 διαγράμματα, τα οποία τοποθετούνται στο φάκελο : "ROC_pngs". Έπειτα, αποθηκεύουμε όλες αυτές τις πληροφορίες στο "EvaluationMetric_10fold.csv". Επίσης, για κάθε μέθοδο κάνουμε fit με το 90% των δεδομένων, και με το υπόλοιπο 10% predict. Στη συνέχεια υπολογίζουμε το ποσοστό των σωστών προβλέψεων, και εμφανίζουμε κατάλληλο μήνυμα.

Τέλος, για τη δημιουργία "testSet_categories.csv" χρησιμοποιούμε μόνο τη μέθοδο NaiveBayes(MultinomialNB), δίνοντας τα πρώτα 10.000 άρθρα ως train, και τα υπόλοιπα για test.

Όσον αφορά στη δικιά μας μέθοδο, με τη χρήση 10-fold Cross Evaluation με μετρική το accuracy, ψάχνουμε να βρούμε για πόσους γείτονες ο K nearest neighbours παρουσιάζει την καλύτερη συμπεριφορά. Και στη συνέχεια κάνουμε ότι και για τις υπόλοιπες μεθόδους.

Σημείωση : Ό,τι αφορά στη μέθοδο SVC απαιτεί πολύ περισσότερο χρόνο από τα υπόλοιπα.

Clustering :

Αρχείο υλοποίησης : Clustering.py

Δημιουργούμε το αρχείο “clustering_KMeans.csv”, το οποίο μας δίνει τα περιεχόμενα του κάθε cluster. Έχει επιλεχθεί ο αριθμός 20 ως ο μέγιστος αριθμός επαναλήψεων, αν δεν έχουν βρεθεί ίδια κέντρα.

Χρησιμοποιούμε μόνο το Content για την εύρεση των clusters. Ως προεπεξεργασία χρησιμοποιούμε την τεχνική LSI πάνω σε αυτό, με την επιθυμητή διάσταση να είναι 200.

Επίσης, λόγω του random για την εύρεση των αρχικών κέντρων, σε κάποιες εκτελέσεις, κάποια clusters είτε περιέχουν πολλά άρθρα από 2 κατηγορίες, είτε είναι σχεδόν άδεια. Αυτό οφείλεται κυρίως στην κατηγορία Technology επειδή περιέχει πολύ λιγότερα άρθα από τα υπόλοιπα.

Σημείωση : Δεν χρησιμοποιούμε πουθενά το αρχείο “test.csv”.