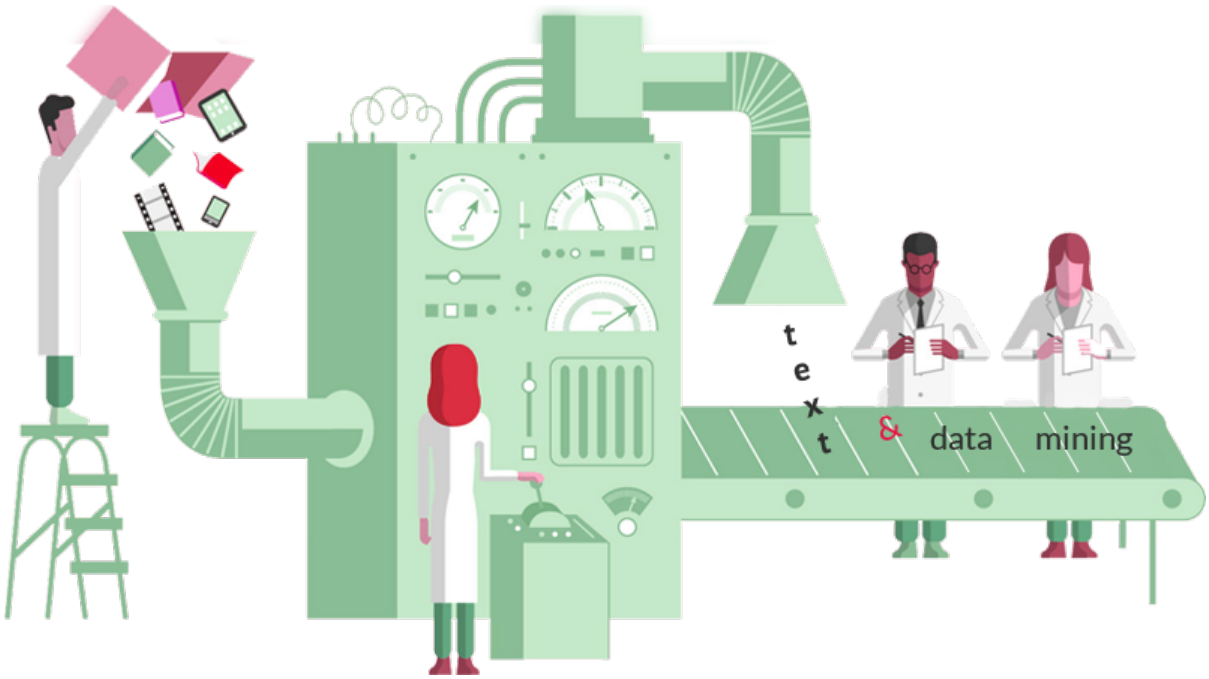
Τεχνικές Εξόρυξης Δεδομένων: Εργασία 2η



Μέλη Ομάδας:

Όνομα:	**Επώνυμο:**	**Αριθμός Μητρώου:**
Αντωνία	Αθανασάκου	1115201400004
Στεφανία	Πάτσου	1115201400156

Περιεχόμενα

Πρόλογος

Σε αυτή την εργασία κληθήκαμε να υλοποιήσουμε την οπτικοποίηση δεδομένων ενός δοκιμαστικού αρχείου “train.tsv” και κατηγοριοποίηση του αρχείου “test_set.csv” με βάση το δοκιμαστικό. Ακόμη, υπολογίσαμε το information gain για κάθε feature και σχεδιάσαμε τα κατάλληλα γραφήματα. Αρχικά δημιουργήσαμε, για κάθε attribute του αρχείου “ train.tsv ”, ένα box plot/histogram ανάλογα με το αν το feature ήταν numerical/categorical αντίστοιχα. Στη συνέχεια, με βάση τα καλύτερα ποσοστά ακρίβειας, επιλέξαμε έναν αλγόριθμο κατηγοριοποίησης και τον εκτελέσαμε στο αρχείο “test.tsv”. Τέλος, με βάση το information gain, σχεδιάσαμε το γράφημα των accuracies αφαιρώντας features και δημιουργήσαμε έναν πίνακα “infoGain.csv”.

Δημιουργία Οπτικοποιημένων Δεδομένων

Για τη δημιουργία των οπτικοποιημένων δεδομένων, αποθηκεύσαμε τις στήλες του ‘Label’ σε good και bad. Ύστερα, από αυτές και για κάθε attribute, δημιουργήσαμε τα box plots και histograms. Αξίζει να σημειωθεί ότι πριν από αυτό το βήμα, είχαμε χωρίσει τις κατηγορίες σε numerical και categorical. Για την δημιουργία των box plots καλέσαμε την συνάρτηση .boxplot της βιβλιοθήκης matplotlib και με κατάλληλες προσθαφαιρέσεις ιδιοτήτων φτάσαμε στο ιδανικό αποτέλεσμα. Η ίδια διαδικασία επαναλήφθηκε για την δημιουργία των histograms μόνο που χρησιμοποιήσαμε την συνάρτηση value_counts() η οποία είναι μέρος της βιβλιοθήκης pandas.Series και η οποία επιστρέφει μία λίστα που περιλαμβάνει τους το σύνολο της κάθε κατηγορίας του feature.

Υλοποίηση Κατηγοριοποίησης (Classification)

Για την κατηγοριοποίηση των δεδομένων, δοκιμάσαμε τις μεθόδους Support Vector Machines (SVM), Random Forests και Naive Bayes. Χρησιμοποιήσαμε “10-fold Cross Validation”(στην οποία στην παράμετρο cv = k_fold εγίνε με βάση το Kfold) για να αξιολογήσουμε την μετρική “Accuracy”. Τα αποτελέσματα των μετρικών έχουν τοποθετηθεί στο “EvaluationMetric_10fold.csv”. Ο πίνακας για τις μετρικές φαίνεται και παρακάτω:

Statistic Measure	**Naive Bayes**	**Random Forest**	**SVM**
Accuracy	0.67125	0.6225	0.64

Όπως προκύπτει από τον πίνακα, τα πιο αποδοτικά αποτελέσματα δίνει ο αλγόριθμος Naïve Bayes.

Προσθέσαμε, επίσης, σε κάθε γραμμή ως πρώτο στοιχείο το ‘ ’ για να φαίνονται καλύτερα τα αποτελέσματα.

Support Vector Machines (SVM)

Η βιβλιοθήκη που χρησιμοποιήσαμε είναι η sklearn και συγκεκριμένα η svm.

Random Forests

Χρησιμοποιήθηκε η βιβλιοθήκη RandomForestClassifier.

Naive Bayes

Χρησιμοποιήθηκε η βιβλιοθήκη GaussianNB.

Υλοποίηση Κατηγοριοποίησης στο “test.tsv ”

Από τα αποτελέσματα της κατηγοριοποίησης, καταλήξαμε ότι πιο αποδοτικός ήταν ο αλγόριθμος Naive Bayes. ‘Ετσι εφαρμόσαμε αυτόν τον αλγόριθμο, για να προβλέψουμε τις κατηγορίες του αρχείου “test.tsv”. Τα αποτελέσματα υπάρχουν σε ένα αρχείο “ testSet_Predictions.csv”.

Υπολογισμός Information Gain

Για τον υπολογισμό του Information Gain για κάθε feature, υπολογίσαμε αρχικά την εντροπία της κατηγορίας Label. Στη συνέχεια υπολογίσαμε για κάθε attribute την εντροπία του attribute με βάση το Label και αφαιρώντας αυτά τα δύο καταλήξαμε στο Information Gain. Πριν υπολογίσουμε την εντροπία για κάποιο numerical attribute, το μετατρέψαμε σε categorical χρησιμοποιώντας τη συνάρτηση pandas.cut (bins = 5).
Αφού ταξινομήσαμε τη λίστα με τα Information Gains, αφαιρέσαμε ένα attribute τη φορά (μέχρι να μείνει στο τέλος ένα) και υπολογίσαμε το accuracy χρησιμοποιώντας τον αλγόριθμο Naive Bayes για κατηγοριοποίηση. Τα αποτελέσματα βρίσκονται σε ένα σχεδιάγραμμα, ενώ δημιουργήσαμε και ένα αρχείο .csv για να εμφανίσουμε το attribute που αφαιρέσαμε και το Information Gain αυτού του attribute.
Παρακάτω φαίνεται ο πίνακας με τα attributes και τα Information Gains αυτών:

Attributes	**Information Gain**
Attribute18	0.000130
Attribute11	0.000221
Attribute19	0.001203
Attribute16	0.002396
Attribute17	0.002940
Attribute10	0.005674
Attribute14	0.007042
Attribute8	0.007331
Attribute20	0.007704
Attribute15	0.011619
Attribute9	0.012747

Attribute13	0.013413
Attribute7	0.014548
Attribute12	0.014906
Attribute5	0.018461
Attribute6	0.022199
Attribute4	0.026897
Attribute2	0.032963
Attribute3	0.037889
Attribute1	0.093828

Επίλογος / Σχόλια

Όλα τα αρχεία με τα αποτελέσματα είναι τοποθετημένα στον φάκελο “Output ”, έτσι ώστε να έχετε τη δυνατότητα να τα εξετάσετε, χωρίς να απαιτείται να εκτελέσετε το πρόγραμμα. Πριν εφαρμόσουμε την κατηγοριοποίηση των δεδομένων, μετατρέψαμε τα Categorical Attributes σε Numerical και χρησιμοποιήσαμε TfidfVectorizer() και TruncatedSVD (n_components=10), για προεπεξεργασία των δεδομένων.