NLP Assignment 1

Στα αρχεία που περιλαμβάνονται στο φάκελο περιέχεται ο κώδικα υλοποίησης των απαιτήσεων της εκφώνησης της εργασίας.

Μέρος 1

Αυτό το τμήμα της εργασίας αφορά στην υλοποίηση με κώδικα των διαφόρων παραλλαγών του αλγορίθμου Naive Bayes, χωρίς τη χρήση των έτοιμων βιβλιοθηκών της python. Ο απαραίτητος κώδικας περιλαμβάνεται στα ακόλουθα αρχεία:

- **1a-b.py**: Υλοποιούνται οι αλγόριθμοι Multinomial Naive Bayes και Binary Multinomial Naive Bayes.
- **1c.py**: Υλοποιείται ο αλγόριθμος Multivariate Bernoulli Naive Bayes.
- **commonModules.py**: Περιέχει τον κώδικα κοινών μεθόδων, που γίνονται εισαγωγή και χρησιμοποιούνται και από τους 3 αλγορίθμους. Πρέπει να βρίσκεται στον ίδιο φάκελο με τα υπόλοιπα αρχεία κώδικα.

Οι μέθοδοι των αλγορίθμων υλοποιήθηκαν εξ' αρχής, σύμφωνα με τις απαιτήσεις της εκφώνησης και όλη η απαραίτητη τεκμηρίωση παρέχεται με τη μορφή docstrings. Για τον έλεγχο των αλγορίθμων χρησιμοποιήθηκαν υποτυπώδη δεδομένα εισόδου από το βιβλίο και τα αποτελέσματα τυπώνονται άμεσα στο Terminal.

Μέρος 2

Αυτό το τμήμα της εργασίας αφορά στην συγκριτική εφαρμογή 4 αλγορίθμων πάνω σε παρεχόμενα έγγραφα. Τα εξεταζόμενα έγγραφα βρίσκονται μέσα στο φάκελο **files**, ο οποίος πρέπει να βρίσκεται στον ίδιο φάκελο με τα υπόλοιπα αρχεία κώδικα.
Οι χρησιμοποιούμενοι αλγόριθμοι είναι οι 3 του πρώτου μέρους μαζί με τον Logistic Regression. Σε αυτό το μέρος της εργασίας οι αλγόριθμοι υλοποιήθηκαν μέσω των έτοιμων βιβλιοθηκών που προσφέρει η python. Ο απαραίτητος κώδικας περιλαμβάνεται στο αρχείο 2.ipynb, το οποίο είναι αρχείο Google Colaboratory Python 3 Notebook, σύμφωνα με τις απαιτήσεις της εκφώνησης.
Οι διαδικασίες που υλοποιούνται με την εκτέλεση του κώδικα είναι οι ακόλουθες:

- Ανάγνωση των αρχείων από το δίσκο και εξαγωγή-αποθήκευση όλων των απαραίτητων για την ανάλυση στοιχείων.
- Έλεγχος της κατανομής των δεδομένων ανά κλάση.
- Εφαρμογή διαφόρων κατηγοριών Preprocessing των δεδομένων.
- Εκτέλεση και καταγραφή-απεικόνιση της απόδοσης των 4 αλγορίθμων με τις εξ' ορισμού τιμές παραμέτρων.
- Εκτέλεση και καταγραφή της απόδοσης των 4 αλγορίθμων με χρήση κατάλληλης βιβλιοθήκης που βρίσκει τις βέλτιστες τιμές παραμέτρων. Η παραπάνω διαδικασία επαναλαμβάνεται ξεχωριστά για τις διάφορες μορφές Preprocessing των δεδομένων.
- Τέλος οι βέλτιστες τιμές παραμέτρων των αλγορίθμων χρησιμοποιούνται για να γίνει συγκριτική μελέτη απόδοσης των αλγορίθμων, όταν των αλγορίθμων προηγείται απλή μετατροπή των δεδομένων σε διανύσματα ή συνδυασμός με κανονικοποίηση των διανυσμάτων (εφαρμογή των αλγορίθμων CountVectorize ή TfidfVectorize αντίστοιχα)

Όλα τα δεδομένα που αφορούν τα αποτελέσματα εκτέλεσης των αλγορίθμων εμφανίζονται σε κατάλληλους πίνακες ή/και γραφήματα μέσα στο Notebook. Επίσης, σύμφωνα με τις απαιτήσεις της εκφώνησης, όλες οι πληροφορίες που αφορούν τις εφαρμοζόμενες μεθόδους, καθώς και οι αξιολογήσεις των αποτελεσμάτων, εμφανίζονται σε κατάλληλα κελιά του Notebook.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Jurafsky, Dan & Martin, James H. - Speech and Language Processing, 3 rd edition
2. Spam filtering with naive Bayes - which naive Bayes?, από τα παρεχόμενα additional readings
3. Το υλικό από τα εργαστηριακά μαθήματα
4. Η επίσημη τεκμηρίωση της βιβλιοθήκης sklearn για κάθε ένα χρησιμοποιούμενο αλγόριθμο και μέθοδο.