ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΓΡΑΠΤΗ ΕΡΓΑΣΙΑ 4
Bayesian Learning

Σκοπός της συγκεκριμένης εργασίας είναι η εξοικείωση με την υλοποίηση του αλγορίθμου Multinomial Naive Bayes, μέσω της αντίστοιχης μεθόδου της βιβλιοθήκης sci-kit της Python. Η εφαρμογή τους έγινε με χρήση του DataSet “20newsgroups”, της βιβλιοθήκης sklearn. Τα δεδομένα, πριν εισαχθούν στον classifier, υφίστανται προεπεξεργασία (preprrocessing), μετατρεπόμενα σε διανύσματα (vectors) σύμφωνα με τη μέθοδο TF-IDF, όπως αυτή υλοποιείται στην βιβλιοθήκη sklearn. Λόγω της ύπαρξης διαδοχικών αλγορίθμων που επιδρούσαν επί των δεδομένων χρησιμοποιήθηκε η δυνατότητα δημιουργίας pipeline διεργασιών, με τη βοήθεια της αντίστοιχης βιβλιοθήκης του sklearn. Οι βιβλιοθήκες των αλγορίθμων χρησιμοποιήθηκαν στην default μορφή τους, εκτός από τη παράμετρο alpha του Naive Bayes, που μελετήθηκε αναλυτικά για την επίδρασή του στην απόδοση της διαδικασίας. Η αξιολόγηση και η σύγκριση της απόδοσης των διαφορετιά παραμετροποιημένων αλγορίθμων έγινε, σύμφωνα με τις οδηγίες της εκφώνησης, με χρήση των ακόλουθων μετρικών (στον υπολογισμό των μετρικών επιλέχθηκε τιμή παραμέτρου average=’macro’):

- **precision**: Ορίζεται ως TP / (TP+FP)
- **recall**: Ορίζεται ως TP / (TP+FN)
- **f1**: Ορίζεται ως 2*precision*recall / (precision+recall)

Α. Περιγραφή και μελέτη του αλγορίθμου

Σύμφωνα με τις απαιτήσεις της εκφώνησης χρησιμοποιήθηκαν συγκεκριμένοι αλγόριθμοι επί των δεδομένων, με χρήση των ακόλουθων βιβλιοθηκών της sklearn:

- **TfidfVectorizer**: Κατεργάζεται κατάλληλα τα κείμενα εισόδου. Οι κυριότερες τροποποιήσεις που επιφέρει, στη default διαμόρφωση, είναι ότι μετατρέπει τους χαρακτήρες σε πεζούς και στη συνέχεια διαχωρίζει το κείμενο σε ανεξάρτητες λέξεις (tokenization), που αποτελούν και τα features της μεθόδου. Τέλος μετατρέπει κάθε έγγραφο εισόδου σε ένα ανεξάρτητο διάνυσμα (vector), με παραμέτρους όλα τα features και τιμές που είναι κατάλληλα κανονικοποιημένες έτσι ώστε οι πιο κοινές και συχνές λέξεις (όπως οι stop words) που έχουν μικρή διακριτική αξία, να παρουσιάζουν υποβαθμισμένη τιμή.
- **MultinomialNB**: Υλοποιεί την ταξινόμηση των επεξεργασμένων κειμένων εισόδου, σύμφωνα με τον αλγόριθμο Naive Bayes. Χρησιμοποιείται στην default μορφή του με εξαίρεση την παρακάτω παράμετρο:
    - alpha: Επιδρά πάνω στην Laplace/Lidstone εξομάλυνση (smoothing). Μελετήθηκσν σταδιακά αυξανόμενες τιμές από 0,01 έως και 1.

Ο αλγόριθμος εκτελέστηκε επαναληπτικά για τις διάφορες τιμές της παραμέτρου alpha, με παράλληλη καταγραφή των τιμών των μετρικών. Ο κώδικας υλοποίησης της επαναληπτικής εκτέλεσης του αλγορίθμου υπάρχει στο συνοδευτικό αρχείο *Iliapikos_ge4_NB_studyAlpha.py*. Η μεταβολή των τιμών των μετρικών φαίνεται στον παρακάτω πίνακα.

Β. Σύγκριση των παραλλαγών του αλγορίθμου

Βλέποντας συγκριτικά τα αποτελέσματα παρατηρώ ότι η απόδοση του αλγορίθμου συνολικά κινείται σε υψηλά επίπεδα και δεν επηρεάζεται δραματικά με τη μεταβολή της παραμέτρου alpha σε όλο το εύρος τιμών της. Αυτό πιθανότατα οφείλεται στο γεγονός ότι ο αλγόριθμος είναι ήδη

alpha	Recall	Precission	F1 score
0,01	0,913	0,917	0,914
0,02	0,911	0,916	0,912
0,05	0,902	0,912	0,903
0,10	0,893	0,907	0,894
0,15	0,887	0,905	0,887
0,20	0,878	0,900	0,878
0,50	0,855	0,889	0,854
1,00	0,829	0,879	0,827

overfitted. Για το σκοπό αυτό προχώρησα στο προαιρετικό τμήμα της εργασίας όπου μας ζητείται να παραμετροποιήσουμε το συνολικό αλγόριθμο κατάλληλα ώστε η απόδοσή του, βασιζόμενοι στη μετρική F1, να πέσει στο επίπεδο του 0.70. Εξετάζοντας τις παραμέτρους του αλγορίθμου Naive Bayes, ξεχώρισα αυτές που θα μπορούσαν, σε κάποια λογική βάση, να επηρεάσουν την εξειδίκευση του μοντέλου στα δεδομένα εκπαίδευσης:

- **max_features**: Περιορίζει το λεξιλόγιο (vocabulary) στις λέξεις (features) που εμφανίζουν την μεγαλύτερη συχνότητα στο corpus. Η χρήση στενότερου λεξιλογίου πιστεύω ότι θα μειώσει την εξειδίκευση στα κείμενα εκπαίδευσης και κατά συνέπεια το overfitting. Δεδομένου επιπλέον της προεπεξεργασίας TF-IDF, η οποία υποβάθμισε τις συχνές λέξεις χαμηλής διακριτικής αξίας, πιστεύω ότι η παράμετρος αυτή θα ξεχωρίσει αποτελεσματικά τις λέξεις που θα συμβάλλουν στην αποτελεσματικότερη ταξινόμηση των εγγράφων. Δοκιμάστηκαν διάφορες τιμές παραμέτρου κυμαινόμενες από 500 έως και 15000.
- **ngram_range**: Καθορίζει το μέγεθος των διάφορων n-grams, που εξάγονται ως features από τα κείμενα εισόδου, παράμετρος που είναι γνωστό ότι επιδρά στην ακρίβεια κατηγοριοποίησης των εγγράφων. Δοκιμάστηκαν όλοι σχεδόν οι δυνατοί συνδυασμοί (min, max) από (1,1) έως και (5,5)

Οι παραπάνω μετρήσεις έγιναν χρησιμοποιώντας τιμή παραμέτρου alpha=0.1. Λόγω του όγκου των δεδομένων δεν θα τα παρουσιάσω αναλυτικά. Κατέληξα στο συμπέρασμα ότι η ζητούμενη από την εκφώνηση απόδοση F1-score~0.70 μπορεί να επιτευχθεί ικανοποιητικά με διάφορους συνδυασμούς τιμών παραμέτρων. Ο γενικός κανόνας που φαίνεται να εμφανίζεται είναι ότι όσο αυξάνει (μέσα σε κάποια όρια) το μέγεθος των εξαγόμενων n-grams τόσο μπορεί (ή πρέπει) να αυξηθεί και το μέγεθος του χρησιμοποιούμενου λεξιλογίου. Κατέληξα στους παρακάτω συνδυασμούς τιμών:

(max_features, ngram_range): (1000,(1,1)), (5000,(2,2)), (12500,(3,3)).

Τέλος κατασκεύασα το Confusion Matrix που αποδίδει ο αλγόριθμος, τόσο με τις παραπάνω παραμετροποιήσεις, όσο και στην default διαμόρφωσή του, πάντα με τιμή alpha=0.1. Γίνεται ξεκάθαρα φανερή η εμφάνιση σκοτεινών περιοχών μακρυά από την κύρια διαγώνιο, στην παραμετροποιημένη εκδοχή του αλγορίθμου. Οι παραπάνω υλοποιήσεις του αλγορίθμου, καθώς και η κατασκευή των γραφημάτων, έγινε σε γλώσσα προγραμματισμού Python και ο κώδικας φαίνεται στο συνοδευτικό αρχείο *lliapikos_ge4_NB.py*.

ΒΙΒΛΙΟΓΡΑΦΙΑ

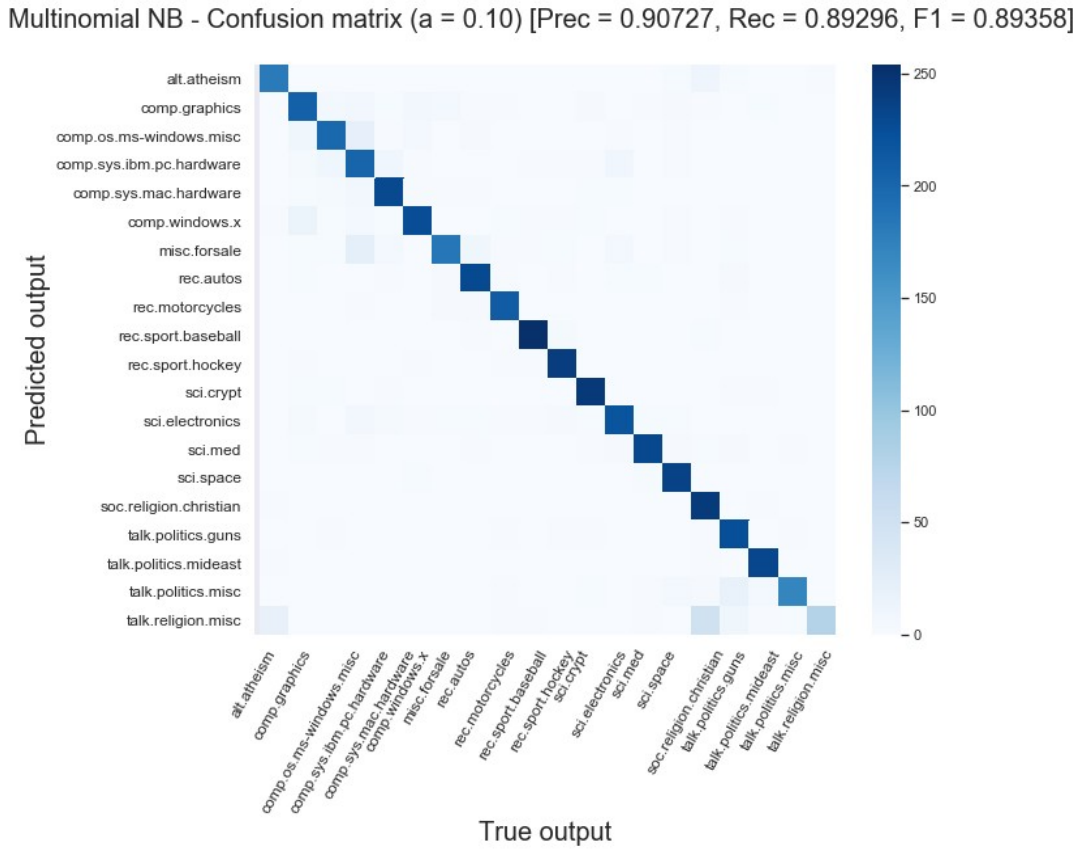
1. Υλικό μαθήματος.

2. Müller, A.C & Guido, S. - Introduction to Machine Learning with Python. A Guide for Data Scientists, O' Reilly, 2017.

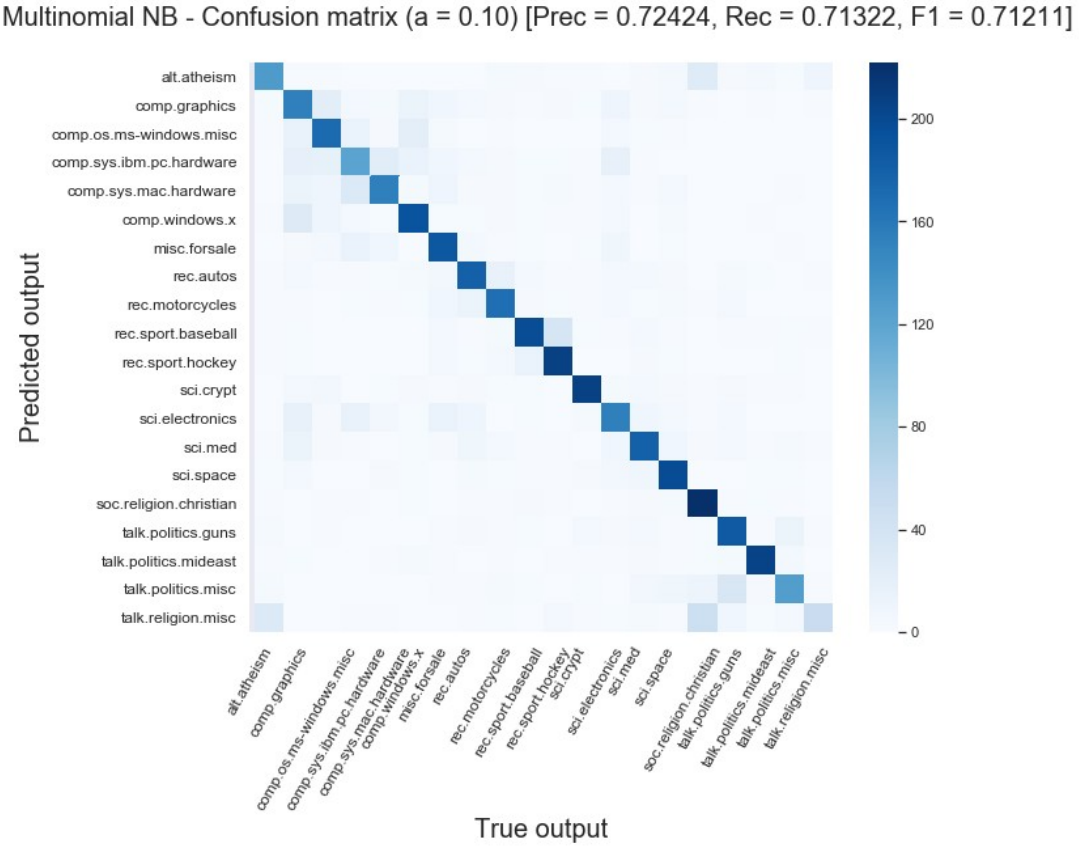
3. Τεκμηρίωση από τον ιστότοπο της βιβλιοθήκης Sklearn.

CONFUSION MATRIX ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

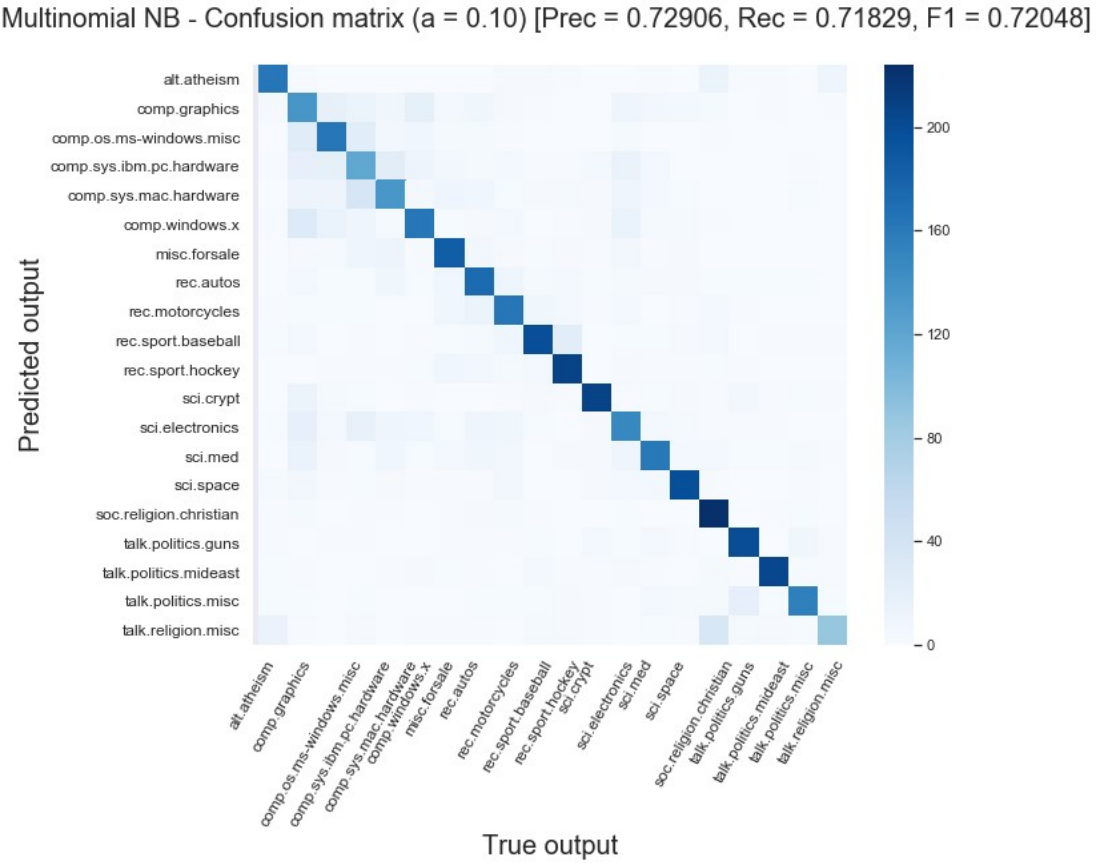
Default διαμόρφωση αλγορίθμου:



Παραμετροποιημένος αλγόριθμος: max_features=1000, ngram_range= (1,1)



Παραμετροποιημένος αλγόριθμος: max_features=5000, ngram_range= (2,2)



Παραμετροποιημένος αλγόριθμος: max_features=12500, ngram_range= (3,3)