Επεξεργασία Φυσικής Γλώσσας – 2022/20223 - Ο.Π.Α

**Εργασία 2**

Φοίβος-Αστέριος Νταντάμης (f3312204 ) ([phivos93@yahoo.com](mailto:phivos93@yahoo.com))

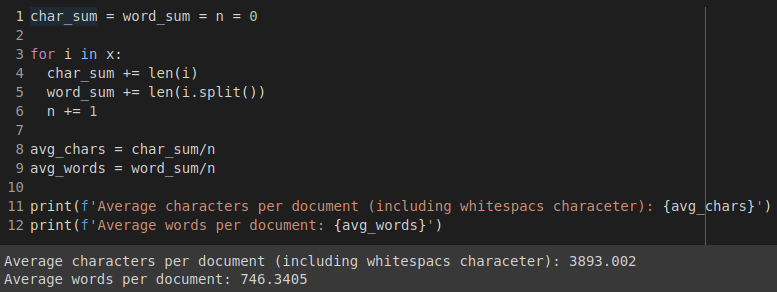
Παναγιώτης Καραστατήρης () ([pan.karastatiris@gmail.com](mailto:pan.karastatiris@gmail.com))

**Άσκηση 15**

**notebook url: https://colab.research.google.com/drive/19BDPgAe6xb88vuniNTEN017Ao5E0Bs6Q#scrollTo=DQJz-3ArZXIg**

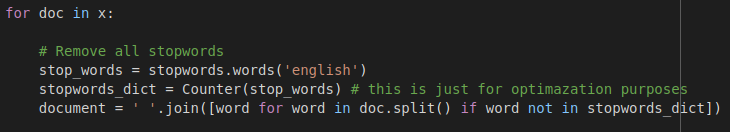
**Περιγραφή Dataset (Φοίβος Νταντάμης)**

Ζητείται να φτιαχτεί ένας sentiment classifier για ένα dataset με δύο κλάσεις. Το dataset που χρησιμοποιήθηκε είναι το Cornell Movie Review Data από το ομώνυμο πανεπιστήμιο. Το dataset αυτό περιέχει κριτικές ταινιών από το αρχείο του IMDb κατηγοριοποιημένες ως θετικές(pos) ή αρνητικές(neg). Το μέγεθος του συνόλου όλων των κριτικών είναι 2000, στις οποίες οι 1000 ανήκουν στην κατηγορία pos και οι υπόλοιπες 1000 στην κατηγορία neg. Ο μέσος όρος χαρακτήρων ανά κριτική είναι 3893.002, ενώ ο μέσος όρο των λέξεων είναι 746.3405.

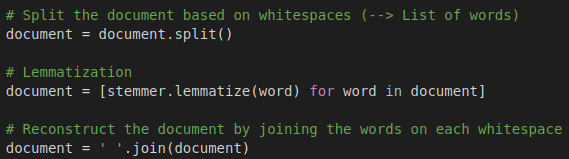


**Προεπεξεργασία (Φοίβος Νταντάμης)**

Το πρώτο βήμα που εφαρμόζουμε είναι να αφαιρέσουμε τα stopwords της αγγλικής γλώσσας, φιλτράροντας όλα τα κείμενα για τις λέξεις που θεωρεί το nltk ως stopwords.



Στη συνέχεια με τη βοήθεια του WordNetLemmatizer αντικαθιστούμε όλες τις λέξεις με τα λήμματά τους.



Συνεχίζοντας, μετατρέπουμε όλους τους χαρακτήρες σε μικρούς.



Σε αυτήν την φάση έχουν προκύψει λέξεις οι οποίες ανήκουν στα stopwords, οπότε θα ξαναφιλτράρουμε τα κείμενα.



Έπειτα αφαιρούμε τα σημεία στίξης και τα σύμβολα που ουσιαστικά δεν αποτελούν κάποια λέξη.



Αφαιρούμε όλες τις λέξεις που αποτελούνται από έναν χαρακτήρα μόνο, μιας και δεν προσφέρουν κάποια ιδιαίτερη πληροφορία για την κατηγοριοποίηση.



Σε αυτό το σημείο και πάλι έχουν προκύψει κάποιες λέξεις που είναι stopwords οπότε φιλτράρουμε εκ νέου.



Αφαιρούμε τώρα όλους τους αριθμούς

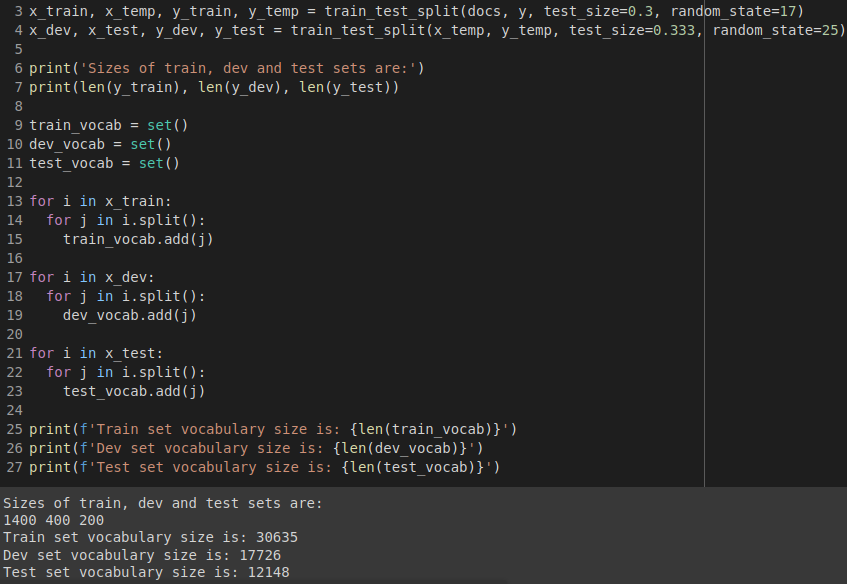


Αφαιρούμε τέλος τα honorifics



**Διαχωρισμός Dataset (Φοίβος Νταντάμης)**

Σπάσαμε το Dataset σε Train, Development και Test sets με ποσοστά 70%, 20%, 10% αντίστοιχα. Προκύπτουν έτσι το Train Dataset με 1400 κριτικές και μέγεθος λεξιλογίου 30635. Το Development Dataset με 400 κριτικές και μέγεθος λεξιλογίου 17726. Το Testing Dataset με 200 κριτικές και μέγεθος λεξιλογίου 12148.

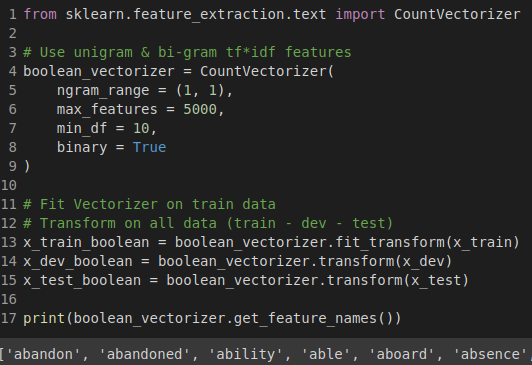


**Feature Exctraction (Φοίβος Νταντάμης)**

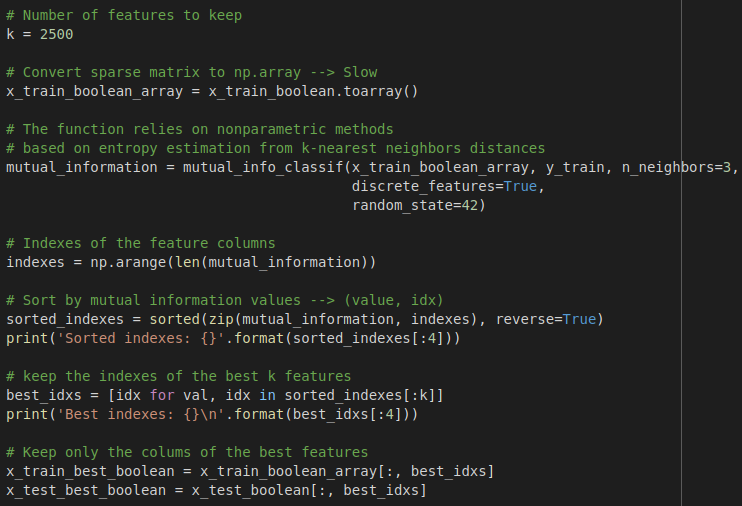
Προκειμένου να μετατρέψουμε τα κείμενα σε διανύσματα χρησιμοποιήσαμε δύο τρόπους. Ο πρώτος είναι δημιουργώντας Boolean Features και ο δεύτερος TF-IDF features.

Boolean Features:

χρησιμοποιούμε τον CountVectorizer με binary times. Κρατάμε μόνο τα μονογράμματα και μόνο αυτά με συχνότητα εμφάνισης τουλάχιστον 10. Τέλος κρατάμε τα 5000 πιο συχνά εμφανιζόμενα μονογράμματα.

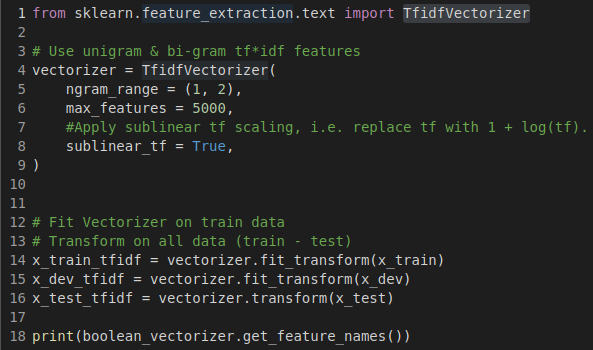


Στη συνέχεια με κριτήριο το Information Gain(Mutual Info) από τα 5000, κρατάμε τα 2500 καλύτερα features.



TF-IDF Features:

χρησιμοποιούμε τον TfidfVectorizer. Κρατάμε τα μονογράμματα και τα διγράμματα και μόνο αυτά με συχνότητα εμφάνισης τουλάχιστον 10. Τέλος κρατάμε τα 5000 πιο συχνά εμφανιζόμενα μονογράμματα ή διγράμματα.



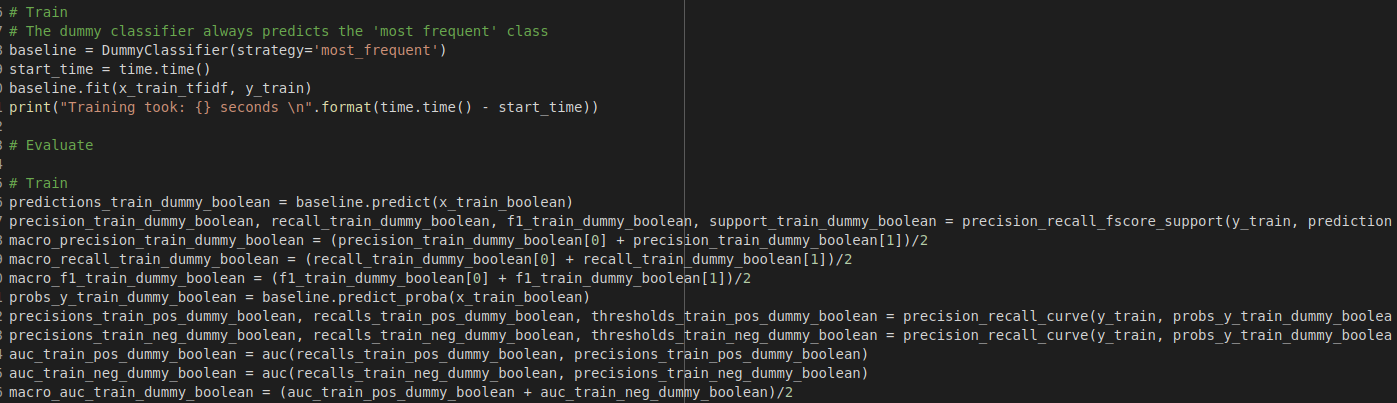
Στη συνέχεια με κριτήριο το Information Gain(Mutual Info) από τα 5000, κρατάμε τα 2500 καλύτερα features με τον ίδιο ακριβώς τρόπο όπως και στα boolean features.

**Classifiers (Φοίβος Νταντάμης)**

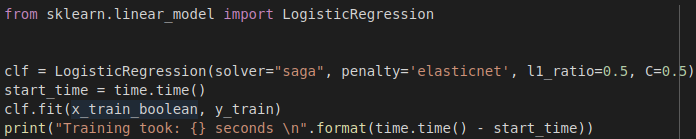
Για την κατηγοριοποίηση των διανυσμάτων μας χρησιμοποιούμε αρχικά μόνο για σκοπούς σύγκρισης τον Dummy Classifier. Στη συνέχεια χρησιμοποιούμε τον LogisticRegression, μία φορά με τα boolean features και μία φορά με τα tfidf features. Υπολογίζουμε για κάθε περίπτωση τα precision, recall, f1, auc, macro precision, macro recall, macro f1 και macro auc για κάθε set από τα train, dev και test χωριστά.

Dummy Classifier:

Για τον DummyClassifier χρησιμοποιούμ την στρατηγική επιλογής της πιο συχνά εμφανιζόμενης κλάσης

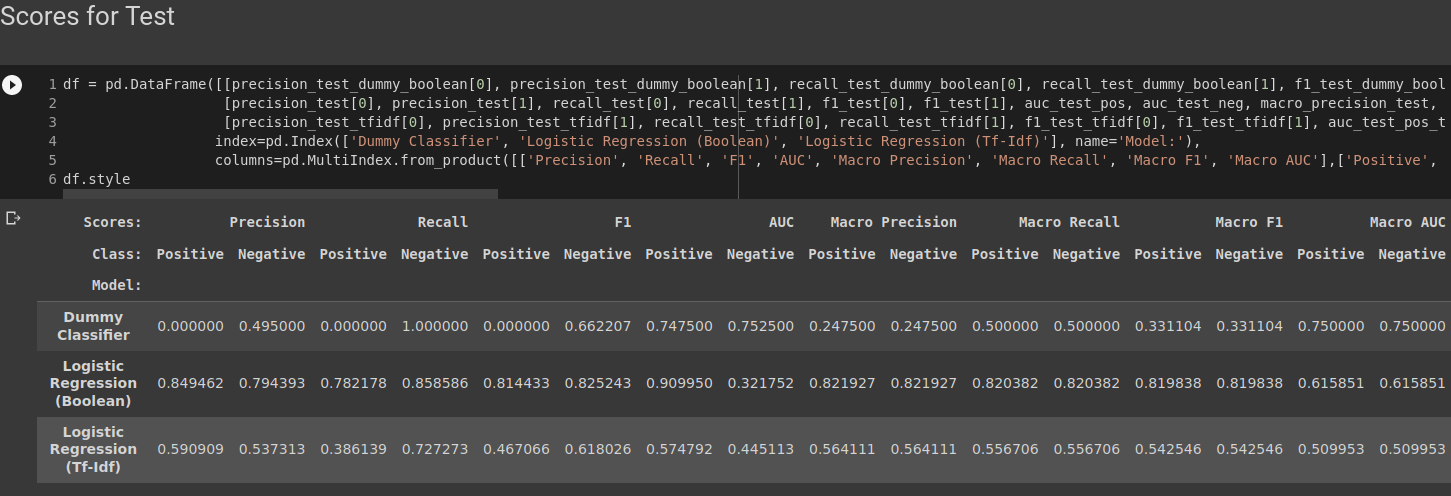


Για τον LogisticRegression χρησιμοποιούμε τον saga solver για να επιλέξουμε το elasticnet ως penalty, ώστε να μπορούμε να ορίσουμε έναν ή και τους δύο τύπους regularization l1 και l2, ορίζοντας κατάλληλα το ratio μεταξύ τους.



**Απεικόνιση αποτελεσμάτων (Φοίβος Νταντάμης)**

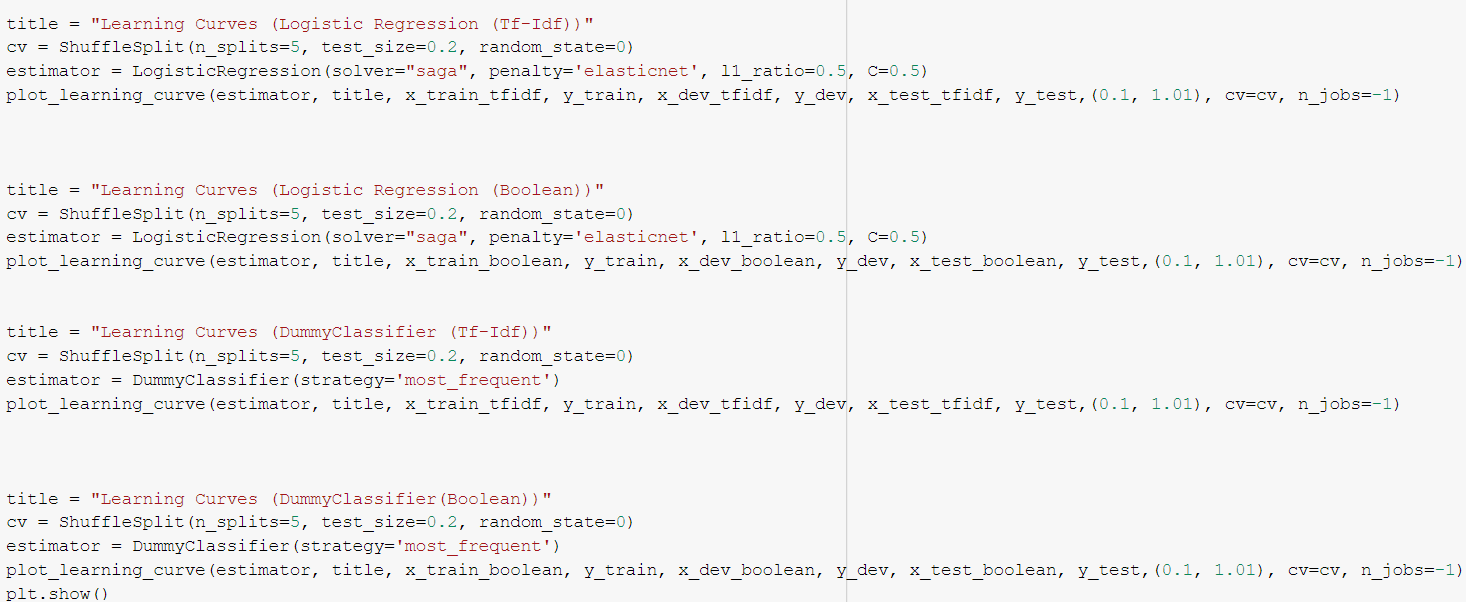
Τελικά αποτυπώνουμε όλα τα παραπάνω σε έναν πίνακα. Στις στήλες των macro μεγεθών δεν υπάρχει νόημα στον διαχωρισμό βάση των κλάσεων και για αυτό τα αποτελέσματα είναι τα ίδια και στις δύο στήλες κάθε φορά για ευκολία υλοποίησης.



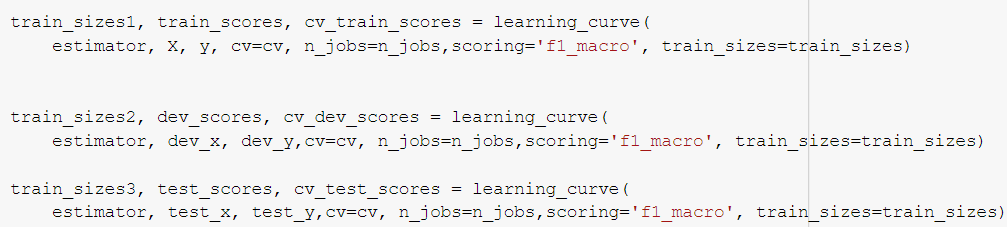
**Learning Curves (Παναγιώτης Καραστατήρης)**

Για να φτιάξουμε για κάθε classifier τις 3 καμπύλες εκμάθησης στα 3 ζητούμενα dataset χρησιμοποιήσαμε τη συνάρτηση plot\_learning\_curve. Την καλούμε 4 φορές, 2 φορές για κάθε classifier (linear και dummy) διότι έχουμε 2 εναλλακτικούς vectorizer, τον Boolean και τον TfidfVectorizer.

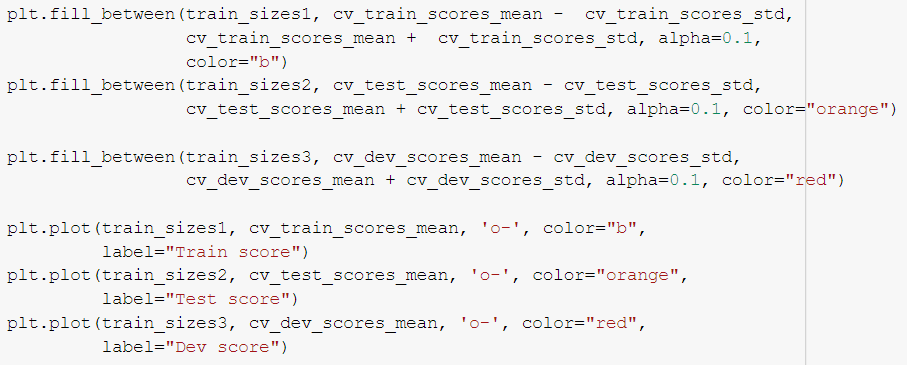
plot\_learning\_curve : Της δίνουμε σαν ορίσματα κάθε φορά τα x\_tfidf ή τα x\_boolean στο train,dev,test dataset με τα αντίστοιχα y.

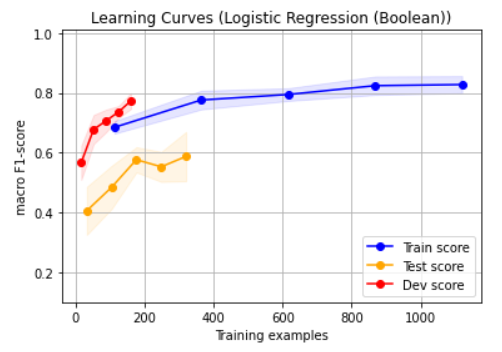
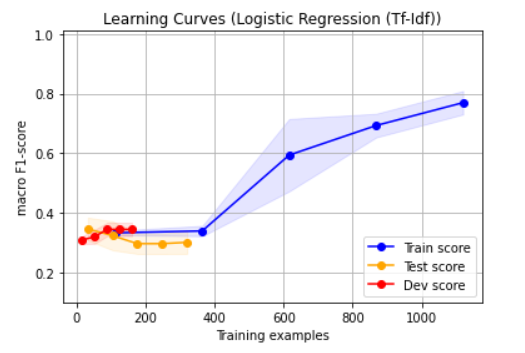


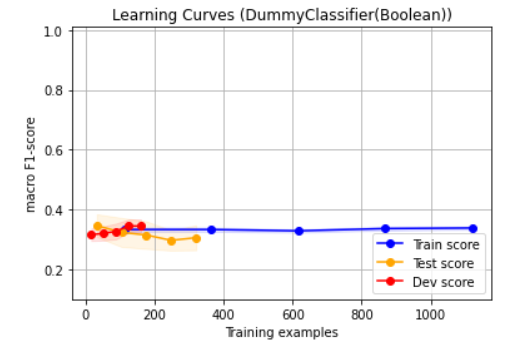
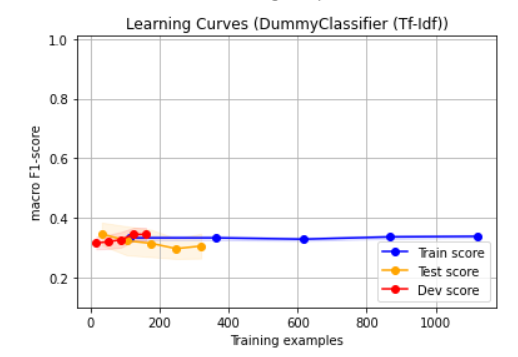
Εσωτερικά καλούμε 3 φορές την learning\_curve μια φορά για κάθε dataset με τα αντίστοιχα ορίσματα και scoring='f1\_macro' όπως και ζητείται.



Με τα scores στη συνέχεια φτιάχνουμε τα αντίστοιχα mean και std scores τα οποία χρησιμοποιούν η plt.fill\_between και η plt.plot.



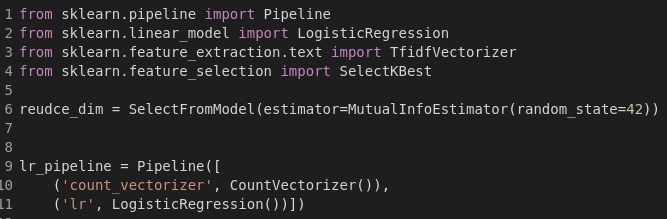




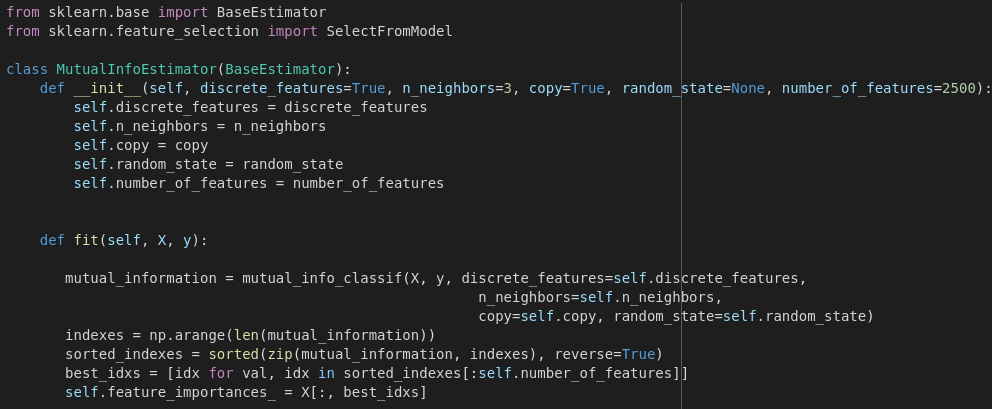
**Hypermaremeter tuning (Φοίβος Νταντάμης)**

Το ερώτημα αυτό είναι ημιτελές.

Προκειμένου να διαλέξουμε τις τιμές των υπερπαραμέτρων c, l1\_ratio αλλά και τον αριθμό των features, που κρατάμε θα έπρεπε να ψάξουμε σε όλο τις διαστάσεις του προβλήματος αυτού και να διαλέξουμε τις καλύτερες τιμές. Προκειμένου να βρούμε τις τιμές αυτές, αρχικά φτιάχνουμε ένα pipeline το οποίο θα τροφοδοτήσουμε σε κάποια συνάρτηση αναζήτησης βέλτιστης λύσης.



Στο pipeline θα θέλαμε εκτός από τη διαδικασία του vectorizing και του solving να έχουμε και τη διαδικασία του feature reduction μέσω mutual\_information. Δεν υπάρχει αυτή η διαδικασία στο sklearn οπότε έπρεπε να φτιάξουμε έναν δικό μας estimator που θα χρησιμοποιεί το mutual\_info\_classif και θα επιστρέφει τα καλύτερα features στο επόμενο stage του pipeline



Στη συνέχεια ο RandomSearch θα δεχόταν τα εύρη των παραμέτρων στα οποία αναζητάμε λύση και θα έβγαζε τον καλύτερο συνδυασμό. Δυστυχώς κάτι σκάει σε αυτήν τη διαδικασία.

