**Μεθοδολογία**

Συλλογή δεδομένων: Γινεται η συλλογη των δεδομενων των κειμενων που εχει περιεχομενο σχολιασμους με συναισθηματα ,και αυτό το συνολο θα χρησιμοποιηθεί για την εκπαίδευση και την αξιολόγηση του μοντέλου εκπαιδευσης.

Προεπεξεργασία δεδομένων: Προ-επεξεργαζονται τα δεδομενα και γινεται η αφαιρεση των stop-words

Εξαγωγή δυνατοτήτων: Λαμβανει χωρα το tokenization και το tfidf δημιουργωντας το bag of words.

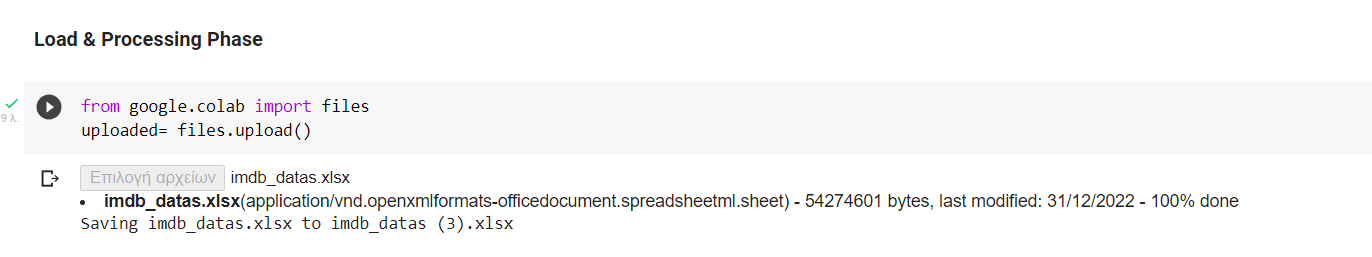
Επιλογή μοντέλου: Επιλογη ενος μοντέλου εκμάθησης που είναι κατάλληλο για την ανάλυση του συναισθήματος. Τα μοντέλο για την ανάλυση συναισθήματος που χρησιμοποιηθηκε είναι το Support Vector Machines (SVM) και το νευρωνικο δικτυο Keras.

Εκπαίδευση μοντέλου: Εκπαιδευση του μοντέλου εκμάθησης στα προεπεξεργασμένα δεδομένα και στους αντίστοιχους σχολιασμούς συναισθημάτων(tokens και υστερα bag of words).

Αξιολόγηση μοντέλου: Αξιολογηση του μοντελου για την απόδοση του σε ένα σύνολο δεδομένων δοκιμής. Αυτό μπορεί να γίνει χρησιμοποιώντας μετρήσεις όπως η ακρίβεια και ο χρονος

**ΣΤΑΔΙΟ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ**

Πρωτα ανεβασαμε το csv στον notebook μεσω αυτής της εντολης :



Υστερα θεσαμε ολες τις βιβλιοθηκες σε ένα κελι και τις ενσωματωσαμε στον κωδικα μας κανοντας τες import.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Στο επομενο βημα εκτυπωσαμε τα πρωτα στοιχεια του excel ώστε να εχουμε μια πρωτη ιδεα για τα δεδομενα που κατεχουμε και για το πως θα πρεπει να κινηθουμε στα επομενα βηματα μας.



Στο επομενο βημα αλλαξαμε τα ονοματα των τιτλων τα από label -> review

& file ->sentiment και κρατησαμε μονο αυτά τα 2 columns γιατι τα υπολοιπα ηταν περιττα.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Με αυτή την χρηση των εντολων που καναμε είναι για να παρατηρησουμε αν το dataset μας είναι balanced ,το οποιο το παρατηρουμε αν είναι σχετικα ισοποσα το πληθος της κάθε τιμης δηλαδη το positive και το negative κυμαινεται το πληθος τους στις 21.000 + ενώ το πληθος του unsup εχει πληθος ανω του 40.000.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Υστερα ξανακανουμε ένα describe του dataset για να δουμε το γενικο του summary. Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Στο επομενο βημα μας υστερα καναμε balance το πληθος του Positive του Negative και του Unsup αφου πρωτα τις ξεχωρισαμε από το πληθος και πηραμε την unsup και την θεσαμε στην τιμη πληθος 22933 το οποιο επιλεχθηκε τυχαια και πληρει το κριτηριο που είναι τιμη κοντα στο 20.000+. Και μετα ενωσαμε αυτές τις τρεις υποκατηγοριες του ‘sentiment’ σε ένα νέο data set.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Και οριστε το νέο sample toy unsup εχοντας το πληθος 22933.

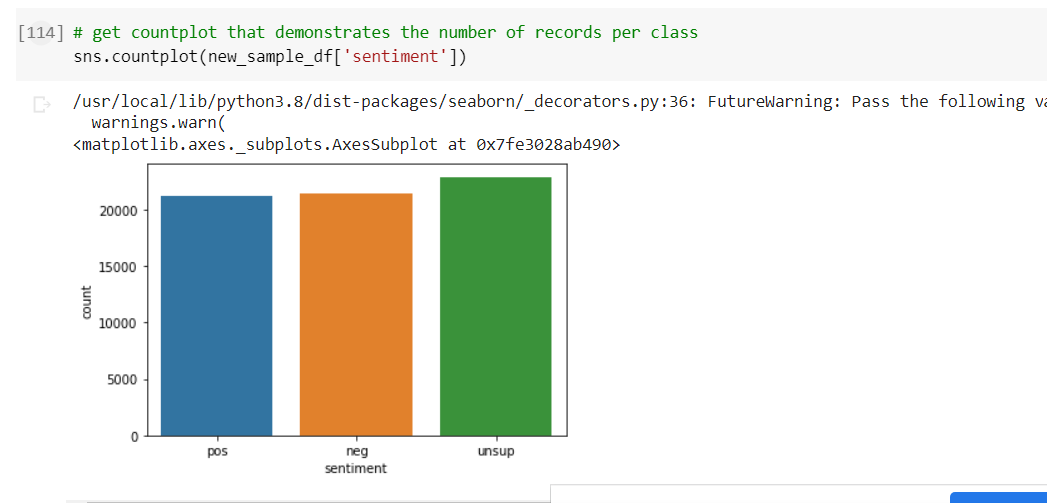
Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Και ετσι πια το dataset μας είναι balanced, το οποιο παρακατω το εχουμε κανει Plot τις τιμες ώστε να φανει η σχετικη ισοποση κατανομη των στοιχειων σε αυτές τις τρεις υποκατηγοριες.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα



Τωρα μπαινουμε στην φαση οπου σπαμε το dataset σε training kai test dataset , αφου πρωτα κανουμε shuffle τα data δηλαδη να τα μπερδεψουμε μεταξυ τους . Και θετω το train dataset σε 0.9 και το test data set σε 0.1 δηλαδη το 10 τοις εκατο.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Στο επομενο μας βημα θα λαβει χωρα η επεξεργασια του dataset και πιο συγκεκριμενα θα θα παμε να αφαιρεσουμε τα stop-words δηλαδη λεξεις όπως το ‘the’ το ‘is’ και αλλα τετοια παρεμφερη και θα χρησιμοποιησουμε τον RegexpTokenizer.Ο οποιος tokenizer μας βοηθα να χωρισουμε το κειμενο σε μικροτερα κομματια τα οποια ονομαζονται tokens ,ώστε αργοτερα να μπορεσουμε να επεξεργαστουμε το κειμενο μας.

Επισης για να μπορεσουμε να υλοποιησουμε ολες αυτές τις λειτουργιες δημιουργησαμε μια συναρτηση την review processing η οποια πρωτα μετατρεπει οσα εφαλαια γραμματα βρισκει σε μικρα γραμματα και υστερα κανει το tokenization στο κειμενο.

Επισης είναι εξυπηρετικη η χρηση της δημιουργιας ενός dictionary το οποιο αντιστοιχιζει την κάθε υποκατηγορια με ένα αριθμο και πιο συγκεκριμενα το positive : 0 negative : 1

Unsup: 2 . Οπου μετα με την χρηση της συναρτησης transform label encoding μετατρεπουμε ετσι τις υποκατηγορις

Και μετα με την χρηση μιας επαναληψης αφαιρουμε όλα τα stop-words. Τα οποια θεωρουνται περιττα για είναι αρκετα frequnt και ετσι αλλοιωνουν την ουσια της διαδικασιας η οποια είναι να βγαλουμε ένα μοντελο για την αναλυση του συναισθηματος οι οποιες λεξεις επισης δε μπορουν να υποδηλωσουν συναισθημα.Οποτε τις αφαιρουμε.

Εικόνα που περιέχει κείμενο

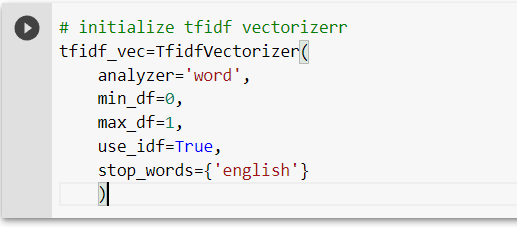
Περιγραφή που δημιουργήθηκε αυτόματα

Και στο επομενο βημα αυτοο που κανουμε είναι ότι επεξεργαζομαστε το train και το test dataset με το να τους κανουμε apply τις δυο αυτές συναρτησεις ώστε υστερα να χουμε το τελικο αποτελεσμα που είναι τα tokens διχως stop-words Και όλα τα γραμματα να ναι μικρα.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Υστερα θα αρχικοποιησουμε ένα vectorizer και πιο συγκεκριμενα τον tfidf vectorizer ,που εχουμε ως απωτερο σκοπο την χρηση αυτου του εργαλειου είναι μετατρεπει το κειμενο σε αριθμητικα χαρακτηριστικα τα οποια αντικατοπτριζουν την εννοια της καθες λεξης δηλαδη φτιαχνει ένα array από νουμερα τα οποια αντιστοιχουν σε μια λεξη και την εννοια της.



Υστερα αυτό που θα κανουμε στο train και το test dataset μας είναι να ενσωματωσουμε την λειτουργια του vectorizer ώστε να φτιαξουμε το τελικο Bag of Words μας το οποιο είναι το συνολο των λεξεων που αντικατοπτριζουν την κάθε κριτικη το οποιο Bag of Words εχει υποστει μετατροπη και είναι μεταφρασμενο σε αριθμους.Την οποια λειτουργια θα την υλοποιησουμε και στις δυο κατηγοριες sentiment και review.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Υστερα αναγνωριζουμε το feature του test set και του training set και επισης θετουμε το πληθς των διαφορετικων ετικετων που χρησιμοποιουμε δηλαδη το positive , negative ,unsup (3) . Το οποιο feature εχει την τιμη 47494. Το οποιο feature σειχνει την ποσοτικοποιηση ολων των λεξεων .

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

**ΣΤΑΔΙΟ ΕΚΠΑΙΔΕΥΣΗΣ**

Πρωτα θα αρχικοποιησουμε τον SVM classifier και θα βαλουμε σαν εισοδους τα τελικα διαμορφωμενα train και test set μας ως εισοδοι στο svm οπου εκει κιολας θα μετρησουεμ και τον χρονο που χρειαστηκε για να ολοκληρωθει η ολη διεργασια του και ο χρονος ηταν 9 minutes και 5 seconds.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

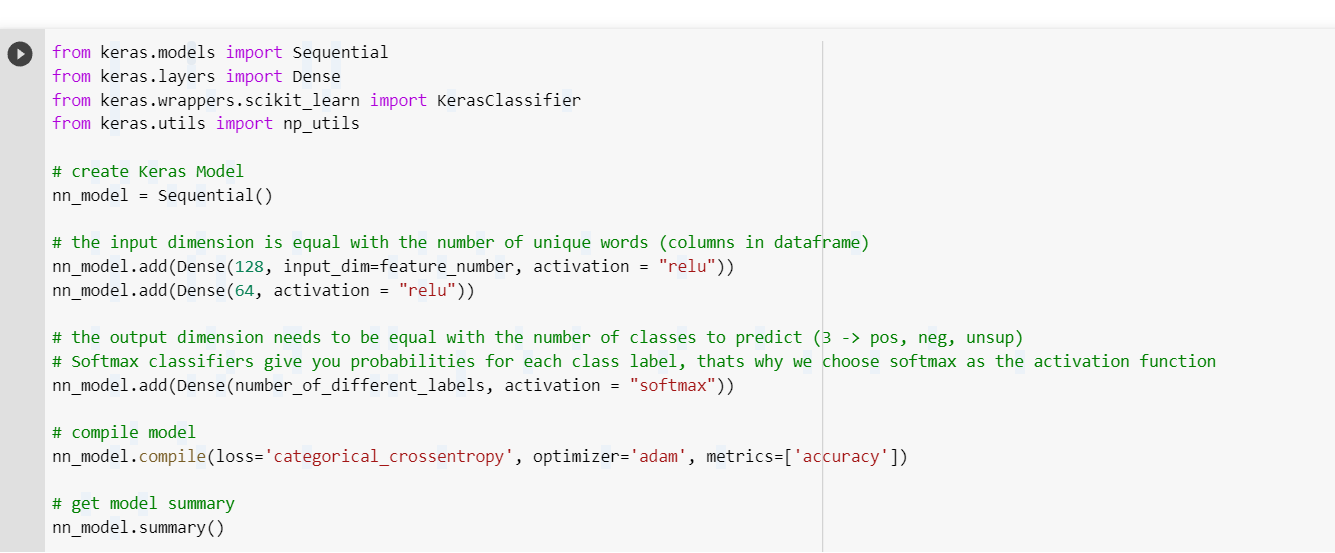
Υστερα θα υλοποιησουμε το νευρωνικο δικτυο Keras

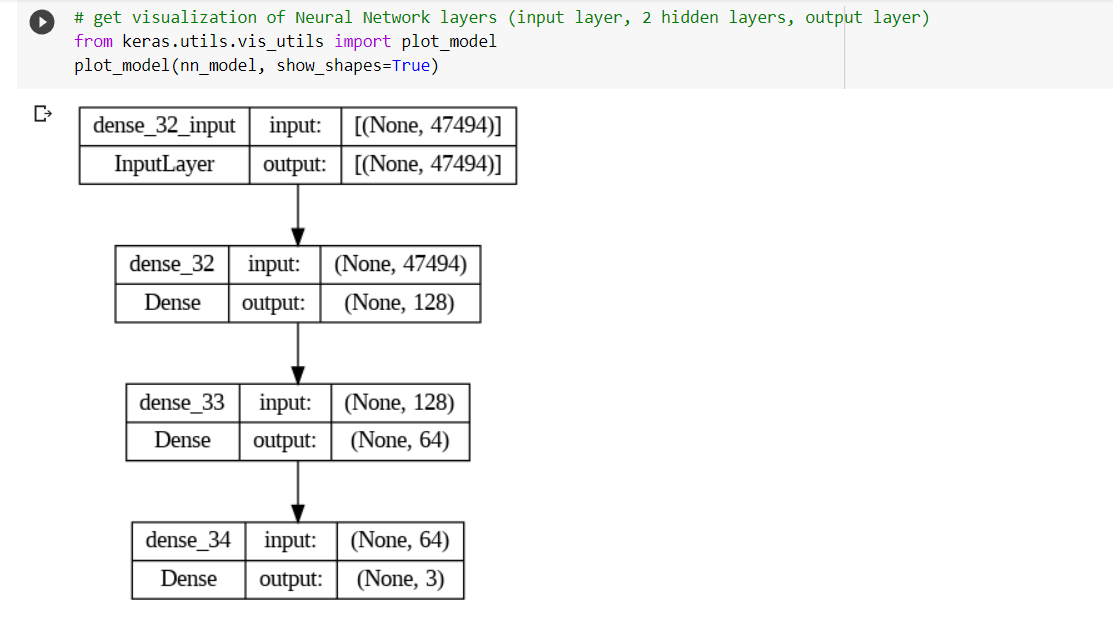
Και αυτή η υλοποιηση θα γινει μεσα από πρωτα την ενσωματωση βασικων βιβλιοθηκων του keras και ετσι πρωτα θα φτιαξουμε το γενικο μοντελο του νευρωνικου δικτου.Στο επομενο βημα θα θεσουμε ότι το input dimension μας είναι ισο με τον αριθμο των unique words δηλαδη οσο τα features =47494

Και βεβαια το output dimension πρεπει να είναι ισο με το πληθος των κατηγοριων που στοχευουμε να χαρακτηριζουν την προβλεψη μας δηλαδη ισο με 3 λογω του pos,neg,unsup.

Το softmax classifier που χρησιμοποιησαμε στο νευρωνικο δικτυο είναι δινει την πιθανοτητα για κάθε κατηγορια ,για αυτό τον λογο την επιλεξαμε ως activation function.

Υστερα θα κανουμε compile to μοντελο και θα ζητησουμε το summary του μοντελου που δημιουργησαμε.



Υστερα, θα κανουμε ένα visualization toy νευρωνικου δικτυο για να ναι πιο κατανοησιμη η υποδομη του. Εχουμε δηλαδη ένα input layer δυο hidden layer και ένα output layer. 

Αφου δημιουργησαμε το νευρωνικο δικτυο ηρθε η ωρα να δοκιμασουμε την επιδοση του και να παρατηρησουμε τον ποσο χρονο χρειαζεται για να ολοκληρωθει εχοντας 10 epochs.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Και θα χρειαστει 1934 δευτερολεπτα κοντα στα 36 λεπτα δηλαδη. Και παρατηρουμε ότι οσο περναει και ολοκληρωνεται ο ενας κυκλος ολο και τοσο αυξανεται η επιδοση του νευρωνικου δικτυου αλλα θα ηταν προτιμοτερη η χρηση του epoch να βρισκοταν στο νουμερο 6 οπου και κατεχει εκει την καλυτερη επιδοση πριν ξεκινησει να είναι ελαφρως πτωτικη η αποδοση του σχεδον μηδαμινη βεβαια.

**ΣΤΑΔΙΟ ΑΞΙΟΛΟΓΗΣΗΣ**

Αφου υλοποιησαμε τον SVM ηρθε η ωρα να τον αξιολογησουμε και να δουμε την επιδοση του και κυριως να δουμε τις την επιδοση του στις προβλεψεις που θα κανει στο test set το οποιο ειχα εξάρχης στος πληθος των 5500 κριτικων οι οποιες είναι randomly choised κριτικες και μεσα σε ολο αυτό το πλαισιο βρισκω και τον χρονο που χρειαζεται ο SVM για να ολοκληρωθει.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Οπου χρειαστηκαν 34.5 seconds για το SVM και στην συνεχεια θα χρησιμοποιησουμε το confusion matrix για να δουμε τα macro accuracys του SVM στα label του pos,neg,unsup.

Και στην συνεχεια θα κανω οπτικοποιησω το confusion matrix του SVM CLASSIFICATION



Ιδια διαδικασια ελαβε χωρα και η αντιστοιχη κατασταση με το NEURAL NETWORK

Οπου εβαλα 5500 τυχαια δειγματα κριτικων τα οποια τα ειχα εξάρχης από το test set μου λογω του shuffle Που χα κανει στην αρχη. Και βρισκω τον χρονο που χρειαστηκε για την ολοκληρωση του αλλα και το accuracy του .

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Και στην συνεχεια βρηκα και οπτικοποιησα το confusion matrix του .

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα