# Οδηγίες χρήσης του κώδικα

Ο φάκελος που περιέχει τον κώδικα, περιλαμβάνει τους εξής υποφακέλους:

* Data: Περιέχει τα δεδομένα εκπαίδευσης και τα pickles
* Greek\_stemmer: Περιέχει το πακέτο Greek\_stemmer που χρησιμοποιείται για stemming των λέξεων κατά το preprocessing.
* Img: Περιέχει τις εικόνες οι οποίες γίνονται export από τα διάφορα διαγράμματα. Οι περισσότερες βρίσκονται σε μορφή .eps
* Models: Περιλαμβάνει ένα σύνολο από μοντέλα που υλοποιήθηκαν για τους σκοπούς του project. Τα μοντέλα αυτά είναι είτε μοντέλα classification (GMM, KNN,MEAN, NB) είτε μοντέλα μείωσης διαστάσεων (LDA,PCA)
* Plotting: Πακέτο που περιέχει τον κώδικα για το evaluation μοντέλων με διαφορετικές παραμέτρους μέσω διαφόρων γραφημάτων.
* Preprocessor: Πακέτο που περιέχει όλη τη διαδικασία του preprocessing και του feature extraction, καθώς επίσης και της κλήσης των κατάλληλων μοντέλων για εκπαίδευση.

Τα βασικά αρχεία διεπαφής με τον χρήστη είναι 2:

1. **run\_single\_model.py:** Αρχείο που επιτρέπει στον χρήστη να τρέξει κάποιο συγκεκριμένο μοντέλο και να δει την απόδοση για το μοντέλο αυτό. Κατά την εκτέλεση του αρχείου αυτού, υπάρχει Menu το οποίο καθοδηγεί το χρήστη για την επιλογή μετασχηματισμών, μεθόδων μείωσης διαστάσεων και μοντέλων κατηγοριοποίησης. Επίσης, αξίζει να αναφερθεί ότι για την αλλαγή των default παραμέτρων με τις οποίες τρέχουν τα μοντέλα, αρκεί να τροποποιηθεί το dictionary param\_dict το οποίο βρίσκεται στο αρχείο αυτό και διαθέτει τις παραμέτρους όλων των μοντέλων. Τέλος, για την ταχύτερη προεπεξεργασία και εκτέλεση των μοντέλων, έχουν χρησιμοποιηθεί pickles τα οποία μπορεί ο χρήστης να επιλέξει να χρησιμοποιήσει μέσα από το menu.
2. **model\_param\_plot.py:** Διαθέτει επίσης μενού διεπαφής με τον χρήστη και επιτρέπει τις εξής ενέργειες:
   1. Απεικόνιση των πιο σημαντικών bigrams σε μορφή εικόνας και αποθήκευση της εικόνας σε αρχείο.
   2. Απεικόνιση των πιο σημαντικών όρων σε μορφή εικόνας και αποθήκευση της εικόνας σε αρχείο.
   3. Δοκιμή των διαφόρων μοντέλων για διαφορετικές παραμέτρους και κατασκευή διαγράμματος που απεικονίζει την απόδοση για τις παραμέτρους αυτές. Τα μοντέλα τα οποία δοκιμάζονται αναγράφονται στη λίστα *methods* του αρχείου **model\_param\_plot** και οι παράμετροι που ελέγχονται για κάθε μοντέλο, μπορούν να τροποποιηθούν από τις αντίστοιχες συναρτήσεις του αρχείου plotting/\_\_init\_\_.py. Παράδειγμα για την τροποποίηση των παραμέτρων που δοκιμάζονται για το ΚΝΝ θα πρέπει να πάμε στη συνάρτηση ΚΝΝ του αρχείου plotting/\_\_init\_\_.py. Τέλος, για την τροποποίηση των μεθόδων που χρησιμοποιούνται κατά το preprocessing (feature transformation, dimensionality redunction κλπ) πρέπει να τροποποιήσουμε τη μέθοδο preprocess() του αρχείο plotting/\_\_init\_\_.py.
   4. Εκπαίδευση όλων των μοντέλων χρησιμοποιώντας 5-fold cross validation και κατασκευή συγκεντρωτικού διαγράμματος απόδοσης για όλα τα μοντέλα. Για την τροποποίηση των μοντέλων που χρησιμοποιούνται ή των παραμέτρων τους μπορούμε να τροποποιήσουμε το dictionary *param*\_list το οποίο βρίσκεται στη συνάρτηση run\_all\_kfold του αρχείου plotting/\_\_init\_\_.py. Επίσης, για την τροποποίηση των μεθόδων που χρησιμοποιούνται για το processing, μπορούμε να τροποποιήσουμε τη συνάρτηση \_\_train\_and\_evaluate\_all() του αρχείου plotting/\_\_init\_\_.py.

Σημειώσεις:

* Αν για κάποιο λόγο το πρόγραμμα σκάει λόγω dimension mismatch, κατά πάσα πιθανότητα ευθύνονται τα pickles. Οπότε στην περίπτωση αυτή, επιλέξτε να τρέξετε το πρόγραμμα χωρίς pickles από το menu και το πρόβλημα θα διορθωθεί. Εναλλακτικά, θέστε το όρισμα ignore\_pickles του constructor του Preprocessor σε True (ignore\_pickles=True).
* Στο δεύτερο αρχείο, αν επιλέξετε την επιλογή run\_all\_models\_kfold, τότε by default γίνονται train και evaluate όλα τα μοντέλα, συμπεριλαμβανομένων των ANN και CNN. Όμως, τα ANN και CNN απαιτούν αρκετό χρόνο για εκπαίδευση, ειδικά σε υπολογιστές χωρίς κάρτα γραφικών. Επομένως, αν θέλετε να τα απενεργοποιήσετε για την ταχύτερη εξαγωγή αποτελεσμάτων, απλά βάζετε σε σχόλιο τις γραμμές 201 και 202 του αρχείου plotting/\_\_init\_\_.py.
* Τα διαγράμματα αποθηκεύονται στον φάκελο img/