

Ειδικά Θέματα Επικοινωνιών & Επεξεργασίας Σήματος Ψηφιακή Επεξεργασία και Αναγνώριση Εγγράφων

Αναγνώριση Ελληνικών Χειρόγραφων
Χαρακτήρων με χρήση ζωνών

Rizart Dona
Σεπτέμβριος, 2018

Περιεχόμενα

1. Εισαγωγή	2
2. Μεθοδολογία	2
2.1 Απευθείας Σύγκριση	3
2.2 Ζώνες	4
2.3 Προβολές	5
2.4 Αναδρομικές Υποδιαιρέσεις	5
3. Πειραματικά Αποτελέσματα	6
4. Εγχειρίδιο Εφαρμογής	8

1. Εισαγωγή

Η παρούσα εργασία αφορά τον σχεδιασμό και την υλοποίηση ενός συστήματος κατηγοριοποίησης τυπωμένων πολυτονικών Ελληνικών χαρακτήρων. Για την εξαγωγή των χαρακτηριστικών από τους χαρακτήρες χρησιμοποιούνται οι αποστάσεις Jaccard, Jule (απευθείας σύγκριση) καθώς και οι τεχνικές των προβολών, των ζωνών και των αναδρομικών υποδιαιρέσεων σε συνδυασμό με τον κατηγοριοποιητή του κοντινότερου γείτονα. Η εφαρμογή υλοποιήθηκε σε περιβάλλον Qt και περιλαμβάνει γραφική επιφάνεια εργασίας με την οποία ο χρήστης αλληλεπιδρά με το πρόγραμμα.

Τα επόμενα κεφάλαια είναι διαρθρωμένα ως εξής. Στο δεύτερο κεφάλαιο αναλύεται η μεθοδολογία της διαδικασίας καθώς και οι παράμετροι της. Στο τρίτο κεφάλαιο παρουσιάζονται ορισμένα πειραματικά αποτελέσματα όπου και αναφέρεται ο πίνακας συσχετισμών (confusion matrix). Τέλος, στο τέταρτο κεφάλαιο παρουσιάζεται μέσω εικόνων ένα demo της εφαρμογής και περιγράφεται ο τρόπος χρήσης της καθώς και οι δυνατότητες της.

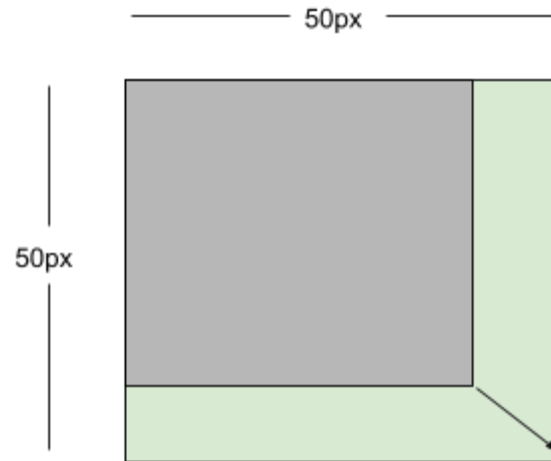
2. Μεθοδολογία

Για την κατηγοριοποίηση των χαρακτήρων χρησιμοποιούνται 2 ξεχωριστές τεχνικές. Η πρώτη αφορά την απευθείας σύγκριση (template matching) και τον υπολογισμό των αποστάσεων Jaccard και Yule. Η δεύτερη τεχνική αφορά την χρήση του κατηγοριοποιητή του κοντινότερου γείτονα (για $n=1$ και με χρήση ευκλείδειας απόστασης) με εξαγωγή χαρακτηριστικών μέσω προβολών, ζωνών και αναδρομικών υποδιαιρέσεων.

Το συνολικό δείγμα των χαρακτήρων αποτελείται από 2764 στοιχεία που είναι το δείγμα εκπαίδευσης (trainset) καθώς και από 2731 στοιχεία που είναι το δείγμα δοκιμής (testset), οι διαφορετικές κλάσεις στις οποίες ανήκουν οι χαρακτήρες μας είναι 67 στο σύνολο.

Οι χαρακτήρες κατά την φόρτωση τους κανονικοποιούνται έτσι ώστε να μπορούμε να εφαρμόσουμε τις απαιτούμενες τεχνικές. Η κανονικοποίηση τοποθετεί τα στοιχεία σε ένα ενιαίο πλαίσιο 50×50 (το οποίο υπολογίστηκε πως καλύπτει όλα τα στοιχεία επαρκώς) όπου στα επιπλέον pixel ανατίθεται η τιμή 0. Τα επιπλέον pixel τοποθετούνται επιμηκύνοντας τον πίνακα του στοιχείου ανατολικά κατά τον άξονα του πλάτους και νότια κατά τον άξονα του ύψους, αυτό συμβαίνει έτσι ώστε να μην αλλοιωθεί η κυρίως πληροφορία του χαρακτήρα. Η παρακάτω εικόνα αποτυπώνει

σχηματικά την κανονικοποίηση ενός χαρακτήρα. Η πράσινη περιοχή παριστάνει την προέκταση των pixel του χαρακτήρα τα οποία λαμβάνουν την μηδενική τιμή.



2.1 Απευθείας Σύγκριση

Η μεθοδολογία της απευθείας σύγκρισης (template matching) θεωρεί ότι τα pixels της εικόνας του χαρακτήρα είναι και το διάνυσμα των χαρακτηριστικών του. Δεδομένων δύο χαρακτήρων X , Y με n στοιχεία έκαστο, ορίζεται σαν n_{IJ} το πλήθος των pixel όπου ο χαρακτήρας X έχει τιμή i και ο χαρακτήρας Y έχει τιμή j , με $i, j \in \{0,1\}$.

Έτσι, έχουμε την απόσταση Jaccard η οποία ορίζεται ως:

- $$D_J = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

και την απόσταση Jule η οποία ορίζεται ως:

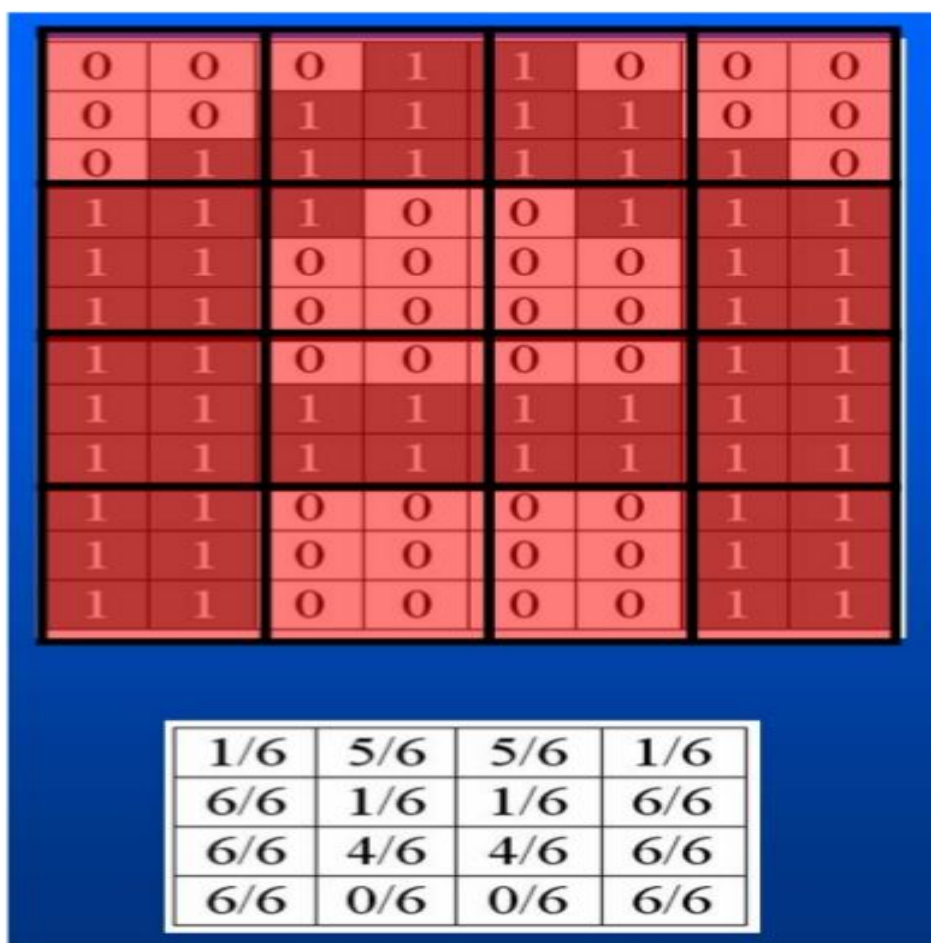
- $$D_Y = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}}$$

Οι παραπάνω αποστάσεις όσο πιο κοντά στο 1 είναι τόσο κοντύτερα είναι οι δύο χαρακτήρες. Συνεπώς, ο αλγόριθμος για κάθε δείγμα του testset υπολογίζει την απόσταση που ο χρήστης επιλέγει (D_J ή D_Y) από κάθε στοιχείο του trainset και

κρατάει την μεγαλύτερη έτσι ώστε να εξαγάγει συμπέρασμα για την κλαση στην οποία ανήκει το στοιχείο.

2.2 Ζώνες

Σύμφωνα με την προσέγγιση χωρισμού σε ζώνες, υπολογίζεται η κατανομή των μαύρων pixels σε διάφορες ζώνες στις οποίες χωρίζεται ο χαρακτήρας. Για την εξαγωγή χαρακτηριστικών γίνεται ο χωρισμός του χαρακτήρα σε ζώνες $n \times m$ pixel και ο υπολογισμός της πυκνότητας των pixels σε κάθε ζώνη. Ουσιαστικά ο χαρακτήρας αποκτά αριθμό χαρακτηριστικών όσες και οι ζώνες. Η παρακάτω εικόνα αποτυπώνει σχηματικά την μέθοδο με ένα παράδειγμα ζωνών 3×2 .



Στην επιφάνεια αλληλεπίδρασης της εφαρμογής ο χρήστης δύναται να επιλέξει τον αριθμό των pixels που θα έχει κάθε ζώνη από ένα σύνολο έτοιμων τιμών, αυτές

αφορούν ζώνες 2×2 , 5×5 , 10×10 και 25×25 . Αφου εξαχθεί η πυκνότητα των pixel κάθε ζώνης για ένα χαρακτήρα X_{test} υπολογίζουμε την ευκλείδεια απόσταση από όλους τους χαρακτήρες εκπαίδευσης και κρατάμε την κλάση εκείνου με την μικρότερη απόσταση (τον κοντινότερο γείτονα δηλαδή).

2.3 Προβολές

Η μέθοδος των προβολών αφορά τον υπολογισμό της πυκνότητας των pixel σε οριζόντιες και κάθετες προβολές στο σώμα του χαρακτήρα. Ο αριθμός των προβολών που επιθυμούμε καθορίζει τον αριθμό των χαρακτηριστικών του χαρακτήρα που θα χρησιμοποιηθούν αργότερα στην φάση της κατηγοριοποίησης.

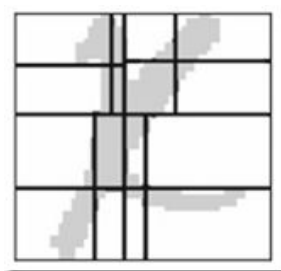
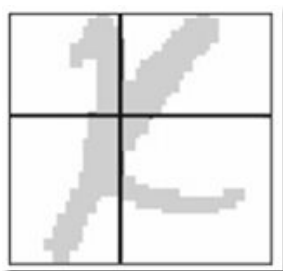
Στην επιφάνεια αλληλεπίδρασης της εφαρμογής ο χρήστης δύναται να επιλέξει τον αριθμό των προβολών (κάθετες και οριζόντιες έχουν την ίδια τιμή) από ένα σύνολο έτοιμων τιμών (2, 5, 10, 25, 50). Αφού εξαχθούν οι προβολές σαν χαρακτηριστικά ενός χαρακτήρα πραγματοποιείται η κατηγοριοποίηση που περιγράψαμε και στο κεφάλαιο 2.2.

2.4 Αναδρομικές Υποδιαιρέσεις¹

Η μέθοδος των αναδρομικών υποδιαιρέσεων αφορά των χωρισμό του χαρακτήρα σε περιοχές βάσει του κέντρου βάρους της καθεμιάς. Πιο αναλυτικά, κάθε χαρακτήρας χωρίζεται αρχικά σε 4 υποπεριοχές όπου η καθεμιά έχει όσο το δυνατόν ίσο αριθμό μαύρων pixel (foreground pixels) με τις υπόλοιπες. Ο χωρισμός γίνεται από μια κατακόρυφη και μια οριζόντια ευθεία, το σημείο που οι δύο ευθείες συναντιούνται ονομάζεται σημείο τομής (division point). Η ίδια διαδικασία πραγματοποιείται έπειτα αναδρομικά για καθεμια απο τις 4 υποπεριοχές που δημιουργήθηκαν.

Για επίπεδο διάσπασης L έχουμε 4^L σημεία τομής καθώς και 4^{L+1} υποπεριοχές. Σαν χαρακτηριστικά των χαρακτήρων χρησιμοποιούνται οι συντεταγμένες των σημείων τομής (άρα έχουμε 2×4^L χαρακτηριστικά για κάθε χαρακτήρα). Στην επιφάνεια αλληλεπίδρασης της εφαρμογής ο χρήστης δύναται να επιλέξει τον αριθμό του L από ένα σύνολο έτοιμων τιμών (0 έως 9). Αφού εξαχθούν οι συντεταγμένες των σημείων τομής σαν χαρακτηριστικά ενός χαρακτήρα πραγματοποιείται η κατηγοριοποίηση που περιγράψαμε και στο κεφάλαιο 2.2. Οι επόμενες τρεις εικόνες δείχνουν σχηματικά την διαδικασία που περιγράφηκε.

¹ "Handwritten character recognition through two-stage ... - Demokritos."
<http://users.iit.demokritos.gr/~bgat/PRHandRec2010.pdf>. Accessed 16 Sep. 2018.



3. Πειραματικά Αποτελέσματα

Τα παρακάτω αποτελέσματα έχουν ως παραμέτρους το σύνολο των μεθόδων - χαρακτηριστικών που μπορεί να επιλέξει ο χρήστης στην εφαρμογή.

Εξαγωγή Χαρακτηριστικών	Ακρίβεια (Accuracy)	Ενδεικτικός Χρόνος Εκτέλεσης (mm:ss)*
Προβολές (2)	41.9627%	00:00
Προβολές (5)	82.3142%	00:00
Προβολές (10)	88.9784%	00:01
Προβολές (25)	90.4797%	00:03
Προβολές (50)	90.8092%	00:06
Ζώνες (2x2)	99.5606%	00:30
Ζώνες (5x5)	99.3409%	00:05
Ζώνες (10x10)	95.13%	00:01
Ζώνες (25x25)	43.4273%	00:00
Απόσταση Jaccard	99.5972%	05:05
Απόσταση Yule	99.3775%	05:08
Αναδρομικές Υποδιαίρέσεις (Features=1, L=0)	40.5346%	00:00

Αναδρομικές Υποδιαιρέσεις (Features=4, L=1)	96.0088%	00:01
Αναδρομικές Υποδιαιρέσεις (Features=16, L=2)	98.9381%	00:03
Αναδρομικές Υποδιαιρέσεις (Features=64, L=3)	97.0707%	00:09
Αναδρομικές Υποδιαιρέσεις (Features=256, L=4)	89.491%	00:25
Αναδρομικές Υποδιαιρέσεις (Features=1024, L=5)	83.669%	01:27

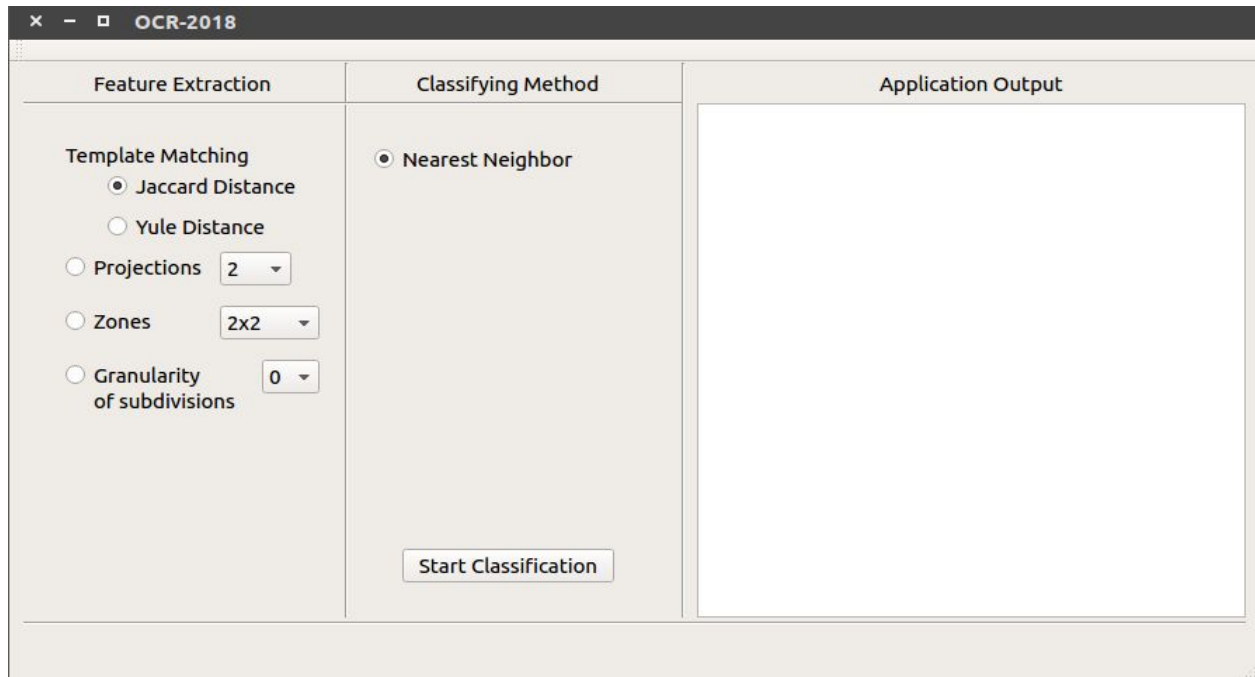
***Setup:** Intel Core i5-6200U CPU @ 2.30GHz × 4, 16GB RAM, Ubuntu 16.04 LTS, Qt 5.8.0

Παρατηρούμε πως την μεγαλύτερη ακρίβεια την προσφέρουν οι αποστάσεις με την απευθείας σύγκριση (Jaccard/Yule), η μέθοδος των ζωνών 2x2 καθώς και η μέθοδος των υποδιαιρέσεων με L=2. Την χειρότερη ακρίβεια αναμενόμενα έχουν οι ζώνες 25x25, οι προβολές με τιμή 2 καθώς και η μέθοδος των υποδιαιρέσεων με L=0.

Στην ουσία συμβαίνει αυτό που διαισθητικά περιμένουμε, όσο πιο πολύ “τεμαχίζουμε” τον χαρακτήρα τα χαρακτηριστικά του πλέον έχουν μεγαλύτερη πληροφορία για ένα σημείο και έτσι αποδίδει καλύτερα η κατηγοριοποίηση. Αντίθετα, όταν ο χαρακτήρας μας έχει λίγα χαρακτηριστικά χάνεται η πληροφορία που τον ξεχωρίζει στο σύνολο.

Το πρόγραμμα μετά από κάθε κατηγοριοποίηση ανοίγει σε νέο παράθυρο της εφαρμογής τον πίνακα συσχετισμών (confusion matrix). Λόγω μεγέθους δεν είναι δυνατό να παρουσιαστεί ο πίνακας σε αυτό το κεφάλαιο, ωστόσο ο αναγνώστης μπορεί να εκτελέσει την μέθοδο που επιθυμεί και να εξερευνήσει στην επιφάνεια της εφαρμογής ποιές κλάσεις μπερδεύονται και σε ποιές περιπτώσεις.

4. Εγχειρίδιο Εφαρμογής

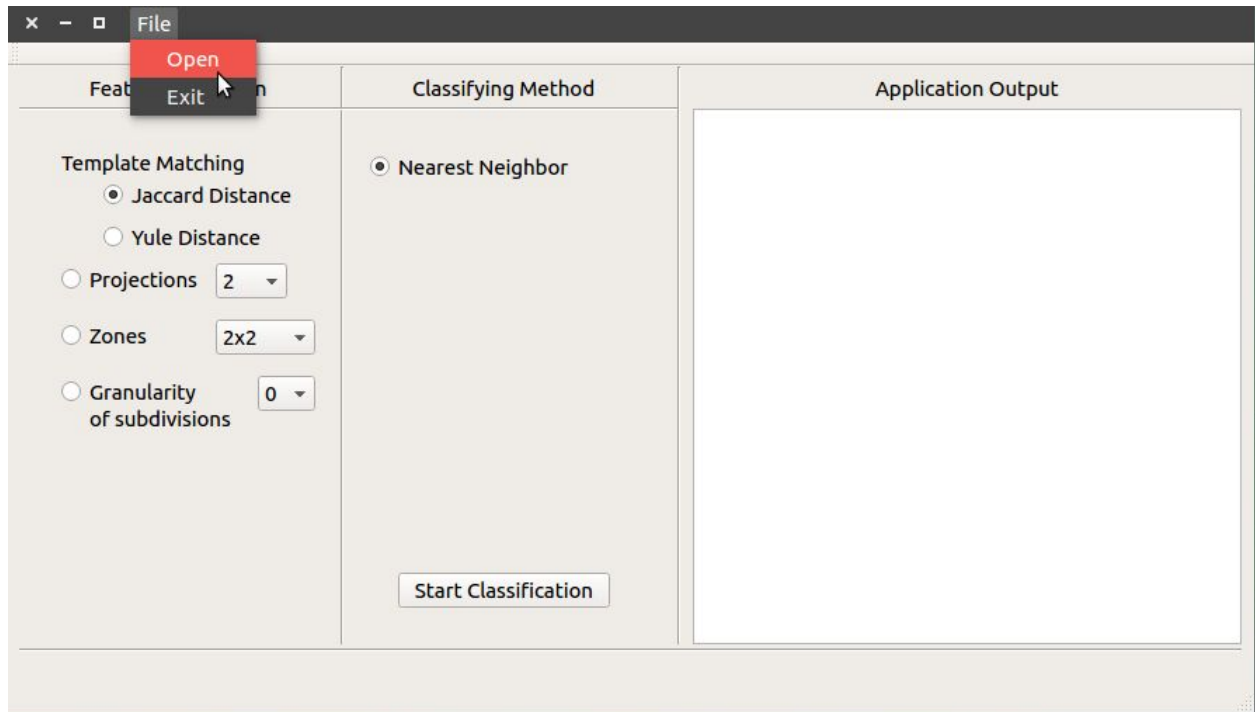


The screenshot displays the OCR-2018 application window. It is divided into three main sections: Feature Extraction, Classifying Method, and Application Output. The Feature Extraction section includes options for Template Matching (Jaccard Distance, Yule Distance), Projections (set to 2), Zones (set to 2x2), and Granularity of subdivisions (set to 0). The Classifying Method section has a radio button selected for Nearest Neighbor and a Start Classification button. The Application Output section is a large empty box for displaying results.

Feature Extraction	Classifying Method	Application Output
<p>Template Matching</p> <p><input checked="" type="radio"/> Jaccard Distance</p> <p><input type="radio"/> Yule Distance</p> <p><input type="radio"/> Projections 2</p> <p><input type="radio"/> Zones 2x2</p> <p><input type="radio"/> Granularity of subdivisions 0</p>	<p><input checked="" type="radio"/> Nearest Neighbor</p> <p><button>Start Classification</button></p>	

Η παραπάνω εικόνα αποτελεί την επιφάνεια εργασίας της εφαρμογής. Στην αριστερή στήλη ο χρήστης καλείται να επιλέξει την μέθοδο εξαγωγής χαρακτηριστικών, στην δεξιά στήλη η επιλογή του κατηγοριοποιητή είναι σταθερή. Επίσης, στο δεξί κομμάτι της επιφάνειας εργασίας ο χρήστης μπορεί να βλέπει μηνύματα ενημέρωσης από την κονσόλα.

Για να φορτώσει τα δεδομένα στην εφαρμογή ο χρήστης ακολουθεί τα βήματα που περιγράφονται για την επόμενη εικόνα.



Το πάτημα του *Open* ανοίγει στον χρήστη ένα παράθυρο για να φορτώσει έναν φάκελο. Η εφαρμογή περιμένει σαν είσοδο έναν φάκελο ο οποίος μέσα έχει τους φακέλους των διάφορων κλάσεων των χαρακτήρων (όπως ο φάκελος *chars* στο παραδοτέο). Κάθε υποφάκελος λοιπόν πρέπει να έχει το όνομα της κλάσης που οι εικόνες μέσα του αντιπροσωπεύουν, αυτό είναι σημαντικό γιατί από αυτά τα ονόματα δημιουργείται ο πίνακας συσχετισμών έπειτα. Αν οι εικόνες δεν φορτωθούν σωστά η κονσόλα θα πετάξει το αντίστοιχο μήνυμα και ο χρήστης μπορεί να ξαναπροσπαθήσει. Μια ενδεικτική εκτύπωση επιτυχούς φόρτωσης των δεδομένων είναι η εξής:

Loading images.. DONE

Trainset size: 2764

Testset size:2731

Normalizing.. DONE

Αν οι τιμές (size) ήταν 0 τότε οι εικόνες δεν θα είχαν φορτωθεί σωστά.

Αφου φορτώσει τα δεδομένα και επιλέξει την μέθοδο ο χρήστης πρέπει να πατήσει το κουμπί *Start Classification* και να περιμένει για το αποτέλεσμα. Το αποτέλεσμα φαίνεται στη εκτύπωση της κονσόλας (η ακρίβεια της μεθόδου) και ταυτόχρονα ανοίγει ένα νέο παράθυρο όπου παρουσιάζεται ο πίνακας συσχετισμών (confusion matrix). Μια αλληλουχία δύο επιτυχών εκτελέσεων καθώς και ένα υποσύνολο του πίνακα συσχετισμών φαίνονται στις επόμενες εικόνες.

Ο τρόπος που πρέπει να διαβαστεί ο πίνακας συσχετισμών είναι ως εξής: Στον άξονα του Χ είναι η πραγματική κλάση ενώ στον άξονα του Υ είναι η κλάση που προβλέφθηκε από το σύστημα. Στα κελιά που συναντιέται η ίδια κλάση (πχ στο πρώτο κελί για “ά”/”ά”) η τιμή μετά το / (πχ. το 66) αποτελεί το σύνολο των χαρακτήρων που δοκιμάστηκαν και ανήκουν σε αυτή τη κλάση. Έτσι βλέπουμε λόγου χάρη πως για την κλάση “ά”, απο τα 66 στοιχεία μόνο τα 27 προβλέφθηκαν σωστά ως “ά”, 19 προβλέφθηκαν λάθος ως “α Βαρεία”, 2 λάθος ως “β”, κ.ο.κ.