**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**
**ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ**

**Επεξεργασία Φωνής και Φυσικής Γλώσσας**
**Χειμερινό Εξάμηνο 2017-18**

# 1η Εργαστηριακή Άσκηση: Μετατροπέας για Greeklish

## ΠΕΡΙΓΡΑΦΗ

Σκοπός είναι η υλοποίηση ενός μετατροπέα για Greeklish. Για αυτό το εργαστήριο θα πρέπει να σχεδιάσετε ένα σύστημα το οποίο δέχεται σαν είσοδο Greeklish (Ελληνικές λέξεις γραμμένες με λατινικούς χαρακτήρες, μία γραφή ιδιαίτερα συνηθισμένη σε e-mail) και Αγγλικά και παράγει Ελληνικά, αφήνοντας χωρίς αλλαγή τις Αγγλικές λέξεις που υπήρχαν στην είσοδο.

Θα σας δωθούν τα κατάλληλα λεξικά για τα Ελληνικά και τα Αγγλικά, καθώς και τα απαραίτητα δεδομένα για την εκπαίδευση και την αποτίμηση του μοντέλου σας. Θα πρέπει να ορίσετε το σύνολο των πιο συνηθισμένων κανόνων για τη μεταγραφή Ελληνικών χαρακτήρων σε λατινικούς χαρακτήρες. Αυτοί οι κανόνες θα χρησιμοποιηθούν για τη μεταγραφή από Greeklish σε Ελληνικά λέξεων που δεν υπάρχουν στο ελληνικό λεξικό.

Οι επιδόσεις του συστήματος θα πρέπει να μετρηθούν χρησιμοποιώντας τα δεδομένα αποτίμησης. Τέλος, θα βελτιώσετε το μοντέλο που δημιουργήσατε ενσωματώνοντας έναν ορθογράφο που θα διορθώνει τις λέξεις που μετατράπηκαν λάθος στα Ελληνικά.

Περιγραφή των δεδομένων:
Για τους σκοπούς του εργαστηρίου θα σας δοθούν τα ακόλουθα δεδομένα:
1) Ελληνικό λεξικό (el_caps_noaccent.dict) το οποίο περιέχει 393949 λέξεις
2) Αγγλικό λεξικό (el_caps_noaccent.dict) το οποίο περιέχει 116344 λέξεις
3) Ελληνικό corpus που θα χρησιμοποιήσετε ως test δεδομένα (test_gr.txt)
4) Greeklish corpus που θα χρησιμοποιήσετε ως test δεδομένα (test_greng.txt)
5) Ελληνικό corpus που θα χρησιμοποιήσετε ως train δεδομένα (train_gr.txt)
6) Greeklish corpus που θα χρησιμοποιήσετε ως train δεδομένα (train_greng.txt)

Μεταξύ των 3,4 και 5,6 υπάρχει 1-1 αντιστοιχία.

## ΠΡΟΕΤΟΙΜΑΣΙΑ

Κατεβάστε τα δεδομένα από το παρακάτω link. Αν ζητηθεί κωδικός για τα δεδομένα εισάγετε τον κωδικό που δόθηκε στο μάθημα:
http://cvsp.cs.ntua.gr/courses/speech_lang_proc/material.shtm

Για την εκτέλεση των βημάτων της άσκησης θα χρειαστεί να εγκαταστήσετε τη βιβλιοθήκη OpenFST (http://www.openfst.org/twiki/bin/view/FST/FstDownload) η οποία επιτρέπει την εύκολη δημιουργία μηχανών πεπερασμένης κατάστασης (FSMs) με χρήση εντολών shell. Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε στο documentation της βιβλιοθήκης (http://www.openfst.org/twiki/bin/view/FST/FstQuickTour).

Σε περίπτωση που θέλετε να υλοποιήσετε την άσκηση σε Python συνίσταται η εγκατάσταση της έκδοσης 1.6.1 της βιβλιοθήκης, καθώς και η εγκατάσταση των αντίστοιχων wrappers (https://pypi.python.org/pypi/openfst/1.6.1)

*Σημείωση: Τα βήματα 1-7 αποτελούν μέρος της προπαρασκευής η οποία έχει προηγηθεί και επαναλαμβάνονται για σκοπούς πληρότητας της άσκησης.*

Βήμα 1
Γράψτε κώδικα σε γλώσσα της προτίμησής σας (προτείνεται να χρησιμοποιήσετε script language, e.g. Perl/Python) προκειμένου να συγκρίνετε τα αρχεία κειμένου "train_gr.txt" και "train_greng.txt" γράμμα προς γράμμα και δημιουργήστε μια αντιστοιχία μεταξύ των των λατινικών και των αντίστοιχων Ελληνικών χαρακτήρων. Σημειώνεται ότι η αντιστοίχιση δεν είναι πάντα 1-1 (Βρείτε πρώτα τις 1-1 αντιστοιχίες και στη συνέχεια τις 2-1 ή 1-2 κ.ο.κ). Για ευκολία έχουν αφαιρεθεί τα σημεία στίξης, οι ειδικοί χαρακτήρες και χρησιμοποιούνται μόνο κεφαλαία γράμματα.

Βήμα 2
Υπολογίστε την πιθανότητα του κάθε κανόντα μετατροπής από λατινικό σε Ελληνικό χαρακτήρα (ο αριθμός των φορών που εμφανίζεται ο καθε κανόνας στο κείμενο, κανονικοποιημένος ως προς το συνολικό αριθμό των κανόνων).

Βήμα 3
Δημιουργήστε ένα μετατροπέα (fst) G που αντιστοιχίζει τους λατινικούς με τους Ελληνικών χαρακτήρων (greeklish μετατροπεας), χρησιμοποιώντας ως κόστη τον αρνητικό λογάριθμο των πιθανοτήτων που υπολογίστηκαν παραπάνω.

Βήμα 4
Δημιουργήστε ένα μετατροπέα (fst) I που αντιστοιχίζει κάθε λατινικό χαρακτήρα στον εαυτό του με πολύ μεγάλο (σταθερό) κόστος. Γιατί χρειάζεται αυτό το βήμα;

Βήμα 5
Δημιουργήστε έναν αποδοχέα (fsa) A1 για το Ελληνικό λεξικό, ο οποίος αποδέχεται όλες τις ακολουθίες Ελληνικών γραμμάτων που σχηματίζουν μια έγκυρη λέξη.

Βήμα 6
Μετατρέψτε τον αποδοχέα A1 στον αντίστοιχο ντετερμινιστικό και ελαχιστοποιήστε τον.

Βήμα 7
Επαναλάβατε για τον αποδοχέα (fsa) A2 για το Αγγλικό λεξικό.

**ΕΚΤΕΛΕΣΗ**

*Τα παρακάτω βήματα δεν αποτελούν μέρος της προπαρασκευής.*

Βήμα 8
Αφού βρείτε την ένωση των A1 και A2 συνθέστε την με την κλειστότητα του G. Στο σημείο αυτό έχετε κατασκευάσει έναν greeklish μετατροπέα T που μετατρέπει τα greeklish σε Ελληνικές ή Αγγλικές λέξεις.

Βήμα 9

Δημιουργήστε έναν αποδοχέα W για κάθε λέξη του αρχείου "Test set in Greeklish" (ακολουθία λατινικών χαρακτήρων που αντιστοιχεί στη λέξη).

Βήμα 10
Συνθέστε τις μηχανές W και T και βρείτε το καλύτερο μονοπάτι. Σε αυτό το σημείο έχετε βρει την πιο πιθανή Ελληνική λέξη για την αντίστοιχη greeklish.

Βήμα 11
Υπολογίστε την ακρίβεια (accuracy) του συστήματός σας, κάνοντας το ίδιο για κάθε λέξη του "Test set in Greeklish". Υπολογίστε τα παρακάτω στατιστικά συγκρίνοντας τα αποτελέσματα με το αρχειο "Test set in Greek": 1) Ποσοστό των σωστών λέξεων, 2) Ποσοστό των Αγγλικων λέξεων που μετατράπηκαν σε Ελληνικές, 3) Ποσοστό των λέξεων που δεν δεν αντιστοιχήθηκαν σε καμία λέξη του Ελληνικού ή Αγγλικού λεξικού. 4) Ποσοστό των λέξεων που μετατράπηκαν λαθος.

Βήμα 12
Βελτιώστε το μοντέλο σας ενσωματώνοντας έναν ορθογράφο για τις Ελληνικές λέξεις.
Τα λαθη που μπορούν να υπάρξουν μπορεί να είναι τεσσάρων ειδών:
1) εισαγωγές χαρακτήρων
2) διαγραφές χαρακτήρων
3) αντικαταστάσεις χαρακτήρων
4) αναδιατάξεις διαδοχικών χαρακτήρων

Αρχικά υποθέστε ότι μπορεί να υπάρχει μόνο ένα λάθος ανα λέξη. Στη συνέχεια γενικεύστε το μοντέλο σας ώστε να καλύπτει και περισσότερα λάθη ανά λέξη

Για την υλοποίση του ορθογράφου ακολουθήστε τα παρακάτω βήματα:

i) Θα σας δοθούν δεδομένα στα ελληνικά (train_wr.txt, train_co.txt)

ii)Γράψτε κώδικα σε γλώσσα της προτίμησής σας (προτείνεται να χρησιμοποιήσετε script language, e.g. Perl/Python) που συγκρίνει τα δύο αρχεία train_wr.txt, train_co.txt λέξη προς λέξη και υπολογίζει το συνολικό αριθμό των λαθών εισαγωγής, διαγραφής, αντικαταστάσης και αναδιατάξης χαρακτήρων.

iii)Υπολογίστε την πιθανότητα εμφάνισης του κάθε λάθους. Υποθέστε πως η πιθανότητες είναι ανεξάρτητες απο τα συμφραζόμενα του κάθε γράμματος/λέξης.

iv) Δημιουργήστε ένα μετατροπέα I ο οποίος αντιστοιχίζει κάθε γράμμα στον εαυτό του με κόστος 0

v) Δημιουργήστε ένα μετατροπέα E ο οποίος αντιστοιχίζει κάθε γράμμα στον εαυτό του με κόστος 0 και επιτρέπει μόνο μια εισαγωγή, διαγραφή, αντικατάσταση ή αναδιάταξη με κόστος τον αρνητικό λογάριθμο των πιθανοτήτων που έχουν υπολογιστεί.

vi) Δημιουργήστε ένα μετατροπέα S1 ο οποίος επιτρέπει μονο μια εισαγωγή, διαγραφή, αντικατάσταση ή αναδιάταξη (με τα κατάλληλα βάρη), δηλαδή το concatenation των I* με το E και I*

vii) Δημιουργήστε ένα μετατροπέα S2 ο οποίος επιτρέπει ακριβώς δύο εισαγωγές, διαγραφές, αντικαταστάσεις ή αναδιατάξεις για κάθε λέξη (με τα κατάλληλα βάρη)

viii) Δημιουργήστε ένα μετατροπέα L ο οποίος αντιστοιχίζει μια ακολουθία γραμμάτων στην αντίστοιχη λέξη του λεξικού (Το βήμα 5 της προπαρασκευής)

x) Γράψτε κώδικα που συγκρίνει τα test αρχεια "test_co.txt" και "test_wr.txt" και για κάθε λέξη που είναι διαφορετική βρίσκει την κοντινότερη λέξη στο λεξικό σύμφωνα με το μοντέλο στο S1 και S2 αντίστοιχα, δηλαδή το καλύτερο μονοπάτι της σύνθεσης των μηχανών (W,S1,L), όπου W είναι η ακολουθία των γραμμάτων σε κάθε λέξη (αναπαριστάται ως ένας αποδοχέας). Ποιο είναι το ποσοστό των (best) λέξεων που προτείνονται από τον ορθογράφο και είναι σωστές (για το S1 και S2)?

x) Χρησιμοποιείστε την έξοδο του μετατροπέα για Greeklish ως είσοδο στον ορθογράφο για να διορθώσετε ορθογραφικά λάθη. Βελτιώθηκε η απόδοση του μετατροπέα για Greeklish?

*Προχωρείστε σε κατ' οίκον ολοκλήρωση των βημάτων εκείνων που δεν προλάβατε κατά τη διεξαγωγή του εργαστηρίου.*

**ΠΑΡΑΔΟΤΕΑ**

Αφορά τα *εκτός προπαρασκευής* βήματα μόνο.

(1) Σύντομη αναφορά (σε pdf) που θα περιγράφει τη διαδικασία που ακολουθήθηκε σε κάθε βήμα, καθώς και τα σχετικά αποτελέσματα.
(2) Κώδικας, ο οποίος περιέχει και τις εντολές του OpenFst (συνοδευόμενος από σύντομα σχόλια).

Συγκεντρώστε τα (1) και (2) σε ένα .zip αρχείο το οποίο πρέπει να αποσταλεί μέσω του mycourses.ntua.gr εντός της καθορισμένης προθεσμίας.