

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Επεξεργασία Φωνής και Φυσικής Γλώσσας
Χειμερινό Εξάμηνο 2017-18

1η Εργαστηριακή Άσκηση: Μετατροπές 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 δοθούν 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.shtml

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 αντίστοιχών wrappers (<https://pypi.python.org/pypi/openfst/1.6.1>)

ΕΚΤΕΛΕΣΗ

Υλοποιήστε τα παρακάτω βήματα χρησιμοποιώντας OpenFst

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 κανόνα μετατροπής από λατινικό σε Ελληνικό χαρακτήρα (ο αριθμός των φορών που εμφανίζεται ο κάθε 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 μετατροπείας), 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

Επαναλάβετε για τον αποδοχέα (fsa) A2 για το Αγγλικό λεξικό.

ΠΑΡΑΔΟΤΕΑ

- (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 πριν την προπαρασκευή του εργαστηρίου.