Machine Learning Assignment using NLTK

Σε αυτό το αρχείο παρουσιάζεται μία μικρή τεκμηρίωση του κώδικα που γράφηκε για την λύση της 5ης εργασίας της μηχανικής μάθησης.

ΕΡΩΤΗΜΑ [Α]

Αρχικά, αφού φορτώσουμε όλες τις βιβλιοθήκες που χρειαστήκαμε, δημιουργούμε μία λίστα από τα δέκα βιβλία που χρησιμοποιήσαμε, για να δημιουργήσουμε ένα λεξικό το οποίο περιλαμβάνει όλες τις λέξεις που εμφανίζονται σε αυτά τα βιβλία.

Στην συνέχεια, αφού βρούμε όλες τις λέξεις και τις προτάσεις που υπάρχουν σε κάθε βιβλίο χρησιμοποιώντας την εντολή gutenberg.words και gutenberg.sents αντίστοιχα, δημιουργούμε τις συναρτήσεις οι οποίες θα μας βοηθήσουν να παράγουμε 10 προτάσεις από κάθε βιβλίο. Οι πρώτες δύο συναρτήσεις παίρνουν σαν όρισμα τις προτάσεις που βρήκαμε πιο πριν από ένα βιβλίο και δημιουργούν ένα λεξικό το οποίο βρίσκει τις πιθανότητες των διγραμμάτων και τριγραμμάτων. Οι συναρτήσεις αυτές βασίστηκαν σε μία εργασία που είναι ανεβασμένη στο διαδίκτυο η οποία βρίσκετα εδώ.

Αφού δημιούργησαμε αυτά τα δύολεξικά τα οποία επιστρέφουν τα λεξικά που θα χρειαστούμε, δημιουργούμε δύο νέες συναρτήσεις . Η συνάρτηση που εκτυπώνει τις προτάσεις από κάθε βιβλίο χρησιμοποιώντας τα διγράμματα παίρνει σαν όρισμα το λεξικό που περιλαμβάνει τις πιθανότητες των διγραμμάτων, ενώ η συνάρτηση που εκτυπώνει τις προτάσεις χρησιμοποιώντας τριγράμματα παίρνει το αντίστοιχο λεξικό.

Για τα διγράμματα, ορίσαμε σαν token αρχής πρότασης την λέξη ‘the’ και σαν τέλος πρότασης την τελεία ‘.’ . Η συνάρτηση βρίσκει το κλειδι του λεξικού ‘the’ και επιλέγει τυχαία, με βάση τις πιθανότητες που έχουν σχηματιστεί λόγω διγραμμάτων, την λέξη που θα ακολουθέι την λέξη ‘the’. Με βάση αυτή την επιλογή, βρίσκει το κλειδί του λεξικού της τελευταίας λέξης που παρήχθησε και επιλέγει ξανά μία νέα λέξη. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να παραχθεί η τελεία, που είναι το token για το τέλος της πρότασης, και τέλος δημιουργέι την πρόταση με όλες τις λέξεις που επιλέχθηκαν μέχρι να βρεθεί η τελεία.

Η συνάρτηση που εκτυπώνει προτάσεις με την βοήθεια των τριγραμμάτων ακολουθεί το ίδιο σκεπτικό, μόνο που τώρα έχουμε ορίσει δύο token αρχής πρότασης, τα ‘and’ και ‘the’. Σαν token τέλους χρησιμοποιήθηκε πάλι η τελεία.

Τέλος, κάνοντας 10 επαναλήψεις για κάθε βιβλίο που διαλέξαμε, εκτυπώνουμε 10 προτάσεις που παρήχθησαν από τα διγράμματα και 10 από τα τριγράμματα. Συνολικά δηλαδή εκτυπώθηκαν 10(αριθμός πρωτάσεων) * 10 (αριθμός βιβλίων που επιλέξαμε) = 100 προτάσεις από τα διγράμματα και 100 από τα τριγράμματα.

Παρακάτω παραθέτουμε μία πρόταση από κάθε βιβλιο με την βοήθεια διγραμμάτων και μία από κάθε βιβλίο χρησιμοποιώντας τριγράμματα.

(α) Προτάσεις χρησιμοποιώντας διγράμματα.

1. the promise me when , with him at the plain was here ; and Anne could not at Winthrop , and terror .
2. the night was an angel is forged iron , Peace , tigers howl for mine .
3. the Civil War - dog to send a mistake about the king named Zelia ' s ring all .
4. the little pool , all about berry pies .
5. the sand with this minute , for going on .
6. the other touches .
7. the air , the moon , reasonless .
8. the World .
9. the witnesse of her Father Hamlet .
10. the while , Children , then the Night .

(β) Προτάσεις χρησιμοποιώντας διγράμματα.

1. and the more lively character were irresistible .
2. and the springs To flourish in eternal vales : they why should Thel complain .
3. and the Elephant ' s pause , and who carried short , and of so fine a quality , that had once sparkled in a bare little room .
4. and the Smiling Pool , and Bobby Coon .
5. and the baby violently up and picking the daisies , when they passed too close , and went in without knocking , and then all the time .
6. and the man he was abnormally tall and quite unmistakable , a large and fat old gentleman with the elephant .
7. and the Pleiades , before the threats Of Gabriel out of darkness answered glad .
8. and the state of things .
9. and the swaggering vpspring reeles , And Iemme of all the strength and Armour of the whole State .
10. and the loyaltie I owe , In the great Hand of God I stand , and mouncht : Giue me my most worthy Friends : my way it lyes .

ΕΡΩΤΗΜΑ [Β]

Σε αυτό το μέρος της εργασίας έπρεπε να ταξινομήσουμε σωστά τις θετικές και τις αρνητικές αξιολογήσεις ταινιών.

Αφού βρούμε την κατανομή των λέξεων που έχουν οι κριτικές, δηλαδή τις λέξεις που υπάρχουν και την συχνότητα που εμφανίζονται, δημιουργούμε μία λίστα από πλειάδες, στην οποία κάθε πλειάδα έχει ως πρώτο στοιχείο τις λέξεις κάθε κριτικής και σαν δεύτερο στοιχείο την κατηγορία της κριτικής, δηλαδή αν είναι θετική ή αρνητική.

Στην συνέχεια δημιουργούμε μία συνάρτηση η οποία θα μας βοηθήσει να δημιουργήσουμε το σετ εκπαίδευσης μας. Πιο συγκεκριμένα, στο τέλος θα έχουμε για κάθε κριτική, το ποιές λέξεις από τις 4000 που εμφανίζονται πιο συχνά, υπάρχουν σε κάθε μία και ποιές όχι. Έτσι θα μπορέσουμε να χωρίσουμε train και test sets.

Αυτό που ακολουθεί είναι μία κλασσική διαδικασία εκπαίδευσης και αξιολόγησης ενός Naïve Bayes clsssifier, ο οποίος από ότι φαίνεται ταξινομεί τις κριτικές του test set με ακρίβεια 0.818.

Οι πηγές που χρησιμοποιήθηκαν από το διαδίκτυο είναι:

1. Πηγή 1
2. Πηγή 2