ΕΡΓΑΣΙΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2019-2020

ΠΑΝΑΓΙΩΤΗΣ ΑΠΟΣΤΟΛΟΠΟΥΛΟΣ(Π17007)

# Θέμα 1ο

# 1. Εισαγωγή

Ένα από τα σύγχρονα θέματα της Επεξεργασίας Φυσικής Γλώσσας (Natural Processing Language) είναι αυτό της κατηγοριοποίησης των διάφορων κειμένων ως θετικά ή αρνητικά. Η ταξινόμηση των κειμένων σε αυτές τις δύο κατηγορίες είναι πολύ σημαντική καθώς μπορούν μεγάλες πλατφόρμες ή μέσα κοινωνικής δικτύωσης να τα χρησιμοποιήσουν με κατάλληλο τρόπο, προκειμένου να παραχθούν χρήσιμα συμπεράσματα. Επιπρόσθετα, να σημειώσουμε ότι η χρήση αντίστοιχων εφαρμογών γίνεται από πλατφόρμες όπως το Facebook, το Twitter, αλλά και από το Imdb για την αποσαφήνιση των κριτικών από τους χρήστες για τις διάφορες ταινίες/σειρές. Για την αντιμετώπιση του συγκεκριμένου θέματος μπορούμε να φτιάξουμε ένα μοντέλο πρόβλεψης. Πιο συγκεκριμένα, ένας ενδεδειγμένος τρόπος προσέγγισης είναι αυτός των **νευρωνικών δικτύων** και **word embeddings**.

# 2. Παρουσίαση κεντρικής ιδέας

Η υλοποίηση του συγκεκριμένου μοντέλου θα μπορούσε να γίνει με την χρήση της γλώσσας προγραμματισμού **Python**, λόγω των έτοιμων βιβλιοθηκών που μας προσφέρει. Ωστόσο, στη συγκεκριμένη παρουσίαση θα δοθεί βάρος στην ανάλυση του θεωρητικού υποβάθρου και όχι του κώδικα, με την παρουσίαση αρκετών χρήσιμων μοντέλων που συμβάλλουν στην επίτευξη του στόχου.

2.1 Κωδικοποίηση κειμένου

Αρχικά, για την επίτευξη του στόχου μας θα πρέπει να έχουμε ένα σύνολο από κείμενα(κριτικές), τα οποία να έχουν κάποια ετικέτα, που να τα χαρακτηρίζει ως θετικά ή αρνητικά. Σε αυτό το σημείο δεν μας ενδιαφέρει το σημασιολογικό περιεχόμενο του κειμένου, δηλαδή δεν θα πραγματοποιήσουμε κάποια σημασιολογική ανάλυση. Επιπρόσθετα, δεν θα ληφθούν υπόψη τα οποία σημασιολογικά σχήματα, όπως οι μεταφορές ή οι παρομοιώσεις.

Μια προσέγγιση για τον «χαρακτηρισμό» ενός κειμένου σε θετικό ή αρνητικό θα μπορούσε να ήταν η δημιουργία ενός λεξικού με όλες τις πιθανές λέξεις, ώστε κάθε φορά να μετράμε το βάρος των λέξεων του κειμένου. Για παράδειγμα σε ένα κείμενο 200 λέξεων θα βρίσκαμε τις θετικές και αρνητικές λέξεις του κειμένου και ανάλογα το πρόσημο, δηλαδή την μεταξύ τους διαφορά, θα εξαγάγαμε το συμπέρασμά μας . Ωστόσο, αυτή η τεχνική εμπεριέχει τον κίνδυνο του καθορισμού του βάρους μιας λέξης με υποκειμενικά κριτήρια και όχι με αντικειμενικά ή ακόμα και την ύπαρξη σαρκασμών και ειρωνειών (σημασιολογική αμφισημία).

Έτσι κάθε κείμενο θα το διαχειριζόμαστε ως μια διαφορετική οντότητα, δίχως να ενδιαφερόμαστε εάν ένα κείμενο είναι θετικό ή αρνητικό, και να φτιάξουμε ένα σύστημα το οποίο να μας υπολογίζει την πιθανότητα να ανήκει σε μία από τις δύο κατηγορίες με βάση τις προηγούμενες προσημειώσεις.

2.2 Αναπαράσταση

Για να μπορέσουμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης θα πρέπει να αναπαραστήσουμε τις λέξεις σε μορφή διανύσματος, ώστε να αυτά να δίδονται σαν είσοδο σε κάποιον αλγόριθμο μηχανικής μάθησης, όπως για παράδειγμα σε ένα νευρωνικό δίκτυο. Πιο συγκεκριμένα, αφού σκανάρουμε όλα τα διαθέσιμα κείμενα, βρίσκουμε τις μοναδικές εμφανίσεις της κάθε λέξης και τις αποθηκεύουμε σε μια λίστα-λεξικό (dictionary), με διαστάσεις **v**, όσες και οι λέξεις. Για κάθε λέξη σε αυτό το λεξικό δημιουργούμε ένα διάνυσμα **v** θέσεων, στο οποίο στη θέση στην οποία αντιστοιχεί η θέση της λέξης, βάζουμε την τιμή **1** και σε όλες τις υπόλοιπες. Για παράδειγμα εάν ο λεξικό μας έχει τις λέξεις [“the”, “big”, “dog”], το διάνυσμα που αντιστοιχεί στη λέξη “the” είναι το [1,0,0], στη λέξη “big” είναι το [0,1,0] και στη λέξη “dog” είναι το [0,0,1]. Συνήθως, κατά την διαδικασία «χτισίματος του λεξικού» για να κρατήσουμε τον αριθμό των λέξεων μικρό μετατρέπουμε τα κείμενα σε lower case και παραλείπουμε τα στοιχεία στίξης, αλλά και αφαιρούμε τις λέξεις με λιγότερες από **k** εμφανίσεις, όπου **k** είναι μια παράμετρος που την θέτουμε εμείς.

2.3 Ομοιότητα διανυσμάτων

Σημαντικό βήμα είναι η αξιοποίηση των διανυσμάτων που δημιουργήσαμε, δηλαδή να είμαστε σε θέση να συγκρίνουμε πόσο δύο διανύσματα στον χώρο είναι «κοντά», δηλαδή σημασιολογικά. Επειδή δύναται να έχουμε διανύσματα με τελείως διαφορετικά μήκη δεν θα χρησιμοποιήσουμε το εσωτερικό γινόμενο, αλλά η γωνία συνημίτονου (Εικόνα 1).

2.4 Στατιστικό γλωσσικό μοντέλο

Με τον όρο αυτό αναφερόμαστε σε μια κατανομή, είναι δηλαδή μια ακολουθία κατανομής λέξεων. Αυτό που μας ενδιαφέρει, είναι η εύρεση της κατανομής της λέξης η οποία «ακολουθεί» μια συγκεκριμένη ακολουθία λέξεων. Η συγκεκριμένη μοντελοποίηση είναι ευρέως γνωστή σε εφαρμογές όπως το Word (spell checking), η αναγνώριση φωνής ακόμα και το autocomplete. Aαναλυτικότερα για την πρώτη εφαρμογή, για μια λέξη που δεν είναι στο λεξικό, αναζητούνται κάποια μέτρα απόστασης, ώστε να βρεθούν οι λέξεις που ταιριάζουν, ταξινομημένες πάντα σύμφωνα με τις πιθανότητές τους. Ωστόσο, στη φυσική γλώσσα υπάρχει η πιθανότητα να δημιουργηθεί μια καινούργια πρόταση, η οποία να μην υπάρχει σε καμία συλλογή οπότε η αντιμετώπιση είναι πιο περίπλοκη.

2.5 Το μοντέλο word2vec

Για την αντιμετώπιση του παραπάνω προβλήματος υιοθετήθηκε η πρόταση του Mikolov, γνωστό και ως το μοντέλο **word2vec**.