Εργασία 1
Documentation

Ερώτημα 3:

Για το 3ο ερώτημα του αλγορίθμου επιλέξαμε τον Classifier KNearestNeighbor που υλοποιήσαμε στο προηγούμενο ερώτημα.

Προκειμένου να βελτιώσουμε την αποτελεσματικότητα του Classifier αλλά και τον χρόνο εκτέλεσης του προγράμματος δοκιμάσαμε τις εξής βελτιωτικές ενέργειες.

Προεπεξεργασία Δεδομένων:

1. Προσθήκη περισσότερων stop_words. (καθυστερεί λίγο η προεπεξεργασία πιο γρήγορα αποτελεσματα στα υπόλοιπα.)
2. Δοκιμή Stemming

Υλοποίηση KNN:

1. Αύξηση γειτόνων (k) (ως ενα οριο βελτιωνει την ακριβεια)
2. Μείωση γειτόνων (k) (μειωνει την ακριβεια αλλα ειναι γρηγορο)
3. Πρόβλεψη κατηγορίας με χρήση βάρους στην πρόσθεση ανάλογα με την απόσταση
4. Αξιοποίηση τίτλου κατά την κατηγοριοποίηση.

Προεπεξεργασία Δεδομένων:

1)Προσθήκη περισσότερων stop_words:

Υλοποιήσαμε τον αλγόριθμο addmore ο οποίος προσθέτει τις "περιττές” λέξεις του πεδίου Content στις stopwords. Ο τρόπος λειτουργίας του είναι πολύ απλός. Ελέγχει κάθε λέξη του πεδίου Content με κάθε λέξη του stopword. Αν η μία λέξη περιέχεται μέσα στην άλλη (για παράδειγμα **The**, **The**re). Αν ισχύει, τότε προσθέτει και την δεύτερη στα stopwords.

Αποτέλεσμα: Η προεπεξεργασία των δεδομένων γίνεται αρκετά πιο αργή, αφού πρέπει για όλες τις λέξεις του αρχείου στο πεδίο Content να ελέγξει όλες τις stopwords. Ωστόσο η εκτέλεση του υπόλοιπου αλγορίθμου, δηλαδη η εκτέλεση των classifiers είναι πιο γρήγορη.

2) Δοκιμή Stemming:

Κατά την προεπεξεργασία των δεδομένων, εγαρμόζουμε την τεχνική stemming με σκοπό να τον περιορισμό των διαφορετικών λέξεων. Ουσιαστικά έτσι δίνεται περισσότερη σημασία στην λέξη παραλείποντας κυρίως τις καταλήξεις. Επίσης όλες οι λέξεις ξεκινούν πλέον από μικρό γράμμα οπότε κατά την σύγκρισή τους με τις stopwords (όπως αναλύσαμε από πάνω), εντοπίζονται ακόμα περισσότερες περιττές λέξεις.

Αποτέλεσμα:

Ενδεικτική εκτέλεση χωρίς stemming:

	Statistic Measure	Naive Bayes	Random Forest	SVM	KNN	My Method
0	Accuracy	89.5339761248852	89.5339761248852	89.5339761248852	93.3333333333333	93.3333333333333
1	Precision	87.7575757575757	87.7575757575757	87.7575757575757		
2	Recall	86.2662337662338	86.2662337662338	86.2662337662338		
3	F-Measure	87.6645218417946	87.6645218417946	87.6645218417946		

Ενδεικτική εκτέλεση με stemming:

	Statistic Meas	Naive Bayes	Random Fore	SVM	KNN	My Method
0	Accuracy	89.5414371	89.5414371	89.5414371	96.6666667	96.6666667
1	Precision	88.1515152	88.1515152	88.1515152		
2	Recall	86.7157872	86.7157872	86.7157872		
3	F-Measure	88.1363735	88.1363735	88.1363735		

Η ακρίβεια των classifiers αυξάνεται, το ίδιο όμως παθαίνει και ο χρόνος εκτέλεσης.
Δίνεται περισσότερη βαρύτητα στο σημαντικό τμήμα των λέξεων και όχι στις καταλήξεις. Μπορεί να εντοπιστεί ομοιότητα σε περισσότερες λέξεις και έτσι να προσεγγύσουμε με τους γείτονες την κατηγορία, έχοντας μικρότερο σφάλμα.
Η ταχύτητα της προεπεξεργασίας μειώνεται, αφού και σε αυτή τη περίπτωση πρέπει να “περάσουμε” από το stemming μία μία τις λέξεις του πεδίου Content.

Υλοποίηση KNN:

1) Αύξηση γειτόνων (k) ή Μείωση γειτόνων (k)

Κάναμε δοκιμές αυξάνοντας τον αριθμό των γειτόνων.

Εκτέλεση με k = 2 : accuracy = 83.3333333333%

Εκτέλεση με k = 5 : accuracy = 90%

Εκτέλεση με k = 10 : accuracy = 93.333333333%

Εκτέλεση με k = 20 : accuracy = 83.3333333333%

Καλύτερη λογική είναι να έχουμε ένα μέσο αριθμό γειτόνων.
Όταν έχουμε μικρό είτε μεγάλο αριθμό γειτόνων μπορεί να γίνει λάθος πρόβλεψη. Στην πρώτη περίπτωση από έλειψη πληροφοριών, ενώ στην δεύτερη περίπτωση από εκμετάλλευση περιττών δεδομένων. Επίσης όταν έχουμε πολλούς γείτονες χρησιμοποιούμε αρκετά περισσότερη μνήμη και ίσως αντιμετωπίσουμε πρόβλημα στη διαχείρησή της. Τέλος με τον έλεγχο περισσότερων γειτόνων σπαταλάμε περισσότερο χρόνο στην εύρεση τους αλλά και στις απαραίτητες μαθηματικές πράξεις.

2) Πρόβλεψη κατηγορίας με χρήση βάρους στην πρόσθεση ανάλογα με την απόσταση:

Στον αλγόριθμο KNN έχουμε την συνάρτηση Predict στην οποία, για κάθε Content του test_data βρίσκουμε τους k κοντινότερος γείτονες. Για κάθε κατηγορία υπολογίζουμε τους ψήφους λαμβάνοντας υπόψην την απόσταση του αντικειμένου που διαθέτει αυτή την κατηγορία.
Βελτιώνει πολύ την αποτελεσματικότητα και την ακρίβεια του classifier. Ο χρόνος δεν επηρεάζεται αφού η πράξεις που γίνονται είναι αμελητέου χρόνου.

Εκτέλεση με βάρος : accuracy 93.3333333333%

Εκτέλεση χωρίς βάρος : accuracy 82.3333333333%

3) Αξιοποίηση τίτλου κατά την κατηγοριοποίηση:

Χρησιμοποιώντας τον τίτλο κατά την κατηγοριοποίηση εντοπίζεται μία μικρή βελτίωση της ακρίβειας. Για να αξιοποιήσουμε με καλύτερο τρόπο τον τίτλο δώσαμε

περισσότερο βάρος. Πιο συγκεκριμένα, αντιμετωπίζουμε τον τίτλο με διπλάσιο βάρος σε σχέση με το υπόλοιπο κείμενο. Όταν χρησιμοποιήσαμε τον τίτλο με περισσότερο από 3πλάσιο βάρος εντοπίσαμε αντίθετα αποτελέσματα.

Ηλίας-Ελιας Μπαρμπαρ Α.Μ. 1115 2012 00118
Χριστίνα Κατσαρλίνου Α.Μ. 1115 2015 00068