

Μηχανική Μάθηση Στην Αναγνώριση Ορθογραφικών Λαθών Λέξεων Εντός Λεξιλογίου Στην Ελληνική Γλώσσα

Νίκη Καπετανοπούλου

ΑΜ: Π2013120

Επιβλέπουσα: Κάτια Κερμανίδου

Τριμελής Επιτροπή

► Επιβλέπουσα:

Κάτια Κερμανίδου, Επίκουρη Καθηγήτρια,
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

► Τριμελής Επιτροπή:

Κάτια Κερμανίδου, Επίκουρη Καθηγήτρια,
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Φοίβος Μυλωνάς, Επίκουρος Καθηγητής,
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Σπύρος Σιούτας, Αναπληρωτής Καθηγητής,
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Σκοπός

- ▶ Ο στόχος της παρούσας πτυχιακής εργασίας είναι η ανάλυση διάφορων κατηγοριών αλγορίθμων μηχανικής μάθησης, η εφαρμογή τους σε δεδομένα που αφορούν κείμενα ελληνικής γλώσσας ηλεκτρονικών εφημερίδων και η ανάδειξη του αλγορίθμου με την καλύτερη απόδοση, κατόπιν μιας σειράς πειραματισμών.

Ορισμός μηχανικής μάθησης

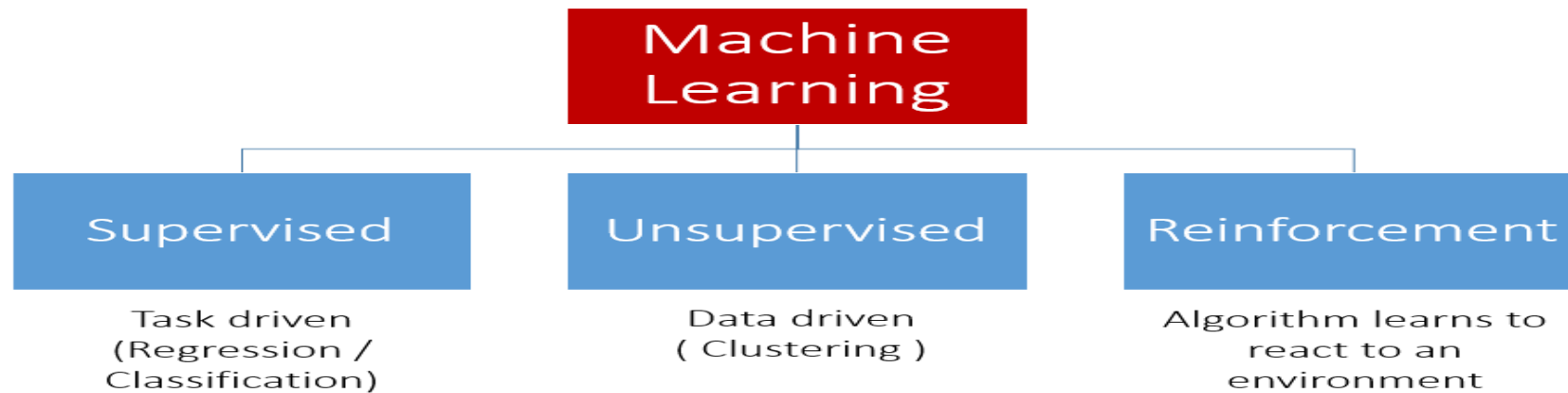
Η «μηχανική μάθηση» (machine learning) αποτελεί έναν κλάδο της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι οι μηχανές πρέπει να είναι ικανές να μαθαίνουν και να προσαρμόζονται μέσω της εμπειρίας.

Σύμφωνα με τον Mitchell, «Δεδομένης μιας εργασίας T , ενός μέτρου απόδοσης P και μιας εμπειρίας E , μία μηχανή 'μαθαίνει' αν το σύστημα βελτιώνει την απόδοσή του P στην εργασία T , ακολουθώντας την εμπειρία E ».

Είδη μηχανικής μάθησης

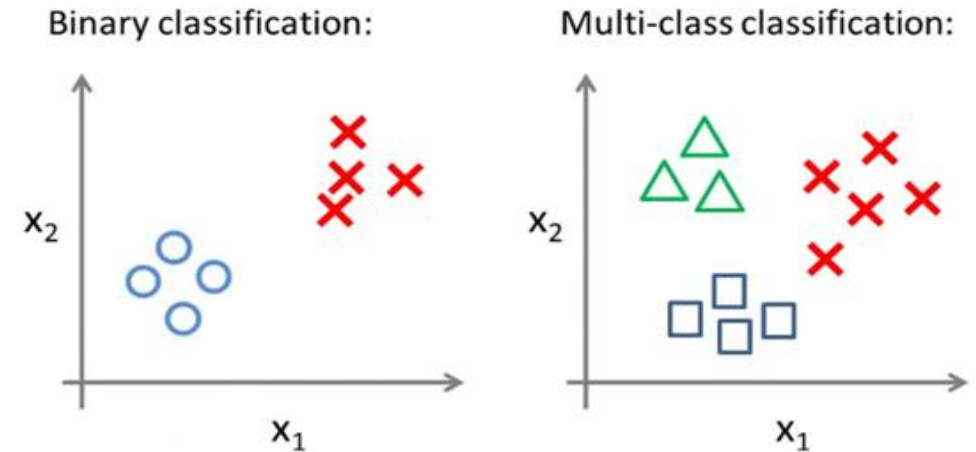


Types of Machine Learning



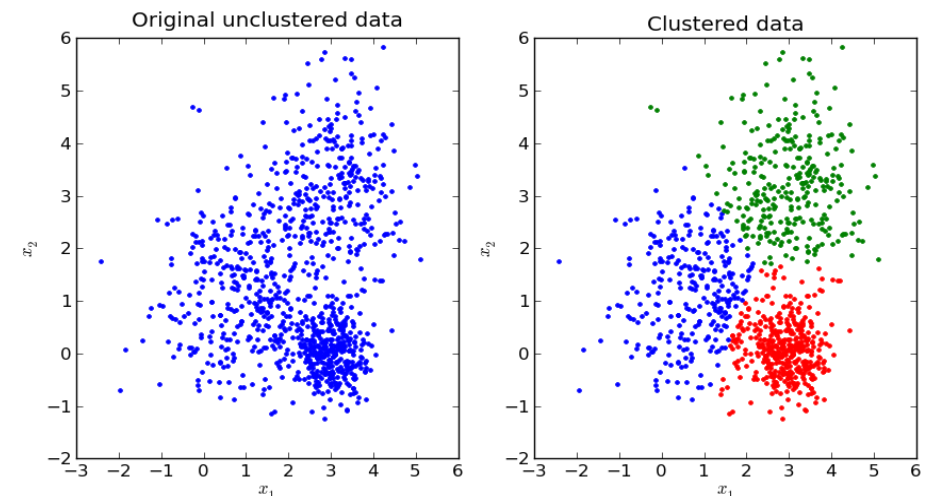
Κατηγορίες προβλημάτων μηχανικής μάθησης

- ▶ Κατηγοριοποίηση (classification)
- ▶ Συσταδοποίηση (clustering)
- ▶ Παλινδρόμηση (regression)
- ▶ Εκτίμηση πυκνότητας (density estimation)
- ▶ Μείωση διαστατικότητας (dimensionality reduction)



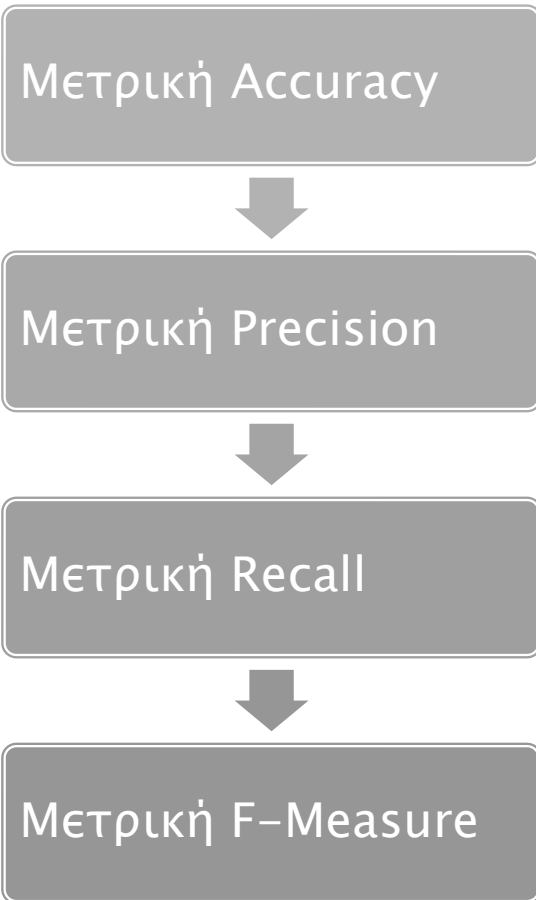
Εικόνα2

Παράδειγμα κατηγοριοποίησης Πηγή: www.holehouse.org



Εικόνα4 Παράδειγμα συσταδοποίησης Πηγή: <http://pypr.sourceforge.net/kmeans.html>

Αξιολόγηση μοντέλων μηχανικής μάθησης



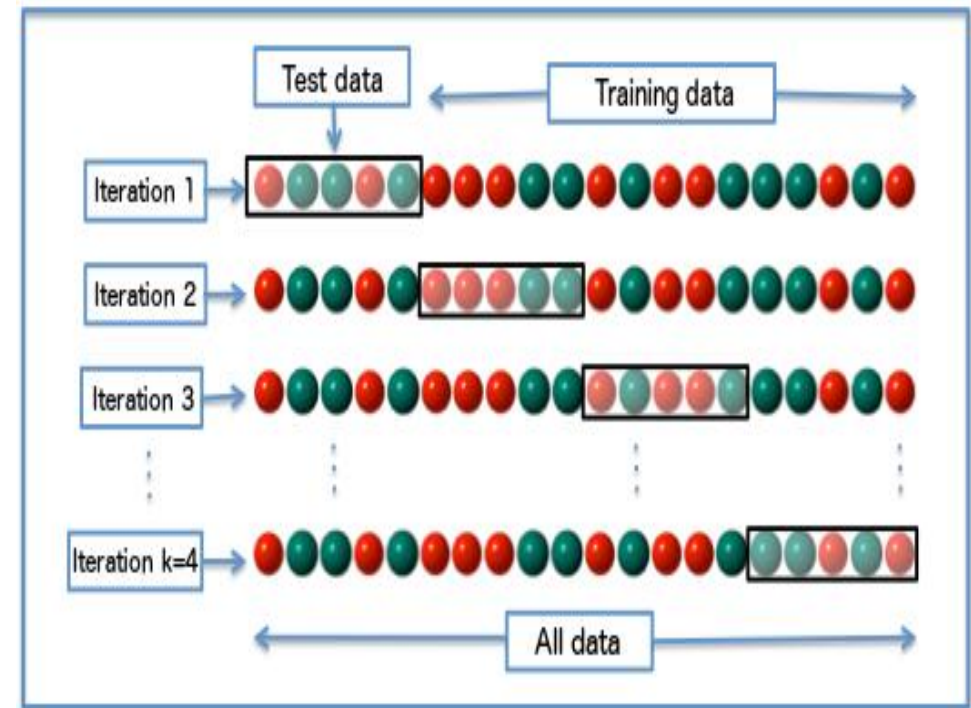
Πίνακας1

Πίνακας σύγχυσης για την αξιολόγηση ταξινομητών

		Κλάση πρόβλεψης	
		Απώλεια	Μη απώλεια
Πραγματική κλάση	Απώλεια	TP	FN
	Μη απώλεια	FP	TN

Cross-validation

- ▶ Το cross-validation είναι μία μέθοδος αξιολόγησης μοντέλων μηχανικής μάθησης, όπου εκτιμάται η ικανότητα γενίκευσης των αποτελεσμάτων μίας στατιστικής ανάλυσης σε ένα ανεξάρτητο σύνολο δεδομένων.
- ▶ Χρησιμοποιείται, κυρίως, σε προβλήματα πρόβλεψης, όπου ο στόχος είναι να εκτιμηθεί ο βαθμός ακρίβειας που θα έχει το μοντέλο πρόβλεψης στην πράξη.
- ▶ Μία πολύ γνωστή κατηγορία του cross-validation είναι το k-foldcross-validation.

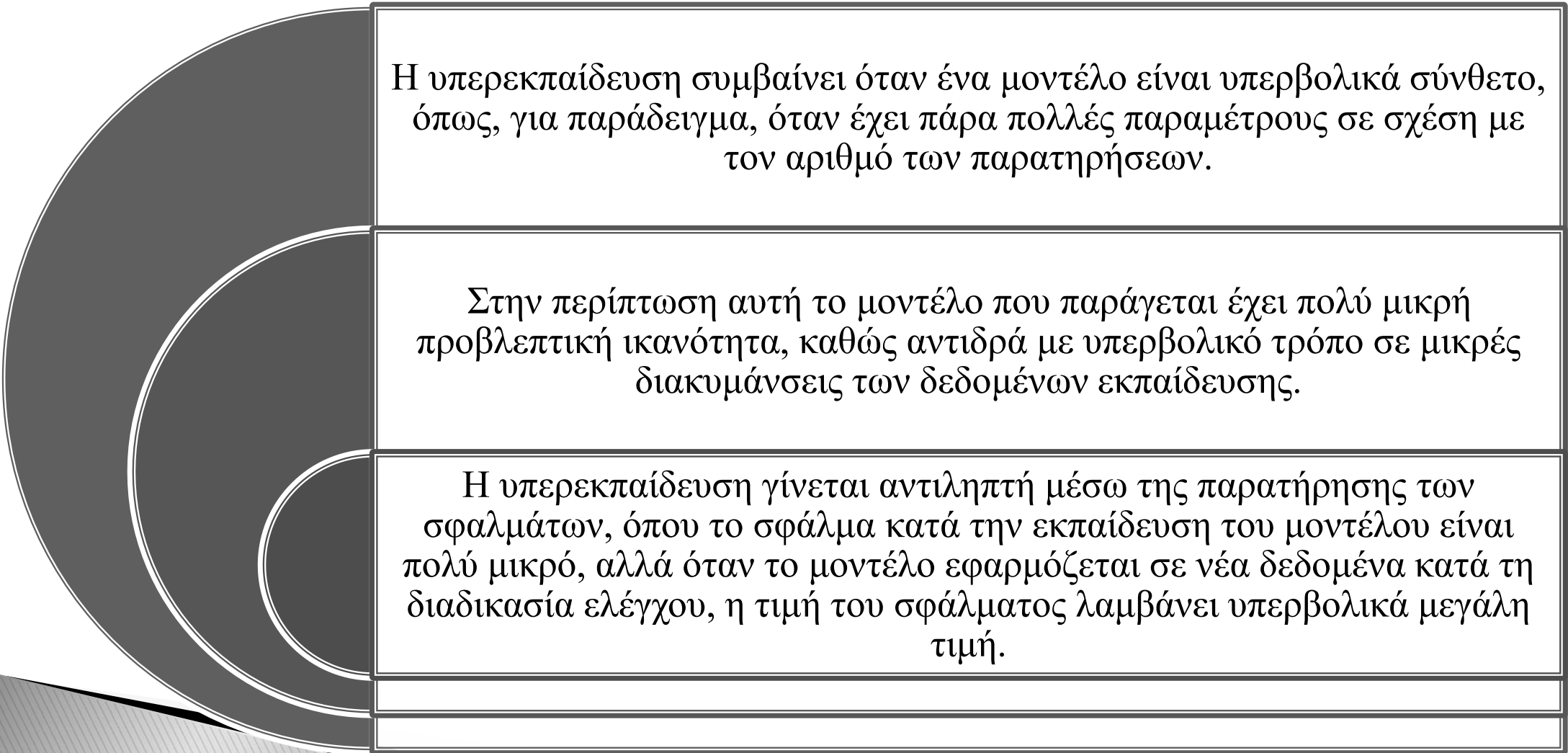


Εικόνα6

Παράδειγμα k-fold cross-validation με k=4

Πηγή: [https://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))

Το φαινόμενο της υπερεκπαίδευσης



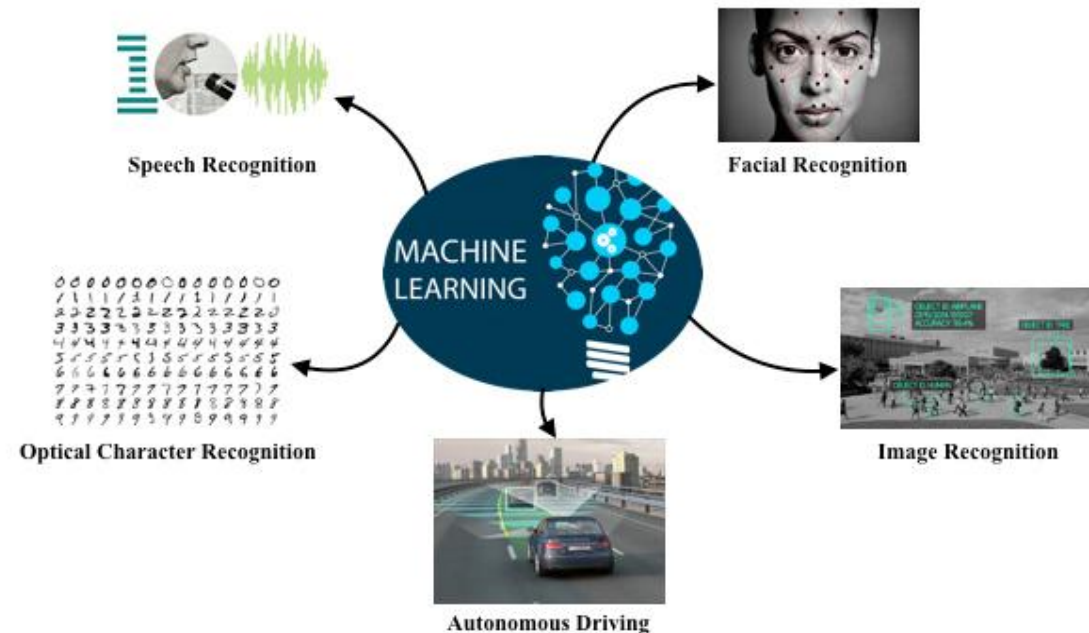
Η υπερεκπαίδευση συμβαίνει όταν ένα μοντέλο είναι υπερβολικά σύνθετο, όπως, για παράδειγμα, όταν έχει πάρα πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων.

Στην περίπτωση αυτή το μοντέλο που παράγεται έχει πολύ μικρή προβλεπτική ικανότητα, καθώς αντιδρά με υπερβολικό τρόπο σε μικρές διακυμάνσεις των δεδομένων εκπαίδευσης.

Η υπερεκπαίδευση γίνεται αντιληπτή μέσω της παρατήρησης των σφαλμάτων, όπου το σφάλμα κατά την εκπαίδευση του μοντέλου είναι πολύ μικρό, αλλά όταν το μοντέλο εφαρμόζεται σε νέα δεδομένα κατά τη διαδικασία ελέγχου, η τιμή του σφάλματος λαμβάνει υπερβολικά μεγάλη τιμή.

Εφαρμογές μηχανικής μάθησης

- ▶ Αναγνώριση φωνής.
- ▶ Υπολογιστική όραση.
- ▶ Βιολογία.
- ▶ Έλεγχος ρομπότ.
- ▶ Εφαρμογή σε εμπειρικές επιστήμες.



Εικόνα 7

Εφαρμογές της μηχανικής μάθησης

Πηγή: https://www1.in.tum.de/lehrstuhl_1/people/people-archive/74-research/research-topics/896-machine-learning-applications

Ορθογραφικά λάθη στην ελληνική γλώσσα

- ▶ Στην περίπτωση της ελληνικής γλώσσας, λόγω του λεξιλογικού της πλούτου, υπάρχουν περισσότεροι από ένας τρόποι για την προφορά ενός φωνήεντος, που δημιουργούν ασάφειες, οι οποίες, αν δεν έχουμε επαρκείς πληροφορίες για τα συμφραζόμενα της πρότασης ή ολόκληρου του κειμένου, είναι δύσκολο να ξεπεραστούν.
- ▶ Τα ορθογραφικά λάθη στην περίπτωση των ομοφώνων είναι ένα συχνό φαινόμενο κατά τη συγγραφή κειμένων στην ελληνική γλώσσα.

Πίνακας2
Ομόφωνα στην ελληνική γλώσσα

Λέξη 1	Νόημα	Λέξη 2	Νόημα
λύπη	συναίσθημα (ουσιαστικό)	λείπει	απουσιάζει (ρήμα)
καλή	καλή (ουσιαστικό, ενικόςαριθμός, θηλυκούγένους)	καλοί	καλοί (ουσιαστικό, πληθυντικόςαριθμός, αρσενικούγένους)
πίνετε	πίνετε (ρήμα, ενεργητικήφωνή, πληθυντικόςαριθμός, 2°πρόσωπο)	πίνεται	πίνεται (ρήμα, παθητικήφωνή, ενικόςαριθμός, 3°πρόσωπο)

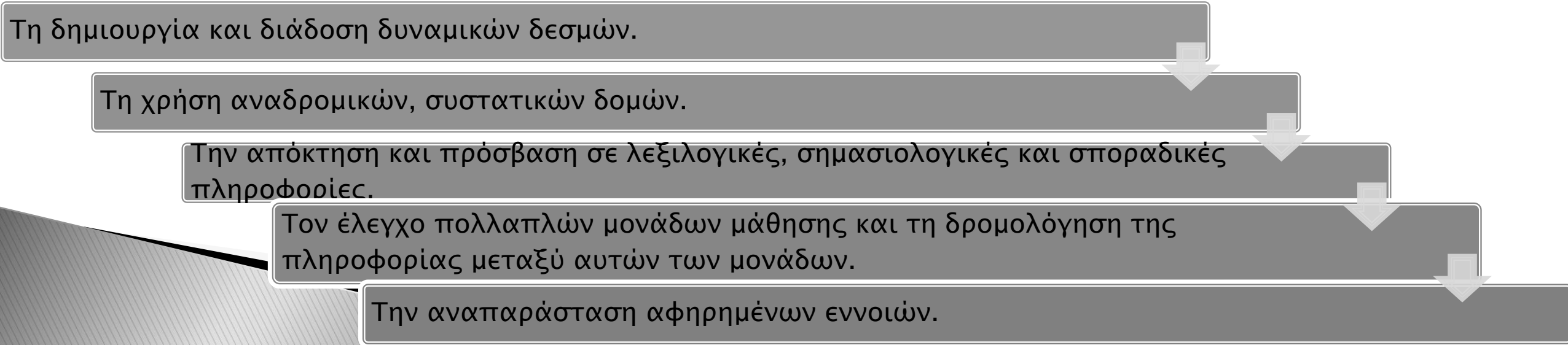
Τεχνικές αυτόματης αναγνώρισης ορθογραφικών λαθών

- ▶ Τα παραδοσιακά εργαλεία διόρθωσης ορθογραφικών λαθών βασίζονται σε λεξικά προκειμένου να εντοπίσουν αστοχίες στην ορθογραφία των λέξεων.
- ▶ Αντικατάσταση της λέξης με τη λανθασμένη ορθογραφία με τη σωστή λέξη από ένα λεξικό, βάσει μίας μετρικής ομοιότητας.
- ▶ Διαδικασία αναζήτησης του κοντινότερου γείτονα για την αυτόματη αναγνώριση ορθογραφικών λαθών.
- ▶ Χρήση ενός πιθανοτικού μοντέλου για την παραγωγή της λανθασμένης λέξης από την αντίστοιχη σωστή.
- ▶ μέθοδο που διορθώνει ταυτόχρονα λάθη μη-λέξεων και πραγματικών λέξεων.
- ▶ Window-based αλγόριθμος, όπου χρησιμοποιούνται δύο τύποι χαρακτηριστικών: τα συμφραζόμενα και οι συμφράσεις.

Εφαρμογή της μηχανικής μάθησης σε προβλήματα επεξεργασίας φυσικής γλώσσας

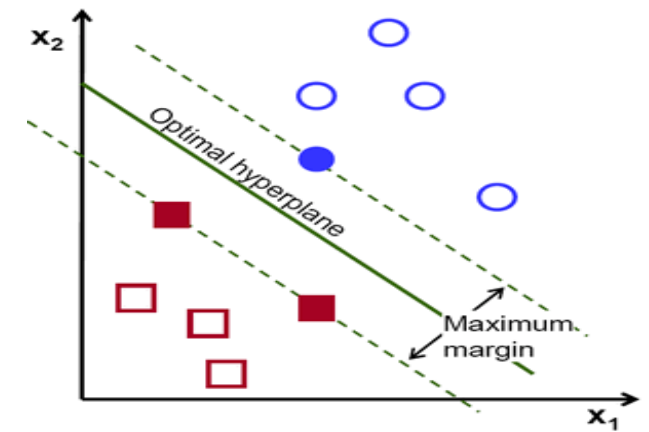
Η Επεξεργασία Φυσικής Γλώσσας (NaturalLanguageProcessing - NLP) αναφέρεται σε ένα εύρος θεωρητικά υποκινούμενων υπολογιστικών τεχνικών για την αυτόματη ανάλυση και αναπαράσταση της ανθρώπινης γλώσσας, έτσι ώστε να διευκολύνεται η επικοινωνία μεταξύ ανθρώπων και μηχανών, να βελτιώνεται η επικοινωνία μεταξύ των ανθρώπων ή απλά να εκτελείται η επεξεργασία του προφορικού ή γραπτού λόγου.

Η Επεξεργασία Φυσικής Γλώσσας απαιτεί :



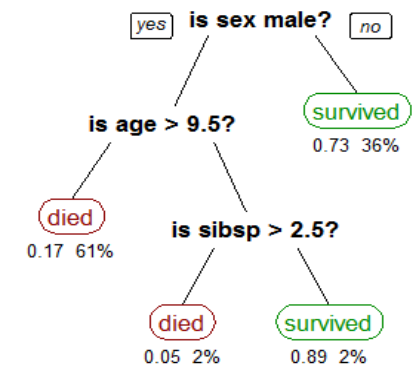
Τεχνικές μηχανικής μάθησης για την αναγνώριση ορθογραφικών λαθών

- ▶ Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM)
- ▶ Δέντρα Απόφασης (Decision Trees - DT)
- ▶ Τυχαία Δάση (Random Forests)
- ▶ Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks – ANNs)
- ▶ Μέθοδος Naïve Bayes
- ▶ Αλγόριθμος κ-κοντινότερων γειτόνων (k-Nearest Neighbors - k-NN)
- ▶ Αλγόριθμοι βασισμένοι σε στιγμιότυπα (instance-based algorithms)
- ▶ Αλγόριθμος ενίσχυσης (boosting)



Εικόνα9 Παράδειγμα Μοντέλου SVM

Πηγή: <http://opencv-python-tutroals.readthedocs.io/en/latest/>



Εικόνα10

Παράδειγμα δέντρου απόφασης

Πηγή: https://en.wikipedia.org/wiki/Decision_tree_learning

Περιβάλλον WEKA

- ▶ Το WEKA (Waikato Environment for Knowledge Analysis) αποτελεί λογισμικό μηχανικής μάθησης που είναι γραμμένο σε Java και αναπτύχθηκε στο πανεπιστήμιο Waikato της Νέας Ζηλανδίας
- ▶ Το WEKA υποστηρίζει διάφορες εργασίες της εξόρυξης δεδομένων και της μηχανικής μάθησης, όπως είναι η προεπεξεργασία των δεδομένων, η συσταδοποίηση, η ταξινόμηση, η παλινδρόμηση, η οπτικοποίηση των δεδομένων και η επιλογή χαρακτηριστικών.
- ▶ Όλες οι τεχνικές του WEKA λειτουργούν με την προϋπόθεση ότι τα δεδομένα είναι διαθέσιμα σε κάποιο προσπελάσιμο αρχείο, όπου κάθε εγγραφή περιγράφεται από ένα συγκεκριμένο αριθμό χαρακτηριστικών, αριθμητικών ή μη.



Εικόνα 15

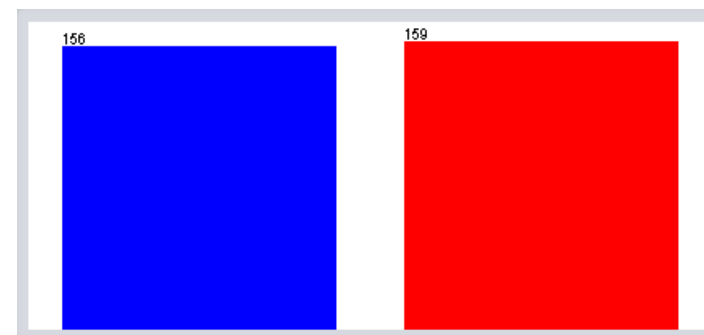
Λογότυπο WEKA

Πηγή: www.google.gr

Σύνολο δεδομένων που χρησιμοποιήθηκε

- ▶ Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν, όπως έχει αναφερθεί, ομόηχα ρήματα με κατάληξη –τε και –ται. Δηλαδή, πρόκειται για λέξεις που έχουν το ίδιο θέμα, αλλά διαφοροποιούνται ως προς την κατάληξη.
- ▶ Οι λέξεις αυτές προήλθαν από κείμενα ηλεκτρονικών εφημερίδων και, πιο συγκεκριμένα, των εφημερίδων ‘Ελευθεροτυπία’, ‘Τα ΝΕΑ’ και ‘Το ΒΗΜΑ’. Ανατρέχοντας σε διαφορετικά άρθρα, έγινε η αναζήτηση προτάσεων που περιείχαν λέξεις που είχαν ενδιαφέρον.

Οι προτάσεις αυτές συνέθεσαν το τελικό κείμενο το οποίο αποτελείται από 3754 λέξεις.

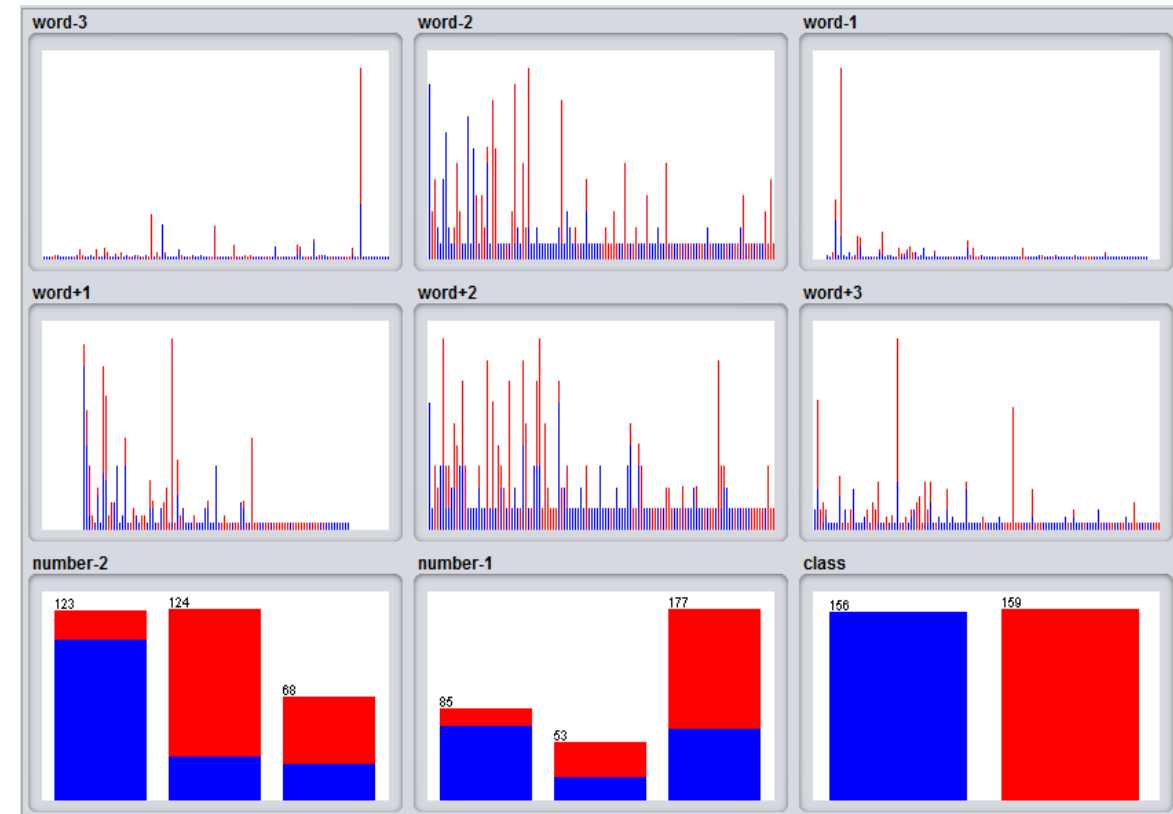


	Ετικέτα	Αριθμός εγγραφών
1	ται	156
2	τε	159
Σύνολο		315

λέξη-3	λέξη-2	λέξη-1	λέξη+ 1	λέξη+ 2	λέξη+ 3	αριθμός- 2	αριθμός- 1	ται/τε
το	ίδιο	φαινόμενο	και	για	τις	ΕΝ	ΕΝ	ται

Σύνολο δεδομένων που χρησιμοποιήθηκε

- ▶ Στην εικόνα φαίνονται τα ιστογράμματα των εννιά χαρακτηριστικών, όπως τα εμφάνισε το πρόγραμμα WEKA.
- ▶ Για κάθε χαρακτηριστικό παρατηρείται η συχνότητα εμφάνισης των διαφορετικών τιμών του αναλογικά με την εκάστοτε τιμή της κλάσης εξόδου.
- ▶ Το μπλε χρώμα αντιστοιχεί στην τιμή ‘ται’ της κλάσης εξόδου, ενώ το κόκκινο χρώμα στην τιμή ‘τε’ αντίστοιχα.



Εικόνα16

Ιστογράμματα των χαρακτηριστικών του συνόλου δεδομένων

Αλγόριθμοι που χρησιμοποιήθηκαν

- ▶ **NaiveBayes:** στον ταξινομητή Naïve Bayes και χρησιμοποιεί κλάσεις εκτιμητών. Πρόκειται για πιθανοτικό, μη-παραμετρικό μοντέλο.
- ▶ **RandomForest:** παράδειγμα συλλογικής μάθησης και χρησιμοποιείται σε προβλήματα ταξινόμησης, όπου κατασκευάζει ένα πλήθος δέντρων απόφασης κατά τη διαδικασία της εκπαίδευσης και εξάγει σαν αποτέλεσμα την κλάση που εμφανίζεται πιο συχνά στο σύνολο των παραγόμενων δέντρων.
- ▶ **Kstar:** Πρόκειται για έναν ταξινομητή που βασίζεται στα στιγμιότυπα (instance-based), όπου η κλάση εξόδου ενός στιγμιοτύπου βασίζεται στην κλάση των στιγμιοτύπων που είναι παρόμοια με αυτό, όπου η ομοιότητα διαπιστώνεται με τη χρήση μιας συνάρτησης ομοιότητας.
- ▶ **DecisionStump:** Το decision stump είναι ένα μοντέλο μηχανικής μάθησης που αποτελείται από ένα δέντρο απόφασης ενός επιπέδου. Πιο συγκεκριμένα, πρόκειται για ένα δέντρο απόφασης που περιλαμβάνει έναν εσωτερικό κόμβο (τη ρίζα του δέντρου) ο οποίος συνδέεται απευθείας με τους τελικούς κόμβους (τα φύλλα του δέντρου).
- ▶ **ID3:** Πρόκειται για κλάση του WEKA που χρησιμοποιείται για την κατασκευή ενός μη-κλαδεμένου (unpruned) δέντρου απόφασης βάσει του αλγορίθμου ID3.
- ▶ **J48:** Κλάση για τη δημιουργία ενός C4.5 δέντρου απόφασης, το οποίο μπορεί να έχει ή να μην έχει υποστεί κλάδεμα (pruning.)

Αποτελέσματα πειραμάτων

Στον πίνακα που ακολουθεί παρουσιάζεται η απόδοση των διαφορετικών αλγορίθμων στο παράθυρο [-3,3,αριθμός], δηλαδή όταν λαμβάνουμε υπόψη όλα τα χαρακτηριστικά του συνόλου δεδομένων.

Πίνακας3 Αξιολόγηση μοντέλων όταν λαμβάνονται υπόψη όλα τα χαρακτηριστικά του συνόλου δεδομένων

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3,αριθμός]	-ται	0,902	0,769	0,830
		-τε	0,802	0,918	0,856
RandomForest	[-3,3,αριθμός]	-ται	0,885	0,744	0,808
		-τε	0,783	0,906	0,840
KStar	[-3,3,αριθμός]	-ται	0,919	0,795	0,852
		-τε	0,822	0,931	0,873
DecisionStump	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798
ID3	[-3,3,αριθμός]	-ται	0,901	0,889	0,895
		-τε	0,938	0,945	0,941
J48	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798

Αποτελέσματα πειραμάτων

Στον πίνακα παρουσιάζονται οι τιμές απόδοσης των μοντέλων όταν αυτά εφαρμόζονται στα χαρακτηριστικά που αντιστοιχούν στις τρεις λέξεις πριν και μετά της λέξης-στόχου. Έχουν αφαιρεθεί, δηλαδή, τα χαρακτηριστικά που αφορούν τον αριθμό των δύο προηγούμενων λέξεων.

Πίνακας4 Αξιολόγηση μοντέλων όταν δε λαμβάνονται υπόψη τα χαρακτηριστικά που αφορούν τον αριθμό των δύο προηγούμενων λέξεων

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3]	-ται	0,939	0,788	0,857
		-τε	0,821	0,950	0,880
RandomForest	[-3,3]	-ται	0,913	0,737	0,816
		-τε	0,783	0,931	0,851
KStar	[-3,3]	-ται	0,929	0,756	0,834
		-τε	0,798	0,943	0,865
DecisionStump	[-3,3]	-ται	0,629	0,936	0,753
		-τε	0,880	0,459	0,603
ID3	[-3,3]	-ται	0,901	0,889	0,895
		-τε	0,938	0,945	0,941
J48	[-3,3]	-ται	0,926	0,481	0,633
		-τε	0,654	0,962	0,779

Αποτελέσματα πειραμάτων

Στον πίνακα, παρατηρώντας τις μετρικές Precision, Recall και F-Measure, διαπιστώνουμε ότι η καλύτερη απόδοση επιτυγχάνεται από το μοντέλο ID3.

Πίνακας 5 Αξιολόγηση μοντέλων όταν λαμβάνεται ο καλύτερος συνδυασμός χαρακτηριστικών και αλγορίθμου

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3]	-ται	0,939	0,788	0,857
		-τε	0,821	0,950	0,880
RandomForest	[-3,3]	-ται	0,913	0,737	0,816
		-τε	0,783	0,931	0,851
KStar	[-3,3,αριθμός]	-ται	0,919	0,795	0,852
		-τε	0,822	0,931	0,873
DecisionStump	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798
ID3	[-2,3,αριθμός]	-ται	0,930	0,904	0,917
		-τε	0,945	0,960	0,953
J48	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798

Αποτελέσματα πειραμάτων

Πίνακας6 Αξιολόγηση του μοντέλουID3 για όλους τους συνδυασμούς των χαρακτηριστικών

Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
[-3,3,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,2,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,1,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,0,αριθμός]	-ται	0,900	0,887	0,894
	-τε	0,937	0,944	0,940
[-2,3,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953
[-2,2,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953
[-2,1,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953
[-2,0,αριθμός]	-ται	0,926	0,900	0,913
	-τε	0,946	0,961	0,953

Αποτελέσματα πειραμάτων

Πίνακας 6 Αξιολόγηση του μοντέλου D3 για όλους τους συνδυασμούς των χαρακτηριστικών

Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
[-1,3,αριθμός]	-ται	0,808	0,778	0,792
	-τε	0,881	0,899	0,890
[-1,2,αριθμός]	-ται	0,778	0,778	0,778
	-τε	0,881	0,881	0,881
[-1,1,αριθμός]	-ται	0,789	0,818	0,804
	-τε	0,914	0,898	0,906
[-1,0,αριθμός]	-ται	0,753	0,810	0,780
	-τε	0,891	0,854	0,872
[0,3,αριθμός]	-ται	0,841	0,828	0,835
	-τε	0,885	0,895	0,890
[0,2,αριθμός]	-ται	0,803	0,828	0,815
	-τε	0,885	0,867	0,876
[0,1,αριθμός]	-ται	0,861	0,818	0,839
	-τε	0,835	0,874	0,854
[-3,3]	-ται	0,901	0,889	0,895
	-τε	0,928	0,945	0,941

Αποτελέσματα πειραμάτων

Πίνακας 7 Αποτελέσματα πριν και μετά της χρήσης SMOTE (Παράθυρο (-3,3,Αριθμός)

Αλγόριθμος	SMOTE	Κλάσεις	Precision	Recall	F-measure
NaiveBayes	No	-ΤΑΙ	0,899	0,744	0,814
		-ΤΕ	0,669	0,662	0,753
NaiveBayes	Yes	-ΤΑΙ	0,902	0,869	0,830
		-ΤΕ	0,802	0,918	0,856
ID3	No	-ΤΑΙ	0,914	0,941	0,928
		-ΤΕ	0,922	0,787	0,904
ID3	Yes	-ΤΑΙ	0,901	0,889	0,895
		-ΤΕ	0,938	0,945	0,941

Σχολιασμός αποτελεσμάτων

- ▶ Ο αλγόριθμος ID3 παρουσίασε την καλύτερη απόδοση στο παράθυρο [-2,3,αριθμός] (την ίδια απόδοση παρουσιάζει και στα παράθυρα [-2,2,αριθμός], [-2,1,αριθμός]).
- ▶ Η απόδοση του μοντέλου σημειώνει ενίοτε σημαντικές διακυμάνσεις όταν αυτό τροφοδοτείται με διαφορετικά σετ χαρακτηριστικών.
- ▶ Τα μοντέλα NaïveBayes, RandomForest και KStar απέδωσαν σε βαθμό συγκρίσιμο μεταξύ τους, αλλά πολύ χειρότερα συγκριτικά με το μοντέλο ID3

Πίνακας8 Παραδείγματα που ταξινομήθηκαν λανθασμένα από τον ταξινομητή ID3

λέξη-3	λέξη-2	λέξη-1	λέξη+1	λέξη+2	λέξη+3	αριθμός-2	αριθμός-1	Πραγματικ ή κλάση	Πρόβλεψ η
ενώ	οι	βάσεις	να	εκδοθούν	το	ΠΛ	ΠΛ	ται	τε
-	να	μη	πολιτικό	άλλοθι	σε	-	-	ται	τε
-	-	-	τις	μνημονιακ ές	υποχρεώ σεις	-	-	τε	ται
-	γνωρίζετε	ότι	λάθος	προς	τα	ΠΛ	-	τε	ται

Σχολιασμός αποτελεσμάτων

- ▶ Προκειμένου να αυξηθεί η ακρίβεια και η αποτελεσματικότητα του ταξινομητή, θα πρέπει να ληφθούν υπόψη όλες οι περιπτώσεις κατά τις οποίες αστοχεί ο ταξινομητής να προβλέψει σωστά την κλάση εξόδου, οι πιο χαρακτηριστικές από τις οποίες αναφέρθηκαν παραπάνω.
- ▶ Αρχικά, μελλοντικά, είναι θεμιτό να τροφοδοτούνται τα μοντέλα με μεγαλύτερο αριθμό δεδομένων, έτσι ώστε να είναι, αναλογικά, πολύ μικρός ο αριθμός των εγγραφών από τις οποίες απουσιάζουν οι τιμές ορισμένων χαρακτηριστικών.
- ▶ Αρκετό ενδιαφέρον παρουσιάζει η περίπτωση των λέξεων με κατάληξη ‘τε’, όπου ο ταξινομητής αστοχούσε συχνά λόγω της ιδιαιτερότητας των ρημάτων με αυτήν την κατάληξη.

Σχολιασμός αποτελεσμάτων

- ▶ Θα μπορούσαμε σε μία μελλοντική επέκταση της πτυχιακής εργασίας, να προσθέσουμε επιπλέον χαρακτηριστικά τα οποία να αφορούν το είδος της πρότασης, δηλαδή αν πρόκειται για κατάφαση ή ερώτηση, καθώς και την έγκλιση, δηλαδή αν έχουμε οριστική, προστακτική ή υποτακτική.
- ▶ Τέλος, τα σημεία στίξης, και συγκεκριμένα το κόμμα, έπαιξε σημαντικό ρόλο αν υπήρχε πριν τη λέξη-στόχο, κυρίως σε ρήματα με κατάληξη ‘τε’, επομένως θα μπορούσε να συμπεριληφθεί σαν ένα επιπλέον χαρακτηριστικό.

Συμπεράσματα

Η μέθοδος που προτείνεται έρχεται να δώσει λύση στην περίπτωση της διόρθωσης λέξεων οι οποίες έχουν λανθασμένη ορθογραφία με βάση τα συμφραζόμενα, ωστόσο αποτελούν έγκυρες λέξεις που μπορούν να βρεθούν σε οποιοδήποτε λεξικό.

Στην παρούσα πτυχιακή εργασία μελετήθηκε μόνο ένα ζεύγος τέτοιων λέξεων, ωστόσο θα μπορούσαν να εξεταστούν μελλοντικά και άλλες περιπτώσεις ομόφωνων, δηλαδή λέξεων που προφέρονται ακριβώς το ίδιο αλλά έχουν διαφορετική κατάληξη.

Μελλοντικές επεκτάσεις

- ▶ Μελλοντικά, θα μπορούσαν να γίνουν πειραματισμοί και με άλλους τύπους χαρακτηριστικών που αφορούν τα συμφραζόμενα.
- ▶ Επιπλέον, σαν μελλοντική επέκταση της παρούσας εργασίας προτείνεται η μελέτη και εφαρμογή και άλλων αλγορίθμων μηχανικής μάθησης, όπως είναι τα μοντέλα SVM και τα νευρωνικά δίκτυα.

Βιβλιογραφία

- ▶ T. Mitchell, «Machine Learning,» 1997.
- ▶ T. Mitchell, «The Discipline of Machine Learning,» αρ. School of Computer Science, Carnegie Mellon University, 2006.
- ▶ «Data mining,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Data_mining.
- ▶ P.N. Tan, M. Steinbach, V. Kumar, «Introduction to Data Mining,» 2006.
- ▶ «Forbes,» [Ηλεκτρονικό]. Available: <https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#298c74b794c9>.
- ▶ «Machine learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Machine_learning.
- ▶ «Supervised learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Supervised_learning.
- ▶ «Reinforcement learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Reinforcement_learning.
- ▶ S. Kotsiantis, «Supervised Machine Learning: A Review of Classification Techniques,» *Informatica*, τόμ. 31, pp. 249-268, 2007.
- ▶ A.K. Jain, M.N. Murty, P.J. Flynn, «Data clustering: a review,» *ACM computing surveys (CSUR)*, 1999.
- ▶ «Confusion matrix,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Confusion_matrix.
- ▶ A.A. Lunsford, K.J. Lunsford, «Mistakes Are a Fact of Life: A National Comparative Study,» *College Composition and Communication*, τόμ. vol. 59, αρ. no. 4, pp. 781-806, 2008.
- ▶ Ch. Desmet, R. Balthazor, «Finding Patterns in Textual Corpora: Data Mining, Research, and Assessment in First-year Composition,» *Conference Proceedings, Computers and Writing 2005: New Writing and Computer Technologies*, pp. 1-8, 2005.
- ▶ M. Flor, Y. Futagi, «On Using Context for Automatic Correction of Non-word Misspellings in Student Essays,» *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, BEA-7 (at NAACL HLT 2012), Montreal, Canada*, pp. 105-115, 2012.
- ▶ M. Flor, «Four types of context for automatic spelling correction,» *Educational Testing Service, Rosedale Road, Princeton, NJ, 08541, USA*, 2012.

Σας ευχαριστώ!