ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΟΛΙΚΟΤΗΤΑ ΚΕΙΜΕΝΟΥ

Διπλωματική Εργασία της

ΜΑΡΙΑ ΟΡΦΑΝΟΥΔΑΚΗ (ΑΕΜ: 03/2436)

Επιβλέπων Καθηγητής: ΦΩΤΗΣ ΚΟΚΚΟΡΑΣ

ΘΕΣΣΑΛΟΝΙΚΗ
ΑΠΡΙΛΙΟΣ 2011

Πρόλογος

Ένα θέμα που μας απασχολεί όλους είναι το «Τι σκέφτονται οι άλλοι;». Κατά τη διάρκεια αυτής της εργασίας θα προσπαθήσουμε να ερευνήσουμε μία μέθοδο η οποία μπορεί σε πολλές περιπτώσεις να απαντήσει σε αυτή την ερώτηση. Η μέθοδος αυτή ονομάζεται ανάλυση συναισθήματος και το εφαρμογή της που μας ενδιαφέρει περισσότερο είναι η πολικότητα ενός κειμένου, αν, δηλαδή, το κείμενο βγάζει θετικά ή αρνητικά συμπεράσματα. Θα αναλυθεί ολόκληρο το πεδίο που καλύπτει αυτό το θέμα, θα παρουσιαστούν διάφοροι αλγόριθμοι και προγράμματα που κάνουν ανάλυση συναισθήματος. Στη συνέχεια θα προσπαθήσουμε να εφαρμόσουμε τα παραπάνω σε ένα πείραμα που θα διεξάγουμε βγάζοντας αποτελέσματα από πραγματικές κριτικές ταινιών. Τέλος, θα μπορέσουμε να διαπιστώσουμε τι μπορεί να μας προσφέρει αφού μελετήσουμε ορισμένες περιπτώσεις όπου η ανάλυση συναισθήματος προέβλεψε ορισμένα περιστατικά.

Η πτυχιακή αυτή εργασία, έγινε στο τμήμα Πληροφορικής του Τ.Ε.Ι. Θεσσαλονίκης με επιβλέποντα καθηγητή τον κ. Φώτη Κόκκορα, τον οποίο θα ήθελα να ευχαριστήσω για την πολύτιμη βοήθειά του και γιατί σε μια κρίσιμη στιγμή μου έδωσε το κουράγιο να συνεχίσω. Επίσης θα ήθελα να ευχαριστήσω την οικογένειά μου και τις φίλες μου για την ψυχολογική υποστήριξή και κυρίως τον θείο μου Γεώργιο Ψωμαδάκη, χωρίς τη βοήθεια του οποίου δεν θα ήμουν εδώ αυτή τη στιγμή. Τέλος, ευχαριστώ απ' τα βάθη της καρδιάς μου τον αδερφό μου Μιχάλη Ορφανουδάκη και την φίλη μου Ελίνα Ανδρουλάκη για τη βοήθεια που μου προσέφεραν!

ΟΡΦΑΝΟΥΔΑΚΗ ΜΑΡΙΑ

Ημερομηνία 20/04/2011

Περιεχόμενα

1 Εισαγωγή

Γενικά οι πληροφορίες κειμένου μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες, τα γεγονότα και τις απόψεις. Τα γεγονότα είναι αντικειμενικές διατυπώσεις για οντότητες, γεγονότα και τα χαρακτηριστικά τους. Οι απόψεις είναι συνήθως υποκειμενικές διατυπώσεις οι οποίες περιγράφουν το συναίσθημα, αξιολογήσεις ή την προτίμησή μας σε οντότητες, γεγονότα και τις ιδιότητες τους. Η έννοια της άποψης είναι πολύ γενική. Σ' αυτή την πτυχιακή θα εστιάσουμε στην έκφραση απόψεων οι οποίες μπορούν να διαχωρίσουν το συναίσθημα σε θετικό ή αρνητικό.

Πάνω στην επεξεργασία πληροφοριών κειμένου έχουν γίνει κατά καιρούς έρευνες που εστιάζουν στην επεξεργασία και ανάκτηση αντικειμενικών πληροφοριών όπως αναζήτηση στο διαδίκτυο, κατηγοριοποίηση κειμένου και άλλα text mining και NLP θέματα. Πολύ λίγη δουλειά είχε γίνει, όμως, για την επεξεργασία απόψεων μέχρι προσφάτως. Παρόλα αυτά οι απόψεις είναι τόσο σημαντικές που κάθε φορά που θέλουμε να πάρουμε μια απόφαση θέλουμε να ακούσομε τη γνώμη των άλλων. Αυτό δεν ισχύει μόνο για τους ανθρώπους στην καθημερινότητά τους αλλά και για τους οργανισμούς και για τις επιχειρήσεις

Ένας απ' τους κύριους λόγους που δεν υπήρχε πολύ μελέτη και έρευνα πάνω σ' αυτό το θέμα είναι το γεγονός ότι υπήρχε πολύ λίγη διαθέσιμη υποκειμενική πληροφορία μέχρι τη δημιουργία του World Wide Web. Πριν το Web, όταν κάποιος ήθελε να πάρει μια απόφαση ζητούσε την άποψη των φίλων του και της οικογένειάς του. Όταν ένας οργανισμός ήθελε να εξετάσει τις απόψεις ή τα συναισθήματα του κοινού για τα προϊόντα και τις υπηρεσίες του, έκανε δημοσκοπήσεις, ερωτηματολόγια και εξέταση συγκεντρωτικών ομάδων. Παρόλα αυτά, με το Web, ειδικά μετά το συνεχές αυξανόμενο περιεχόμενο των χρηστών τα τελευταία χρόνια, ο κόσμος άλλαξε

Το Web έχει αλλάξει δραματικά τον τρόπο που οι άνθρωποι εκφράζουν τις απόψεις τους. Τώρα μπορούν να δημοσιεύσουν κριτικές για προϊόντα σε σελίδες πωλήσεων και να εκφράσουν τη γνώμη τους για οτιδήποτε στα forum, στις ομάδες συζητήσεων ή στα blogs. Αυτή η συμπεριφορά αναπαριστά καινούριες και μετρήσιμες πηγές πληροφορίας με πολλές πρακτικές εφαρμογές. Τώρα αν κάποιος θέλει να αγοράσει ένα προϊόν, δεν

περιορίζεται πια στο να ρωτήσει τους φίλους και την οικογένειά του, επειδή υπάρχουν τόσες πολλές κριτικές στο διαδίκτυο που παρουσιάζουν τις απόψεις των χρηστών που το έχουν ήδη χρησιμοποιήσει. Έτσι και για μία εταιρία, δεν χρειάζεται πια να εξάγει δημοσκοπήσεις ή να προσλαμβάνει εξωτερικούς συμβούλους για δει τι σκέφτονται οι καταναλωτές για τα προϊόντα της ή τους ανταγωνιστές της γιατί το περιεχόμενο του διαδικτύου μπορεί να της δώσει της πληροφορίες που χρειάζεται.

Παρόλα αυτά όμως, το να βρει κάποιος πηγές και να τις ελέγξει είναι αρκετά δύσκολο θέμα γιατί υπάρχει ένας τεράστιος αριθμός από διάφορες πηγές και κάθε μία απ' αυτές μπορεί να έχει τεράστια ποσότητα κειμένου. Σε πολλές περιπτώσεις, οι απόψεις κρύβονται σε δημοσιεύσεις forum ή blogs. Είναι δύσκολο για τον αναγνώστη να βρει σχετικές πηγές, να εξάγει σχετικές προτάσεις που να περιέχουν τις απόψεις, να τις διαβάσει, να τις συνοψίσει και να τις οργανώσει σε μορφή που να μπορεί να τις χρησιμοποιήσει. Γι' αυτό το λόγο χρειάζονται συστήματα που να ανακαλύπτουν αυτόματα τις απόψεις και να τις συνοψίζουν. Η ανάλυση συναισθήματος αναπτύχθηκε μέσα απ' αυτή την ανάγκη. Είναι μία πρόκληση για τον κλάδο της επεξεργασίας φυσικής γλώσσας και το text mining. Λόγω της μεγάλης ανάγκης για πρακτικές εφαρμογές, υπάρχει μια συνεχής ανάπτυξη στην έρευνα στην ακαδημαϊκή κοινότητα και στη βιομηχανία εφαρμογών. Αυτή τη στιγμή υπάρχουν τουλάχιστον 20-30 εταιρίες που προσφέρουν υπηρεσίες ανάλυσης συναισθήματος μόνο στις ΗΠΑ.

Ο σχετικά νέος αυτός κλάδος ενδιαφέρει, όπως διαφάνηκε εξ αρχής, τις μεγάλες εταιρίες με τα παγκόσμια εμπορικά σήματα, που θέλουν να ξέρουν τι νιώθουν οι πελάτες και οι υποψήφιοι αγοραστές για τα προϊόντα και τις υπηρεσίες τους. Ήδη ομάδες εταιριών πληροφορικής έχουν αναπτύξει τέτοιου είδους λογισμικό, που αθόρυβα ψάχνουν σε blogs και σε άλλες ιστοσελίδες (π.χ. κοινωνικής δικτύωσης) για να πιάσουν το συναισθηματικό σφυγμό όσων γράφονται στο διαδίκτυο σε σχέση με κάποια επιχείρηση και το προϊόν της.

Μία σωστή ανάλυση συναισθημάτων αποτελείται από τα παρακάτω τρία συστατικά. Την ικανότητα της σωστής αξιολόγησης στην ιδέα, στο αντικείμενο αλλά και στο επίπεδο θέματος. Οι σωστοί κανόνες και η πολύ καλή γνώση του εταιρικού πελάτη και του προφίλ του από το πρόγραμμα που χρησιμοποιεί την ανάλυση συναισθήματος μπορούν να αποδώσουν ένα ακριβές αποτέλεσμα. Έτσι ένα αξιόλογο πρόγραμμα μπορεί να ξεχωρίσει τα συναισθήματα μιας σειράς συζητήσεων ενός forum

εάν αρχικά αναφέρονται σε υπαρκτό πρόβλημα/παράπονο και στη συνέχεια εάν ευθύνεται όντος η εταιρία γι' αυτό.

Ένα άλλο τεχνικό συστατικό είναι η ικανότητα του προγράμματος στην προκειμένη περίπτωση να βλέπει και να αναλύει πέρα από τις λέξεις-κλειδιά. Χρειάζεται η απαραίτητη γλωσσολογία που θα ασχοληθεί με το πλαίσιο και το νόημα, ώστε να αξιολογηθεί σωστά το συναίσθημα. Υπάρχει αρκετή πολυπλοκότητα λέξεων, και γλώσσας με άμεση επίδραση στα συναισθήματα.

Αρχικά θα αναλύσουμε την έννοια συναίσθημα ,πως επηρεάζει την καθημερινότητα αλλά και πως επηρεάζεται από αυτήν και από πότε ο άνθρωπος προσπαθεί να αναλύσει το συναίσθημα. Θα γίνει αναφορά στους τρεις βασικούς κλάδους που βασίζεται η ανάλυση συναισθήματος και ανάλυση της έννοιας. Θα γνωρίσουμε τον κλάδος της μηχανικής μάθησης ο οποίος ήταν ένας από τους λόγους ανάπτυξης της ανάλυσης συναισθημάτων. Θα αναλύσουμε την λειτουργία της ταξινόμηση πολικότητας συναισθήματος και κάποιων σημαντικών όρων. Θα εξετάσουμε τεχνικές ανάλυσης συναισθήματος και τον Ο αλγόριθμος Thumbs Up Thumbs Down.Θα αναφερθούμε και σε άλλες εφαρμογές της sentiment analysis στο διαδίκτυο, εκτός από την πολικότητα κειμένου.

Αφού ερευνήσουμε θεωρητικά τον κλάδο της ανάλυσης συναισθήματος θα επιχειρήσουμε να παρουσιάσουμε τα πιο δημοφιλή προγράμματα που υπάρχουν αυτή τη στιγμή. Μιλώντας για τα χαρακτηρίστηκα τους, τα πλεονεκτήματα και τα μειονεκτήματα και σε ορισμένες περιπτώσεις παρουσιάζοντας το γραφικό τους περιβάλλον, ελπίζουμε ο αναγνώστης να βρει αυτό που ταιριάζει περισσότερο με τα ανάγκες του.

Στη συνέχεια υλοποιείται ένα πείραμα με τη βοήθεια του Rapidminer για την κατανόηση της ανάλυσης κειμένου .Θα πάρουμε ένα σύνολο δεδομένων (datasets) και θα το κατατάξουμε σε θετικό ή αρνητικό με την ανάλυση συναισθήματος. Γίνεται μία αναφορά στο τι είναι το σύνολο δεδομένων και μία επεξήγηση στο πείραμα μέχρι να φτάσουμε στο αποτέλεσμα. Υπάρχουν αναλυτικά τα βήματα τα οποία έγιναν για την υλοποίηση του πειράματος τα οποία είναι εισαγωγή δεδομένων, επεξεργασία δεδομένων, δημιουργία λέξεων διανυσμάτων και κανόνες συσχέτισης ,ομοιότητα μεταξύ των εγγράφων, κατηγοριοποίηση και ταξινόμηση κειμένου. Το κάθε βήμα θα συνοδεύεται από διάφορα screen shots και γραφήματα για να γίνεται καλύτερα κατανοητό τι ακριβώς γίνεται και στην πράξη.

Τέλος, γίνεται μια περιγραφή ερευνών/περαμάτων που έχουν διεξαχθεί στο παρελθόν, στα οποία η ανάλυση συναισθήματος έχει δώσει σωστές προβλέψεις αρκετό καιρό πριν αυτά τα γεγονότα συμβούν στην πραγματικότητα. Μερικά απ' αυτά είναι η εύρεση του νικητή του Moudial 2010,ανάλυση για χρηματιστηριακούς δείκτες, εύρεση του νικητή στα αποτελέσματα των εκλογών του 2011 της Ιρλανδίας και εύρεση των επιτυχιών του Box Office από το Twitter.

2 Ανάλυση Συναισθημάτων

2.1 Τι είναι το Συναίσθημα;

Από επιστημονικής άποψης θα μπορούσαμε να ορίσουμε το συναίσθημα σαν την ψυχική κατάσταση του κάθε ατόμου, μια συγκεκριμένη στιγμή που έχει σχέση με την αντιμετώπιση της ζωής, των συνανθρώπων του και αυτών που συμβαίνουν γύρω του. Συναίσθημα είναι η χαρά, η λύπη, ο πόνος, η ευχαρίστηση και η ικανοποίηση, η θλίψη, ο ενθουσιασμός, ο θυμός και άλλα. Το συναίσθημα μερικές φορές είναι τόσο δυνατό που συνοδεύεται με οργανική λειτουργία, τέτοια, που γίνεται φανερό προς τα έξω. Μια έντονη συγκίνηση μπορεί να επιφέρει ένταση στη λειτουργία της καρδιάς, ή τρίξιμο των δοντιών, δάκρυα, πόνο στο στομάχι κ.α.

Από πού, όμως, παίρνουμε ερεθίσματα για να νοιώσουμε έτσι; Ο χαρακτήρας του κάθε ανθρώπου είναι διαφορετικός, αυτό σημαίνει ότι διαφορετικά πράγματα ευχαριστούν ή δυσαρεστούν διάφορους ανθρώπους. Για παράδειγμα, η ποίηση μπορεί να κάνει κάποιους να συγκινηθούν, να ερωτευτούν, να αρχίζουν να ελπίζουν ενώ κάποιους άλλους να βαρεθούν. Παρόλα αυτά, καθημερινά δεχόμαστε ερεθίσματα, απ' το εξωτερικό μας περιβάλλον, που μας προκαλούν διάφορα συναισθήματα θετικά και αρνητικά. Στα θετικά μπορούμε να συμπεριλάβουμε το ουράνιο τόξο μετά τη βροχή, την αναγνώριση της δουλειάς μας απ' το αφεντικό μας, όταν ακούμε ένα τραγούδι που μας αρέσει, τη στιγμή που, μετά από μια κουραστική μέρα, θα συναντήσουμε τους φίλους μας ή ανθρώπους που μας αγαπάνε και μας νοιάζονται ή ακόμα τη στιγμή που η εθνική μας ομάδα κέρδισε το Euro του 2004. Στα αρνητικά συναισθήματα μπορούμε να συμπεριλάβουμε, τον τσακωμό μας με κάποιον φίλο, ή τον αποχωρισμό από κάποιο αγαπημένο πρόσωπο, την ξαφνική βροχή που θα μας πιάσει ενώ είμαστε στο δρόμο, την ήττα της αγαπημένης μας ομάδας ή την αδικία που νοιώθουμε όταν δεν αναγνωρίζεται η δουλειά μας.

Όπως είπαμε, όμως, πιο πριν το συναίσθημα είναι καθαρά υποκειμενική αξία και δεν μπορούμε να ξέρουμε με σιγουριά τι σκέφτεται ή καλύτερα τι νοιώθει κάποιος άλλος όσο καλά κι αν τον γνωρίζουμε. Είναι γενική παραδοχή ότι όλοι έχουμε πιάσει τον εαυτό μας να αναρωτιέται 'Τι σκέφτεται αυτός;' αν όχι σε αυτή τη μορφή τότε ίσως

‘Του αρέσει η δουλειά μου’ ή για μας τους φοιτητές (αρκετά συχνά) ‘Θα με περάσει ο καθηγητής;’. Κατά καιρούς έχουν γίνει προσπάθειες για να απαντηθούν τέτοιου είδους ερωτήσεις, ένα απλό παράδειγμα είναι το τεστ αλήθειας που χρησιμοποιούν στην αστυνομία για να εξακριβώσουν αν ο ύποπτος είναι ένοχος ή όχι. Ο ύποπτος απαντάει τις ερωτήσεις που του κάνουν οι αστυνομικοί ενώ ταυτόχρονα καταγράφονται ενδείξεις για την πίεση του, τους παλμούς της καρδιάς κ.α. και αν υπάρχουν μεγάλες αποκλίσεις απ’ τα φυσιολογικά σημαίνει ότι ψεύδεται. Τον τελευταίο καιρό όμως έχει αρχίσει να αναπτύσσεται ένας νέος κλάδος της πληροφορικής, σε συνεργασία με την γλωσσολογία και την ψυχολογία ο οποίος ονομάζεται *ανάλυση συναισθήματος* και σκοπό έχει την κατηγοριοποίηση κειμένου που έχει γραφτεί από τους χρήστες, σε θετικό ή αρνητικό.

2.2 Ιστορία

Η ιδέα για την ανάλυση συναισθήματος υπάρχει από τότε που υπάρχει και ο άνθρωπος και είχε τη μορφή του «τι σκέφτονται οι άλλοι». Σκεφτείτε μόνο πόσες φορές έχετε ζητήσει από γνωστούς σας να σας συστήσουν έναν καλό υδραυλικό, πόσες φορές συζητήσατε ποιους σκοπεύετε να ψηφίσετε στις επερχόμενες εκλογές, ή ακόμα πόσες φορές αναζητήσατε στο internet την γνώμη άλλων για έναν ταξιδιωτικό προορισμό ή διαβάσατε κριτικές μιας ταινίας ή ενός προϊόντος.

Παρόλα αυτά, η ανάπτυξη του sentiment analysis σαν επιστήμη έγινε πολύ αργότερα, με την ανάπτυξη του Web 2.0. Με την άνθιση των blogs και του social network, οι χρήστες άρχισαν να καταγράφουν τις απόψεις τους για διάφορα θέματα, όπως προϊόντα, υπηρεσίες, εμπειρίες, κι έτσι γεννήθηκε, η ανάγκη στις εταιρίες να βρουν τρόπους να διαφημίσουν τα προϊόντα τους και να διαχειριστούν τη φήμη τους. Αυτό οδήγησε στην βαθύτερη εξέταση των καταγεγραμμένων απόψεων των χρηστών, έγιναν προσπάθειες για την κατανόηση των συζητήσεων, την εύρεση του περιεχομένου που σχετίζεται με το αντικείμενο το οποίο πραγματεύονται και τη δημιουργία μιας αυτοματοποιημένης διαδικασίας φιλτραρίσματος του θορύβου, με λίγα λόγια την αρχή του κλάδου της sentiment analysis.

Σημαντικό ρόλο σ’ αυτό, έπαιξε και η άνθιση μεθόδων μηχανικής μάθησης στη φυσική γλώσσα επεξεργασίας (NLP) και στην αναζήτηση πληροφοριών όπως επίσης η πληθώρα ομάδων δεδομένων που υπάρχουν πια στο διαδίκτυο.

2.3 Βασικές Αρχές

Για να κατανοήσουμε την ανάλυση συναισθήματος θα ξεκινήσουμε αναλύοντας τους τρεις βασικούς κλάδους πάνω στους οποίους είναι «χτισμένη», την *επεξεργασία της φυσικής γλώσσας*, την *λανθάνουσα σημασιολογική ανάλυση*, το *text mining* και την *υπολογιστική γλωσσολογία*. Εξετάζοντας τους θα μπορέσουμε να καταλάβουμε πως λειτουργεί η ανάλυση συναισθήματος.

2.3.1 Επεξεργασία Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language processing, NLP) είναι ένας κλάδος της επιστήμης υπολογιστών και της γλωσσολογίας η οποία πραγματεύεται την αλληλεπίδραση της γλώσσας υπολογιστών και της ανθρώπινης. Σκοπός αυτού του κλάδου είναι ο σχεδιασμός και υλοποίηση υπολογιστικών μοντέλων της φυσικής γλώσσας, τα οποία θα έχουν την ικανότητα αναγνώρισης ή κατανόησης της, καθώς και παραγωγής φυσικής γλώσσας ή σύνθεση. Μερικές εφαρμογές μπορεί να είναι ο διάλογος με τον υπολογιστή, η μηχανική μετάφραση (από γλώσσα σε γλώσσα, από γλώσσα σε βάση δεδομένων και αντιστρόφως), το browsing ή filtering από κάποιον agent κτλ. Δεν είναι, λοιπόν, δύσκολο να καταλάβουμε πόσο σημαντικό και βασικό ρόλο έχει παίξει η ανάπτυξη αυτού του κλάδου στην ανάλυση συναισθημάτων

Για να κατανοήσουμε καλύτερα την NLP θα πρέπει να αναλύσουμε τα επίπεδα της φυσικής γλώσσας:

1. Φωνολογικό
2. Μορφολογικό
3. Λεξικολογικό
4. Συντακτικό
5. Σημασιολογικό
6. Συζήτησης
7. Πραγματολογικό

Φωνολογικό Επίπεδο

Το *φωνολογικό επίπεδο* έχει να κάνει με την ερμηνεία του λόγου μέσα από τις λέξεις. Ποιο συγκεκριμένα για το πώς ακούγονται οι λέξεις. Υπάρχουν τρεις κανόνες στη φωνητική ανάλυση: Ο φωνητικός κανόνας που έχει να κάνει με το πώς ακούγεται ο ήχος μέσα από τις λέξεις, ο κανόνας των φωνημάτων που έχει να κάνει με τις παραλλαγές της προφοράς όταν χρησιμοποιούμε τις λέξεις σαν μία πρόταση και τέλος ο προσωδιακός κανόνας που έχει να κάνει με τη διακύμανση και τον τόνο της φωνής σε μία πρόταση. Για παράδειγμα, μια πρόταση μπορεί να έχει πολλές ερμηνείες

ανάλογα με τον τρόπο που θα την εκφέρουμε. Την πρόταση «Θα βρέξει σήμερα» μπορούμε να την εκφέρουμε σαν μια απλή δήλωση, με θαυμασμό, απογοήτευση, εκνευρισμό, με προσμονή ή σαν μια ερώτηση. Σε ένα σύστημα NLP που δέχεται δεδομένα ομιλίας στην είσοδο, τα κύματα ήχου αναλύονται και κωδικοποιούνται σε ένα ψηφιακό σήμα προκειμένου να ερμηνευτούν μέσω των κανόνων ή μέσω σύγκρισης απ' το συγκεκριμένο μοντέλο γλώσσας που χρησιμοποιείται.

Μορφολογικό Επίπεδο

Το *μορφολογικό επίπεδο* έχει να κάνει με τη σύσταση των λέξεων, δηλαδή με τις μονάδες από τις οποίες συντίθενται. Για παράδειγμα, η λέξη «προκαταβάλω» αποτελείται από το πρόθεμα προ- και τις λέξεις –κατά- και –βάλω. Από τη στιγμή που το νόημα των λέξεων δεν αλλάζει μπορούμε να διασπάσουμε οποιαδήποτε λέξη σε συνθετικά για να βρούμε το νόημά της. Όμοια, ένα σύστημα NLP αναγνωρίζει το νόημα κάθε συστατικού για να βρει τι σημαίνει η λέξη η οποία ζητείται.

Λεξικολογικό Επίπεδο

Στο *λεξικολογικό επίπεδο*, τα NLP συστήματα ερμηνεύουν τις λέξεις μεμονωμένα. Λέξεις που έχουν μία σημασία μπορούν να αντικατασταθούν από μια σημασιολογική αναπαράσταση της έννοιάς τους, η φύση της οποίας μπορεί να ποικίλει σύμφωνα με τη σημασιολογική θεωρία που χρησιμοποιείται από το NLP σύστημα. Ακολουθεί η σημασιολογική αναπαράσταση της λέξης πλοίο:

Πλοίο = Ένα μεγάλο, πλωτό, κοίλο σκάφος το οποίο χρησιμοποιείται για να μεταφέρει ανθρώπους σε ποταμούς, λίμνες ή λιμάνια.

Class: σκάφος

Properties: μεγάλο, πλωτό, κοίλο

Purpose: predication: class: μεταφέρει objects: ανθρώπους

Όπως είναι κατανοητό στο λεξικολογικό επίπεδο απαιτούνται λεξικά απλά ή σύνθετα που να βοηθούν τα συστήματα στη σημασιολογική αναπαράσταση.

Συντακτικό Επίπεδο

Το *συντακτικό επίπεδο* αναλύει τις λέξεις μιας πρότασης και βρίσκει τη γραμματική της δομή. Το αποτέλεσμα της επεξεργασίας είναι μια αναπαράσταση της πρότασης η οποία μας δείχνει τη σχέση εξάρτησης της δομής ανάμεσα στις λέξεις.

Σημασιολογικό Επίπεδο

Στη *σημασιολογική επεξεργασία* καθορίζεται το πιθανό νόημα μιας πρότασης εστιάζοντας στις αλληλεπιδράσεις ανάμεσα στο νόημα των λέξεων που έχει βρεθεί απ' τα προηγούμενα επίπεδα. Ακόμα περιλαμβάνει την ερμηνεία των λέξεων με πολλές σημασίες. Η σημασιολογική προσέγγιση επιτρέπει να γίνεται επιλογή μόνο μίας σημασίας αυτών των λέξεων για να συμπεριληφθεί στη σημασιολογική αναπαράσταση της πρότασης. Για παράδειγμα, η λέξη δίσκος άλλες φορές χρησιμοποιείται σαν το γνωστό μας βινύλιο και άλλες για το αντικείμενο που χρησιμοποιούμε για να σερβίρουμε. Αν απαιτούνται πληροφορίες για την ερμηνεία , το σημασιολογικό επίπεδο θα κάνει την ερμηνεία αυτή με τη βοήθεια μιας μεγάλης ποικιλίας μεθόδων .

Επίπεδο Συζήτησης

Το *επίπεδο συζήτησης* δουλεύει με μονάδες κειμένου μεγαλύτερα μιας πρότασης. Εστιάζει στα περιεχόμενα του κειμένου σαν μια ολότητα που εκφράζει το νόημα με το να δημιουργεί συνδέσεις ανάμεσα στα συστατικά των προτάσεων. Δύο τύποι της επεξεργασίας αυτού του επιπέδου είναι οι anaphora resolution και discourse/text structure recognition. Σύμφωνα με τον πρώτο τύπο γίνεται αντικατάσταση λέξεων όπως αντωνυμίες, οι οποίες δεν έχουν σημασιολογική αξία, με τις κατάλληλες οντότητες στις οποίες αναφέρονται. Ο δεύτερος τύπος αποφασίζει για τις λειτουργίες των προτάσεων του κειμένου τις οποίες και προσθέτει στην νοηματική παράστασή του.

Πραγματολογικό Επίπεδο

Το *πραγματολογικό επίπεδο* μπορούμε να πούμε ότι κατά κάποιο τρόπο συμπληρώνει το σημασιολογικό. Υπάρχουν κάποιες γλωσσικές σημασίες που δεν έχει κατορθώσει να εκφράσει η σημασιολογία. Για παράδειγμα, οι ερωτήσεις, οι διαταγές, οι παρακλήσεις, οι ευχές δεν έχουν ακόμη βρει τα σημασιολογικά τους αντίστοιχα για συστήματα που στηρίζονται στη συνθετικότητα και τις τιμές αλήθειας. Γι' αυτές τις περιπτώσεις υπάρχει το πραγματολογικό επίπεδο.

2.3.2 Εξαγωγή Πληροφορίας από Κείμενο

Το text mining (εξαγωγή πληροφορίας από κείμενο) [5], γνωστό και ως data mining, είναι μια διαδικασία εξαγωγής καινούριας μη δομημένης πληροφορίας από μια συλλογή κειμένων. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για να δημιουργηθεί μια περίληψη των λέξεων του εγγράφου, για την εξαγωγή προτύπων (patterns) και την εκτίμηση και διερμηνεία του αποτελέσματος. Όπως ορίζουν οι

Καρανίκας και Θεοδουλίδης [1] «Το text mining είναι ένα βήμα στη διαδικασία του KDT (Knowledge Discovery in Text) που αποτελείται από ιδιαίτερους αλγορίθμους του data mining και του natural language processing που κάτω από μερικούς αποδεκτούς υπολογιστικούς περιορισμούς αποδοτικότητας παράγουν έναν αριθμό από patterns μέσα από ένα σύνολο μη δομημένων δεδομένων κειμένου».

Στόχος, λοιπόν, του text mining είναι να ανακαλυφθούν οι μέχρι τώρα άγνωστες ή κρυμμένες πληροφορίες, τις οποίες κανένας ακόμα δεν γνωρίζει και έτσι ακόμα δεν έχει γράψει γι' αυτό.

Πρακτικά, το text mining, μπορεί να παρουσιαστεί σαν μια διαδικασία «νουμεροποίησης» του κειμένου. Όλες οι λέξεις που υπάρχουν στο προς εξέταση έγγραφο δημιουργούν έναν πίνακα με συχνότητες ο οποίος μετράει πόσες φορές εμφανίζεται κάθε λέξη. Μ' αυτή τη μέθοδο αποκλείονται λέξεις όπως είναι τα άρθρα και υπολογίζονται αυτές που έχουν την ίδια ρίζα όπως οι λέξεις υπολογίζω, υπολογιστής, υπολογισμός. Όταν, λοιπόν, δημιουργηθεί αυτός ο πίνακας μπορούμε να εφαρμόσουμε διάφορες στατιστικές μετρήσεις ή data mining τεχνικές ανάλογα με τη μορφή του αποτελέσματος που θέλουμε να βγει.

Βέβαια, η «νουμεροποίηση» του κειμένου μειονεκτεί σε ορισμένα σημεία. Για παράδειγμα, ενώ αυτή η μέθοδος λειτουργεί πάρα πολύ καλά σε πολλά αλλά μικρά έγγραφα δεν μπορούμε να πούμε το ίδιο για τα λίγα αλλά μεγάλα έγγραφα. Επίσης, πρόβλημα μπορεί να εμφανιστεί όταν κάποιες λέξεις εννοούν κάτι άλλο απ' αυτό που πραγματικά σημαίνουν, όπως η λέξη window που σημαίνει παράθυρο, όταν όμως χρησιμοποιείται σαν Microsoft Windows η σημασία της αλλάζει και έτσι τα αποτελέσματα του πίνακα βγαίνουν διαφορετικά.

Παρόλα αυτά όμως, το text mining ήταν και εξακολουθεί να είναι ένα πολύ χρήσιμο εργαλείο τεχνητής νοημοσύνης το οποίο εξελίσσεται διαρκώς με αποτέλεσμα αυτά που μας προσφέρει να υπερτερούν των μειονεκτημάτων του.

2.3.3 Υπολογιστική Γλωσσολογία

Ένα πεδίο το οποίο πολλές φορές συγχέεται με την Επεξεργασία Φυσικής Γλώσσας είναι η Υπολογιστική Γλωσσολογία (Computational Linguistics) η οποία συνδυάζει τη γλωσσολογία με την επιστήμη υπολογιστών, την ψυχολογία και την λογική. Ασχολείται με τη στατιστική και λογική μοντελοποίηση της φυσικής γλώσσας με μία υπολογιστική οπτική. Με λίγα λόγια, δημιουργούνται προγράμματα τα οποία διδάσκουν στους

ηλεκτρονικούς υπολογιστές πώς να επικοινωνούν με τους ανθρώπους. Είναι ένα παρακλάδι της τεχνητής νοημοσύνης το οποίο ερευνά προβλήματα όπως η αναγνώριση λόγου, η κατανόηση της φυσικής γλώσσας, η παραγωγή φυσικής γλώσσας, η σύνθεση λόγου, η εύρεση και εξαγωγή πληροφορίας, αλλά και η διόρθωση ορθογραφίας, ο έλεγχος γραμματικής και η αυτόματη μετάφραση.

Η Υπολογιστική Γλωσσολογία χρησιμοποιεί προγράμματα επεξεργασίας του λόγου για να παράγει αποτελέσματα. Καθοριστικός όμως είναι και ο ρόλος της ψυχολογικής κατάστασης του ομιλητή. Λόγω της πολυπλοκότητάς της, αυτή η διαδικασία θα πρέπει να χωριστεί σε ορισμένα στάδια. Σε κάθε στάδιο δίνονται στη γλώσσα κάποια τυπικά χαρακτηριστικά και δομές. Για να γίνει αυτό πραγματικότητα, χρησιμοποιούνται τα μοντέλα της υπολογιστικής γλωσσολογίας, τα οποία είναι επίσης υπεύθυνα για το πώς αυτά τα χαρακτηριστικά πρέπει να αλληλεπιδρούν σε κάθε στάδιο της γλωσσικής μετατροπής του υπολογιστή. Κάποια απ' αυτά τα μοντέλα είναι τα νευρογλωσσολογικά μοντέλα (neurolinguistic models), τα ψυχογλωσσικά μοντέλα (psycholinguistic models), τα λειτουργικά μοντέλα (functional models) και τα μοντέλα αναζήτησης (research models).

Νευρογλωσσολογικά Μοντέλα

Τα *νευρογλωσσολογικά μοντέλα* προσπαθούν να ανακαλύψουν τις σχέσεις μεταξύ της εξωτερικής δραστηριότητας της ομιλίας του ανθρώπου και της αντίστοιχης χυμικής δραστηριότητας των νεύρων του εγκεφάλου. Αυτή όμως η τεχνική βρίσκεται ακόμα σε πολύ αρχικά στάδια και ως εκ τούτου, εάν δεν γίνουν επαναστατικές αλλαγές, αυτού του τύπου τα μοντέλα δεν δύναται να δώσουν αξιόπιστα αποτελέσματα στο κοντινό μέλλον.

Ψυχογλωσσικά Μοντέλα

Τα *ψυχογλωσσικά μοντέλα* ερευνούν την ομιλία του ανθρώπου, την αντίληψη και τις μορφές του λόγου μέσω ψυχολογικών μεθόδων.

Λειτουργικά Μοντέλα

Θεωρώντας ότι έχουμε μια μηχανή ηχογράφησης, μπορούμε να της κάνουμε μερικές ερωτήσεις και να ηχογραφήσουμε τις απαντήσεις. Έτσι μπορούμε να πούμε ότι έχουμε δύο διαδικασίες την αναλυτική και τη συνθετική. Στην αναλυτική διαδικασία επεξεργάζεται ο λόγος ενώ στη συνθετική παράγονται οι αντιδράσεις. Τα λειτουργικά μοντέλα χρησιμοποιούν τους κανόνες της συζήτησης των εισερχόμενων γλωσσικών πληροφοριών στις πληροφορίες εξόδου χωρίς να γίνεται καμία προσπάθεια

αναπαραγωγής των εσωτερικών μηχανισμών της δραστηριότητας του εγκεφάλου. Κατά τη διάρκεια αυτής της διαδικασίας δεν αναζητούνται ανθρωπομορφικά χαρακτηριστικά και δεν χρησιμοποιείται προσομοίωση για την εξέταση της εγκεφαλικής αντίδρασης. Παρόλα αυτά, τα αποτελέσματα είναι όσο το δυνατόν πιο κοντά με αυτά του ανθρώπινου εγκεφάλου. Μέχρι στιγμής αυτά τα μοντέλα έχουν αποδειχθεί τα πιο επιτυχημένα ίσως επειδή βασίζονται σε αληθινά δεδομένα, είναι εύκολα προσπελάσιμα και έχουν ατελείωτο απόθεμα σε κείμενα και ηχογραφημένη ομιλία.

Μοντέλα Αναζήτησης

Αυτού του είδους τα μοντέλα παίρνουν σαν είσοδο κείμενα στη φυσική γλώσσα και σαν έξοδο παράγουν άλλα κείμενα συνήθως αυστηρώς διαμορφωμένα τα οποία αναπαριστούν περιεχόμενα λεξικών, γραμματικών πινάκων, ή οτιδήποτε μοιάζει με τμήμα λειτουργικού μοντέλου. Όπως μπορούμε να καταλάβουμε, τα μοντέλα αναζήτησης είναι εργαλεία κατασκευής λειτουργικών μοντέλων.

2.4 Λανθάνουσα Σημασιολογική Ανάλυση

Η Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis γνωστή και σαν Latent Semantic Indexing (LSI)) είναι μια θεωρία και μέθοδος για την εξαγωγή και αναπαράσταση του νοήματος των λέξεων με στατιστικούς υπολογισμούς οι οποίοι εφαρμόζονται σε ένα μεγάλο μέρος κειμένου (Landauer and Dumais, 1997). Πρόκειται για μια μαθηματική/στατιστική τεχνική ανάκτησης πληροφορίας η οποία λαμβάνει υπόψη της αλληλεξαρτήσεις μεταξύ των όρων. Στην ουσία ψάχνει να βρει το νόημα των υπό εξέταση εγγράφων εξετάζοντας τις λέξεις που περιέχουν. Φυσικά λέξεις όπως άρθρα, σύνδεσμοι και ρήματα απορρίπτονται. Κάθε λέξη όμως μπορεί να έχει παραπάνω από μία σημασίες, γι’ αυτό η LSA κοιτάζει το νόημα όλων των σημαντικών λέξεων προσπαθώντας να τις συνδυάσει.

Αυτή η τεχνική συμπεριφέρεται στα έγγραφα σαν μια συλλογή από λέξεις, όπου η σειρά δεν έχει καμία σημασία, παρά μόνο το πόσες φορές εμφανίζεται κάθε λέξη στο έγγραφο. Οι έννοιες αναπαρίστανται σαν πρότυπα λέξεων τα οποία εμφανίζονται μαζί στα έγγραφα. Για παράδειγμα, οι λέξεις λουρί, κέρασμα και υπάκουος συνήθως υπάρχουν σε έγγραφα που έχουν να κάνουν με εκπαίδευση σκύλου. Τέλος αυτή η τεχνική υποθέτει ότι οι λέξεις έχουν μόνο μία σημασία.

Πρακτικά, η LSA δημιουργεί έναν πίνακα με όρους και έγγραφα. Σ’ αυτόν τον πίνακα υπολογίζονται ποιες σημαντικές λέξεις υπάρχουν σε κάθε έγγραφο και πόσες

φορές. Σημαντικές θεωρούνται οι λέξεις οι οποίες εμφανίζονται σε πάνω από δύο έγγραφα και δεν είναι άρθρα, σύνδεσμοι και ρήματα όπως είπαμε παραπάνω. Για παράδειγμα, για τα παρακάτω έγγραφα θα δημιουργηθεί ο πίνακας Μ:

1. The Neatest Little Guide to Stock Market Investing
2. Investing For Dummies, 4th Edition
3. The Little Book of Common Sense Investing: The Only Way to Guarantee Your Fair Share of Stock Market Returns
4. The Little Book of Value Investing
5. Value Investing: From Graham to Buffett and Beyond
6. Rich Dad's Guide to Investing: What the Rich Invest in, That the Poor and the Middle Class Do Not!
7. Investing in Real Estate, 5th Edition
8. Stock Investing For Dummies
9. Rich Dad's Advisors: The ABC's of Real Estate Investing: The Secrets of Finding Hidden Profits Most Investors Miss

Πίνακας 1 Οι σημαντικές λέξεις στον πίνακα Μ

Σημαντικές Λέξεις	Έγγραφα								
	E1	E2	E3	E4	E5	E6	E7	E8	E9
Book			1	1					
Dad’s						1			1
Dummies		1						1	
Estate							1		1
Guide	1					1			
Investing	1	1	1	1	1	1	1	1	1
Market	1		1						
Real							1		1
Rich							2		1
Stock	1		1					1	
Value				1	1				

Όταν δημιουργηθεί ο πίνακας χρησιμοποιείται μια τεχνική η οποία λέγεται Singular Value Decomposition (SVD), για να αναλύσει τα δεδομένα. Σύμφωνα μ' αυτή την τεχνική δημιουργούνται τρεις νέοι πίνακες οι U, S και V^T. Ο πίνακας U περιέχει τα ιδιοδιανύσματα[1] του πίνακα $M \cdot M^T$ (έστω M ένας τετραγωνικός πίνακας διαστάσεων n*n. Η τιμή λ καλείται ιδιοτιμή (eigenvector) του M εάν υπάρχει ένα μη μηδενικό διάνυσμα u έτσι ώστε να ισχύει η σχέση: M*u=λ*u). Στην περίπτωση αυτή το διάνυσμα u καλείται ιδιοδιάνυσμα. Στην ουσία περιγράφει τη μετάβαση απ' τους όρους (terms) στις έννοιες (concepts). Ο πίνακας S περιγράφει τη συνεισφορά της κάθε έννοιας (με χρήση των ιδιαζουσών τιμών), και ο πίνακας V^T τη μετάβαση από τις έννοιες στα έγγραφα. Το γινόμενο αυτών των τριών πινάκων μας δίνει τον αρχικό πίνακα M. Εδώ δεν θα αναλυθεί περαιτέρω η SVD παρόλα αυτά δίνονται παρακάτω οι τρεις πίνακες του παραδείγματός μας:

Πίνακας 2: Πίνακας U

Book	0.15	-0.27	0.04
Dad's	0.24	0.38	-0.09
Dummies	0.13	-0.17	0.07
Estate	0.18	0.19	0.45
Guide	0.22	-0.09	0.46
Investing	0.74	-0.21	0.21
Market	0.18	-0.30	-0.28
Real	0.18	0.19	0.45
Rich	0.36	0.59	-0.34
Stock	0.25	-0.42	-0.28
Value	0.12	-0.14	0.23

Πίνακας 3: Πίνακας S

3.91	0	0
0	2.61	0
0	0	2.00

Πίνακας 4: Πίνακας V^T

E1	E2	E3	E4	E5	E6	E7	E8	E9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0.00	0.34

Υπάρχουν βέβαια ορισμένοι περιορισμοί που μπορούν να θεωρηθούν και μειονεκτήματα της συγκεκριμένης μεθόδου. Καταρχάς, η LSA δεν μπορεί να χειριστεί καλά τις λέξεις με πολλές έννοιες. Θεωρεί ότι κάθε λέξη έχει την ίδια σημασία όπου κι αν βρίσκεται. Έπειτα, εξαρτάται πάρα πολύ απ' τη μέθοδο SVD η οποία βασίζεται σε υπολογισμούς και ενημερώνεται δύσκολα όταν εμφανίζονται καινούρια έγγραφα. Παρόλα αυτά, αρχίζουν να βρίσκονται λύσεις για τα προβλήματα καθώς γίνονται συνεχώς έρευνες για τη βελτίωση της.

2.5 Τι είναι η Ανάλυση Συναισθήματος

Η *ανάλυση συναισθήματος (Sentiment Analysis)* ή αλλιώς *opinion mining* προσδιορίζει τις απόψεις που υπάρχουν σε ένα κείμενο. Με απλά λόγια, προσπαθεί να διακρίνει τα συναισθήματα του συγγραφέα και τα κατηγοριοποιεί. Για την κατηγοριοποίηση αυτή χρησιμοποιούνται δύο στάδια τα οποία αποτελούνε και τις δυο κύριες περιοχές έρευνας της Sentiment Analysis.

Το πρώτο στάδιο είναι η αντικειμενική/υποκειμενική ταυτοποίηση (*subjectivity/objectivity identification*) κατά την οποία κατατάσσουμε το κείμενο, που συνήθως είναι μια πρόταση σε αντικειμενικό ή υποκειμενικό. Ωστόσο, αυτό το στάδιο μπορεί να αποδειχθεί πιο δύσκολο απ' την πολικότητα κειμένου καθώς ένα υποκειμενικό έγγραφο μπορεί να περιέχει αντικειμενικές προτάσεις. Παρόλα αυτά, έχει

αποδειχθεί ότι αν αφαιρέσουμε τις αντικειμενικές προτάσεις από ένα έγγραφο μπορούμε να έχουμε καλύτερα αποτελέσματα.

Στη συνέχεια ελέγχουμε την πολικότητα κειμένου, δηλαδή την κατάταξη των συμπερασμάτων σε θετικά (positive), αρνητικά (negative) και ουδέτερα (neutral). Στην ουσία παίρνει το αντικείμενο το οποίο προέκυψε προηγουμένως, ένα κομμάτι κειμένου ή μια πρόταση, εξετάζει τις σημαντικές λέξεις μία μία και προσπαθεί να «καταλάβει» αν το συμπέρασμα είναι θετικό αρνητικό ή ουδέτερο. Μ' αυτόν τον τρόπο λοιπόν θα μπορούσαμε να δούμε αν μια ταινία είναι καλή ή κακή, αν αξίζει να πάμε σε ένα εστιατόριο, αν αξίζει να κάνουμε την επένδυση που έχουμε στο μυαλό μας, πως αντιδρά το κοινό σε ένα καινούριο προϊόν και άλλα πολλά.

Συνήθως η ανάλυση συναισθήματος πραγματεύεται μόνο θετικά ή αρνητικά συναισθήματα και όχι μεικτά όπως για παράδειγμα η έκπληξη. Δεν εντοπίζει την ένταση του συναισθήματος αν και μερικές φορές χρησιμοποιεί την ένταση των σχέσεων ανάμεσα στις λέξεις για να αποδώσει την κατάσταση (αν είναι θετικό ή αρνητικό)(Kaji & Kitsuregawa, 2007). Επίσης δεν αναγνωρίζει ταυτόχρονα και αρνητικά και θετικά συναισθήματα.

Πριν όμως γνωρίσουμε καλύτερα την ανάλυση συναισθημάτων θα πρέπει να αναλύσουμε μια έννοια η οποία έχει δώσει πολλά στον κλάδο και που χωρίς τη βοήθειά της η ανάλυση συναισθημάτων δεν θα είχε τα ίδια αποτελέσματα, τη μηχανική μάθηση.

2.6 Μηχανική Μάθηση

Όπως έχουμε αναφέρει ένας από τους λόγους ανάπτυξης της ανάλυσης συναισθημάτων ήταν η σημαντική πρόοδος που γνώρισε ο κλάδος της μηχανικής μάθησης με τη σταδιακή δημιουργία και βελτίωση νέων μεθόδων. Η διαδικασία κατά την οποία οι υπολογιστές εξελίσσουν την συμπεριφορά τους μέσω εμπειρικών δεδομένων, με τη βοήθεια ειδικών αλγορίθμων και μεθόδων, λέγεται *μηχανική μάθηση* (machine learning). Υπάρχουν δύο είδη μηχανικής μάθησης, *μάθηση με επίβλεψη* και *μάθηση χωρίς επίβλεψη*.

2.6.1 Μάθηση με Επίβλεψη

Σύμφωνα με αυτή τη μέθοδο ο αλγόριθμος δημιουργεί μια συνάρτηση που απεικονίζει γνωστές εισόδους για επιθυμητές εξόδους, έτσι λοιπόν, το σύστημα καλείται να μάθει κάθε βήμα μέχρι να φτάσει στην επιθυμητή έξοδο.

Υπάρχουν διάφοροι αλγόριθμοι και μέθοδοι τους οποίους χρησιμοποιεί η μηχανική μάθηση με επίβλεψη, ο αλγόριθμος απαλοιφής υποψήφιων, τα δέντρα ταξινόμησης/ απόφασης και ο αλγόριθμος ID3.

2.6.2 Μάθηση χωρίς Επίβλεψη

Κατά τη μάθηση χωρίς επίβλεψη, ο αλγόριθμος κατασκευάζει ένα μοντέλο με κάποιες εισόδους χωρίς να γνωρίζει τις επιθυμητές εξόδους, έτσι το σύστημα πρέπει να «μάθει» μόνο του τις συσχετίσεις και τα βήματα για να φτάσει στο επιθυμητό αποτέλεσμα. Για το σκοπό αυτό δημιουργεί κάποια πρότυπα, μερικά από τα οποία είναι οι κανόνες συσχετίσεις κα οι ομάδες οι οποίες προκύπτουν απ' τη διαδικασία ομαδοποίησης.

Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης έχουν να κάνουν με τις σχέσεις μεταξύ των αντικειμένων. Ένα απλό παράδειγμα είναι και ο λόγος για τον οποίο δημιουργήθηκαν, δηλαδή, ως τεχνική ανάλυσης του καλαθιού αγορών. Εδώ ψάχνουμε να βρούμε ποια αντικείμενα αγοράζονται συνήθως μαζί. Για παράδειγμα, όποιος αγοράσει ζαμπόν και τυρί, συνήθως θα αγοράσει και ψωμί.

Ένας αλγόριθμος που χρησιμοποιούν οι κανόνες συσχέτισης είναι ο Αλγόριθμος Apriori ο οποίος βασίζεται στην ιδιότητα της μονοτονίας (Αν ένα σύνολο αντικειμένων S είναι συχνό, τότε όλα τα υποσύνολα του S είναι επίσης συχνά.). Σύμφωνα μ' αυτόν δημιουργείται ένα σύνολο αντικειμένων από το οποίο, στη συνέχεια, θα βρεθούν, με τη βοήθεια του ορίου υποστήριξης (η πιθανότητα να βρεθεί το συγκεκριμένο σύνολο αντικειμένων στη βάση δεδομένων), τα σύνολα συχνών αντικειμένων ή τα μέγιστα συχνά σύνολα αντικειμένων.

Τέλος, για τη δημιουργία κανόνων συσχέτισης χρησιμοποιείται η εμπιστοσύνη (η πιθανότητα να βρεθεί ένα αντικείμενο Α στο σύνολο αντικειμένων) και επιλέγουμε τους κανόνες με εμπιστοσύνη που ξεπερνά το όριο που έχει θέσει ο χρήστης.

Πρότυπο Ομάδες

Κατά το πρότυπο ομάδες(clusters) τα δεδομένα ομαδοποιούνται με τέτοιο τρόπο ώστε, δεδομένα ίδιας ομάδας να μοιάζουν όσο το δυνατόν περισσότερο και δεδομένα διαφορετικών ομάδων να διαφέρουν όσο το δυνατόν περισσότερο. Οι αλγόριθμοι

ομαδοποίησης διακρίνονται σε τρεις κατηγορίες, τους αλγόριθμους βασισμένους σε διαχωρισμούς, τους ιεραρχικούς αλγόριθμους και του πιθανοκρατικούς.

Οι αλγόριθμοι βασισμένοι σε διαχωρισμούς προσπαθούν να βρουν τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων. Σ' αυτή την κατηγορία ανήκει και ο αλγόριθμος των Κ-μέσων. Αρχικά επιλέγονται Κ τυχαία σημεία δεδομένων για κέντρα των ομάδων. Ο αριθμός Κ έχει καθοριστεί πριν την έναρξη του αλγορίθμου. Στη συνέχεια κάθε σημείο θα μπει στην ομάδα της οποίας το κέντρο είναι πιο κοντά του. Τέλος υπολογίζεται ο μέσος όρος όλων των σημείων κάθε ομάδας και τίθεται αυτό ως το νέο κέντρο της. Αυτά τα βήματα επαναλαμβάνονται είτε για ένα προκαθορισμένο αριθμό βημάτων είτε μέχρι να μην υπάρχει πλέον διαχωρισμός.

Οι ιεραρχικοί αλγόριθμοι προσπαθούν με ιεραρχικό τρόπο να ανακαλύψουν τον αριθμό και τη δομή των ομάδων. Συνδυάζουν τις ομάδες σε μεγαλύτερες ή διαιρούν τις μεγάλες ομάδες σε μικρότερες. Έτσι λοιπόν, προκύπτουν δύο ακόμα ομάδες αλγορίθμων, οι αλγόριθμοι συγχώνευσης και οι αλγόριθμοι διαίρεσης.

Τέλος, έχουμε τους πιθανοκρατικούς αλγόριθμους οι οποίοι βασίζονται σε μοντέλα πιθανοτήτων.

2.6.3 Η Μηχανική Μάθηση για την Ανάλυση Συναισθημάτων

Συχνά οι αλγόριθμοι της sentiment analysis χρησιμοποιούν μεθόδους μηχανικής μάθησης για να αναγνωρίσουν στοιχεία τα οποία συνδέονται με αρνητικά ή θετικά συναισθήματα. Αυτά τα στοιχεία μπορεί να είναι υποσύνολα λέξεων του εγγράφου, μέρη του λόγου ή φωνήματα, συλλαβές, γράμματα ακόμα και ολόκληρες λέξεις ή φράσεις (n-grams) και παρέχουν μια μικρή αλλά σημαντική αύξηση της απόδοσης. Δύο προβλήματα που πρέπει να λύσει η μηχανική μάθηση είναι η εύρεση αυτών των στοιχείων και η επιλογή του αλγορίθμου ταξινόμησης.

Κατά τη διαδικασία επιλογής στοιχείων επεξεργάζονται δεδομένα για να απομακρυνθούν τα λιγότερο χρήσιμα n-grams. Έχει αποδειχθεί ότι μ' αυτόν τον τρόπο έχουμε μια μικρή βελτίωση της απόδοσης κατά την ταξινόμηση. Μικρές βελτιώσεις μπορούν, επίσης, να πραγματοποιηθούν αν διαχωρίσουμε απ' το σύνολο των στοιχείων τα απλά στοιχεία, τα οποία έχουν τιμές κέρδους πληροφορίας αρκετά μεγάλες. Ένας αλγόριθμος που αποδίδει πολύ καλά σ' αυτές τις περιπτώσεις είναι η μέγιστη εντροπία.

Όσο αφορά την επιλογή αλγορίθμου ταξινόμησης, ο SVM χρησιμοποιείται ευρέως, επειδή αποδίδει το ίδιο ή πολλές φορές και καλύτερα από άλλες μεθόδους μηχανικής μάθησης.

Μια ακόμα μέθοδος εύρεσης συναισθήματος σε κείμενο είναι η δημιουργία λεξικού με αρνητικές και θετικές λέξεις και ο υπολογισμός του πόσο συχνά εμφανίζονται. Μπορούμε να δημιουργήσουμε μόνοι μας το λεξικό χρησιμοποιώντας όποιες λέξεις θέλουμε. Η συχνότητα εμφάνισης τους θα μας πει αν το κείμενο έχει αρνητικό συμπέρασμα ή θετικό.

Υπάρχουν αρκετές μέθοδοι μηχανικής μάθησης που χρησιμοποιεί η ανάλυση συναισθημάτων, πολλές απ' αυτές μοιάζουν μεταξύ τους. Παρακάτω αναλύουμε τρεις απ' αυτές τον αλγόριθμο ταξινόμηση Naive Bayes, την κατηγοριοποίηση μέγιστης εντροπίας και τις μηχανές υποστήριξης διανύσματος (Support Vector Machines, SVM) για τις οποίες οι Bo Pang, Lillian Lee και Shivakumar Vaithyanathan[3] σε πρόσφατη μελέτη τους απέδειξαν ότι δεν είναι τόσο αποτελεσματικές στην κατηγοριοποίηση συναισθημάτων όσο στην topic-based categorization.

Για να αναλύσουμε αυτές τις μεθόδους θα χρησιμοποιήσουμε το $\{f_1,f_2,\ldots f_m\}$ το οποίο είναι ένα σύνολο από m προκαθορισμένα στοιχεία τα οποία υπάρχουν στο κείμενο, το $n_i(d)$ το οποίο είναι ο αριθμός των εμφανίσεων του f_i στο κείμενο d και το διάνυσμα του κειμένου d , $\vec{}$ d:= $(n_1(d), n_2(d), \ldots, n_m(d))$.

Naive Bayes

Ο κανόνας του Bayes ορίζεται από τη σχέση $P(c \mid d)=P(c)P(d \mid c)/P(d)$. Όπου c είναι μια κλάση του d όπου ισχύει $c^*=\arg\max_c P(c \mid d)$. Σε περίπτωση που τα στοιχεία του κειμένου είναι ανεξάρτητα μεταξύ τους, ο τύπος αυτός αναλύεται περεταίρω και γίνεται

$$P_{NB}(c \mid d):= P(c)(\Pi^m_{i=1}P(fi \mid c)^{ni(d)})/ P(d)$$

Γενικά ο συγκεκριμένος αλγόριθμος χαρακτηρίζεται από απλότητα και τις υποθέσεις για ανεξαρτησία που κάνει και δίνει αρκετά καλά αποτελέσματα σε προβλήματα στα οποία τα στοιχεία είναι εξαρτημένα μεταξύ τους. Παρόλα αυτά, όμως, δεν είναι τόσο αποτελεσματικός στις περισσότερες των περιπτώσεων, πράγμα που καταφέρνουν οι επόμενες δύο μέθοδοι, η μέγιστη εντροπία και ο SVM.

Μέγιστη Εντροπία

Η κατηγοριοποίηση μέγιστης εντροπίας είναι μια τεχνική η οποία χρησιμοποιείται σε πολλές εφαρμογές της φυσικής γλώσσας επεξεργασίας (Natural Language Processing, NLP) και αρκετές φορές έχει αποδειχθεί πολύ αποτελεσματική, ειδικά σε περιπτώσεις που ο Naive Bayes αποτυγχάνει, όταν, δηλαδή, δεν μπορούν να γίνουν υποθέσεις για την ανεξαρτησία των συστατικών.

Ορίζεται από τον τύπο:

$$P_{ME}(c \mid d) := (1/Z(d))\exp(\Sigma_i \lambda_{i,c} F_{i,c}(d,c))$$

Όπου Z(d) είναι μια κανονικοποιημένη συνάρτηση, η $F_{i,c}$ είναι μια συνάρτηση των συστατικών f_i και τα $\lambda_{i,c}$ είναι παράμετροι μέτρησης των συστατικών για τα οποία ισχύει ότι όσο μεγαλύτερο είναι το $\lambda_{i,c}$ τόσο πιο αποτελεσματικός δείκτης είναι το f_i για την κλάση c. Τέλος, η κλάση c ορίζεται ως εξής:

$$F_{i,c}(d,c') := 1,\ ni(d)>0 \text{ και } c' = c \text{ ή } 0 \text{ για κάθε άλλη περίπτωση}$$

Οι τιμές των παραμέτρων πρέπει να έχουν οριστεί έτσι ώστε να μεγιστοποιείται η εντροπία της συγκεκριμένης κατανομής με τον περιορισμό ότι οι αναμενόμενες τιμές των συναρτήσεων των συστατικών που αναφέρονται στο μοντέλο, ισούται με τις αναμενόμενες τιμές που αναφέρονται στα προς εξέταση δεδομένα.

Μηχανές Υποστήριξης Διανύσματος

Οι μηχανές υποστήριξης διανύσματος (Support Vector Machines, SVM) έχουν αποδειχθεί πολύ αποτελεσματικές στην κατηγοριοποίηση κειμένου ξεπερνώντας κατά πολύ τις επιδόσεις του Naive Bayes (Joachims, 1998) ενώ δεν ταξινομούν πιθανοκρατικά όπως η μέγιστη εντροπία.

Η βασική ιδέα πίσω απ' αυτή τη μέθοδο είναι να βρεθεί ένα υπερεπίπεδο που αναπαρίσταται με το διάνυσμα $(\rightarrow)w$ που διαχωρίζει τα διανύσματα του εγγράφου της μιας κλάσης από την άλλη με c_j , που είναι η κατάλληλη κλάση του εγγράφου να παίρνει δύο τιμές {-1,1}. Αυτές οι δύο τιμές αντιστοιχούν στις καταστάσεις θετικό και αρνητικό. Έτσι έχουμε:

$$(\rightarrow)w := \Sigma_j a_j c_j (\rightarrow)d_j,\ (\rightarrow)a_j > 0$$

Τα a_j και d_j ονομάζονται διανύσματα υποστήριξης καθώς είναι τα μόνα διανύσματα του εγγράφου που συνεισφέρουν στο $(\rightarrow)w$.

2.7 Ταξινόμηση Κειμένου και Πολικότητα

Η δυαδική ταξινόμηση ενός εγγράφου η οποία εκφράζει είτε μία γενικά θετική είτε μία γενικά αρνητική άποψη ονομάζεται *ταξινόμηση πολικότητας συναισθήματος* (sentiment polarity classification ή polarity classification). Έχει γίνει αρκετή δουλειά πάνω σ' αυτό το θέμα, κυρίως στο περιεχόμενο κριτικών, όπως ο αλγόριθμος thumbs up, thumbs down που θα αναλύσουμε στη συνέχεια.

Ενώ το περιεχόμενο μιας θετικής ή αρνητικής άποψης συνήθως χαρακτηρίζεται σαν «μ' αρέσει » ή «δεν μ' αρέσει», σε ορισμένες περιπτώσεις μπορεί να έχει άλλη ερμηνεία. Για παράδειγμα, στην περίπτωση που θέλουμε να προσδιορίσουμε αν ο λόγος που βγάζει κάποιος πολιτικός σε ένα ντιμπέιτ συμφωνεί ή είναι αντίθετος με ένα συγκεκριμένο θέμα. Εδώ οι απόψεις θα μπορούσαν να ταξινομηθούν σαν «μπορεί να κερδίσει» ή «δεν μπορεί να κερδίσει».

Ένας εναλλακτικός τρόπος να συνοψίσουμε τις κριτικές είναι να εξάγουμε πληροφορίες για τον λόγο τον οποίο στους χρήστες αρέσει ή δεν αρέσει ένα προϊόν. Οι Kim και Hovy [8] σε σχετική έρευνά τους έχουν αναφέρει ότι οι εκφράσεις με πλεονεκτήματα και μειονεκτήματα μπορούν να διαφέρουν από τις θετικές και αρνητικές απόψεις, παρόλο που και οι δύο τρόποι, η άποψη 'Αυτό το αυτοκίνητο είναι καταπληκτικό' και η αιτία 'Αυτό το αυτοκίνητο κοστίζει μόνο 8000€' υπάρχουν για το σκοπό της ανάλυσης κειμένου στενά συνδεδεμένου. Η τεχνική της αναζήτησης της αιτίας μπορεί να βοηθήσει στο κατά πόσο οι κριτικές είναι εξυπηρετικές.

Ένα πιο γενικό πρόβλημα στην αξιολόγηση των συμπερασμάτων είναι όταν κάποιος προσπαθεί να προσδιορίσει την εκτίμηση του συγγραφέα με μία κλίμακα με πολλές βαθμίδες (ένα έως πέντε αστέρια), το οποίο μπορεί να καταλήξει σαν ταξινόμηση κειμένου με πολλές κλάσεις. Από την άλλη μεριά όμως η πρόβλεψη του βαθμού της θετικότητας παρέχει μια πιο λεπτομερή αξιολόγηση αφού κάθε κλάση μπορεί να έχει το δικό της λεξικό.

2.8 Σημαντικοί Όροι

Είναι σύνηθες στην ανάκτηση πληροφοριών (Information Retrieval, IR) να αναπαριστούμε ένα κομμάτι κειμένου με ένα χαρακτηριστικό διάνυσμα όπου οι είσοδοι αντιστοιχούν σε ξεχωριστούς όρους. Η term frequency (πόσες φορές εμφανίζεται ένας

όρος) ήταν πάντα μία πολύ σημαντική μονάδα μέτρησης στην IR, παρόλα αυτά όμως, ο Pang στο έργο του "Thumbs up? Sentiment Classification using Machine Learning Techniques" κατάφερε να κερδίσει καλύτερη απόδοση χρησιμοποιώντας *παρουσίαση* (presence) παρά συχνότητα. Τα δυαδικά διανύσματα, στα οποία οι είσοδοι απλώς δείχνουν αν ο όρος συμβαίνει (τιμή 1) ή όχι (τιμή 0) έχουν δημιουργήσει έναν πιο αποτελεσματικό τρόπο στην ταξινόμηση πολικότητας κειμένου απ' ότι τα *διανύσματα* πραγματικών τιμών στα οποία κάθε τιμή εισόδου αυξάνεται με την συχνότητα κατανομής του ανάλογου όρου. Αυτό μπορεί να δείξει μια ενδιαφέρουσα διαφορά μεταξύ της κλασικής θεματική ταξινόμηση κειμένου και της ταξινόμησης πολικότητας κειμένου. Ενώ σε ένα είναι πιθανόν να δίνεται έμφαση από συχνά γεγονότα συγκεκριμένων λέξεων κλειδιών, το γενικό συναίσθημα μπορεί να μην προβάλλεται μέσα από την επαναλαμβανόμενη χρήση των ίδιων όρων.

Σημαντικό ρόλο στην αναζήτηση πληροφοριών παίζει καμιά φορά και η *θέση των πληροφοριών* τις οποίες επεξεργαζόμαστε. Το αν ένας όρος βρίσκεται στη μέση ή προς το τέλος του εγγράφου μπορεί να έχει μια σημαντική επιρροή στο πόσο επηρεάζει το συνολικό συναίσθημα ή ακόμα και την κατάσταση υποκειμενικότητας του κειμένου που εξετάζουμε. Γι' αυτό το λόγο οι πληροφορίες θέσης κωδικοποιούνται κι αυτές στα διανύσματα που χρησιμοποιούμε.

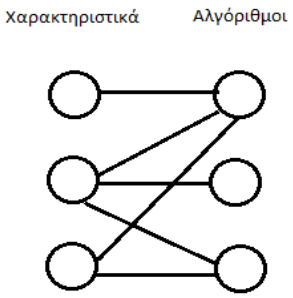
Οι *πληροφορίες για τα μέρη του λόγου* (Part-of-speech, POS) χρησιμοποιούνται εξίσου συχνά στην ανάλυση συναισθημάτων και με πολύ καλά αποτελέσματα. Για παράδειγμα, τα επίθετα, για τα οποία έχει αποδειχθεί ότι είναι πολύ καλοί δείκτες συναισθήματος. Σχετική έρευνα έχει αποκαλύψει ότι υπάρχει μεγάλη σχέση μεταξύ της παρουσίας των επιθέτων σε μία πρόταση και την αντικειμενικότητά της. Αυτό όμως δεν σημαίνει πως και τα άλλα μέρη του λόγου δεν συνεισφέρουν το ίδιο. Στην έρευνά του ο Pang, για την εύρεση πολικότητας σε κριτικές ταινιών, χρησιμοποίησε μόνο επίθετα σαν δείχτες και παρατήρησε πως αυτή η μέθοδος λειτουργούσε χειρότερα απ' ότι αν χρησιμοποιούσε και άλλα μέρη του λόγου όπως ρήματα και ουσιαστικά. Παρόμοιες έρευνες έχουν γίνει και από τους Farah Benamara, Carmine Cesarano και Diego Reforgiato [9].

Η διαχείριση της άρνησης είναι ένα σημαντικό ζήτημα στην ανάλυση συναισθημάτων. Σκεφτείτε ότι οι προτάσεις 'Μου αρέσει αυτό το βιβλίο' και 'Δεν μου αρέσει αυτό το βιβλίο' θεωρούνται όμοιες απ' τα ευρέως χρησιμοποιούμενα μέτρα σύγκρισης ομοιότητας. Η μόνη διαφορετική λέξη είναι αυτή της άρνησης (δεν) που

όμως εξαναγκάζει αυτές τις δύο προτάσεις να βρίσκονται σε διαφορετικές κλάσεις. Ένας άλλος τρόπος έκφρασης άρνησης είναι με σαρκασμό ή ειρωνεία, τα οποία όμως, τις περισσότερες φορές είναι δύσκολο να εντοπιστούν. Έχουν γίνει διάφορες προσπάθειες για τη διαχείριση της άρνησης, τα ποσοστά επιτυχίας, όμως στις περισσότερες, ήταν πολύ χαμηλά.

2.9 Τεχνικές Ανάλυσης Συναισθήματος

Υπάρχουν πολλές μέθοδοι για την επίτευξη της ανάλυσης συναισθήματος. Εδώ θα προσπαθήσουμε να τις εξετάσουμε μέσα από δύο διαστάσεις, τα χαρακτηριστικά της ανάλυσης συναισθήματος και τους αλγορίθμους. Σαν χαρακτηριστικά ορίζουμε μια μετρήσιμη ιδιότητα των εγγράφων που είναι έτοιμα για ανάλυση συναισθήματος, τέτοια είναι η πολικότητα, η συχνότητα και οι πληροφορίες για τα μέρη του λόγου. Σαν αλγόριθμο της ανάλυσης συναισθήματος ορίζουμε μια μέθοδο που είναι βασισμένη στα χαρακτηριστικά για την επιλογή της πολικότητας του εγγράφου. Στο σχήμα 1 θα δούμε τη σχέση ανάμεσα στα χαρακτηριστικά και τους αλγόριθμους. Οι κόμβοι της αριστερής στήλης παρουσιάζουν διάφορα χαρακτηριστικά ενώ τις δεξιάς αλγορίθμους. Οι γραμμές που ενώνουν αυτούς τους κόμβους σημαίνουν ότι ο συγκεκριμένος αλγόριθμος χρειάζεται αυτό το χαρακτηριστικό για να γίνει η ανάλυση. Πρέπει να σημειώσουμε ότι ένα χαρακτηριστικό μπορεί να χρησιμοποιείται από πολλούς αλγορίθμους και ένας αλγόριθμος μπορεί να χρησιμοποιεί πολλά χαρακτηριστικά.



Εικόνα 1 Σχέση Χαρακτηριστικών και Αλγορίθμων

Τα χαρακτηριστικά χωρίζονται σε δύο κατηγορίες, τα *χαρακτηριστικά λεξικού συναισθήματος* και τα *χαρακτηριστικά χωρίς λεξικό συναισθήματος*. Η πρώτη κατηγορία επιτυγχάνεται με τα λεγόμενα λεξικά συναισθήματος τα οποία χρησιμοποιούν όρους με πληροφορίες πολικότητας, συνήθως είναι ένας αριθμός που μας δείχνει πόσο θετική ή αρνητική είναι η συγκεκριμένη λέξη. Η δεύτερη κατηγορία δεν επιτυγχάνεται μέσω τέτοιων λεξικών αλλά με χρήση των πληροφοριών συντακτικού, term frequency (πόσο συχνά εμφανίζεται ο όρος) και term presence (η παρουσία του όρου).

Δύο από τις κυριότερες κατηγορίες αλγορίθμων είναι οι *αλγόριθμοι βασισμένοι στα λεξικά συναισθήματος* και οι *αλγόριθμοι βασισμένοι στη μηχανική μάθηση* για τους οποίους μιλήσαμε προηγουμένως. Οι περισσότεροι αλγόριθμοι της ανάλυσης συναισθήματος ανήκουν σε μία απ’ τις δύο κατηγορίες.

Οι αλγόριθμοι της πρώτης κατηγορίας κατασκευάζουν μεθόδους βασισμένες σε χαρακτηριστικά λεξικού συναισθήματος και μέσω των αριθμό θετικότητας των λέξεων υπολογίζουν την πολικότητα του κειμένου. Μια απλή μέθοδος για εύρεση της πολικότητας είναι η σύγκριση του μέσου όρου των θετικών λέξεων ενός κειμένου με τον μέσο όρο των αρνητικών. Αν ο μέσος όρος των θετικών λέξεων είναι μεγαλύτερος τότε το κείμενο έχει θετική πολικότητα, αν όχι έχει αρνητική. Αν κατά τη σύγκριση διαπιστώσουμε ότι η διαφορά δεν είναι μεγάλη τότε το κείμενο χαρακτηρίζεται ουδέτερο.

Αρκετές φορές έχει αναφερθεί ότι επιτυγχάνουμε καλύτερα αποτελέσματα με ένα μόνο αλγόριθμο ο οποίος χρησιμοποιεί όσα περισσότερα χαρακτηριστικά γίνεται. Παρόλα αυτά δεν έχει γίνει κάποια έρευνα που να αποδεικνύει κάτι τέτοιο. Κάθε αλγόριθμος έχει τα δυνατά του σημεία και τις αδυναμίες του. Ίσως αν εκμεταλλευόμασταν τα δυνατά σημεία διαφόρων αλγορίθμων οι οποίο συνολικά χρησιμοποιούν όλα τα χαρακτηριστικά, το αποτέλεσμα να μας εξέπληττε.

2.10 Ο Αλγόριθμος Thumbs Up Thumbs Down

Σκοπός του συγκεκριμένου αλγορίθμου είναι να κατατάξει μια σειρά από κριτικές σε προτεινόμενες (thumbs up) ή μη προτεινόμενες (thumbs down) [2]. Σαν είσοδο παίρνει μια κριτική (ένα κομμάτι κειμένου) και σαν έξοδο την ταξινομεί ανάλογα με το νόημά της. Αρχικά, αναγνωρίζει τις προτάσεις του κειμένου που περιέχουν επίθετο ή επίρρημα και στη συνέχεια, προσδιορίζει το σημασιολογικό προσανατολισμό (semantic orientation) κάθε μίας εξ αυτών. Μια πρόταση έχει θετικό σημασιολογικό προσανατολισμό όταν έχει καλές συσχετίσεις (π.χ. καλόγουστα σκηνικά) ενώ αρνητικό όταν έχει κακές συσχετίσεις (π.χ. απογοητευτικές ερμηνείες). Έπειτα, κατατάσσεται η κριτική σε μία κλάση, προτεινόμενες ή μη προτεινόμενες κριτικές. Αν, τελικά, ο μέσος όρος είναι θετικός τότε το αντικείμενο το οποίο αφορούν οι κριτικές προτείνεται, αν όχι δεν προτείνεται.

Στο πρώτο σκέλος του αλγόριθμου επιλέγονται οι προτάσεις που περιέχουν επίθετο ή επίρρημα. Ένα επίθετο όμως μπορεί να έχει και αρνητική και θετική σημασία ανάλογα με τη λέξη που ακολουθεί. Γι’ αυτό το λόγο ο αλγόριθμος ελέγχει δύο συνεχόμενες λέξεις, το επίθετο ή το επίρρημα και τη λέξη που ακολουθεί. Αρχικά, οι σημαντικές λέξεις της κάθε πρότασης καθορίζονται ανάλογα με το τι είναι: JJ για επίθετα, NN για ουσιαστικά, RB για επιρρήματα και VB για ρήματα. Δύο συνεχόμενες λέξεις εξάγονται αν οι ετικέτες τους συμφωνούν με τον παρακάτω πίνακα.

Πίνακας 5 Συνδυασμοί συνεχόμενων λέξεων για να εξεταστεί η πιθανότητα εξαγωγής

Πρώτη Λέξη	Δεύτερη Λέξη	Τρίτη Λέξη
JJ	NN ή NNS	οτιδήποτε
RB ή RBR ή RBS	JJ	όχι NN ή NNS
JJ	JJ	όχι NN ή NNS

NN ή NNS	JJ	όχι NN ή NNS
RB ή RBR ή RBS	VB ή VBD ή VBN ή VBG	οτιδήποτε

Για παράδειγμα, στην τρίτη σειρά του πίνακα βλέπουμε ότι αν η πρώτη λέξη είναι επίρρημα, η δεύτερη επίθετο και η τρίτη οτιδήποτε άλλο εκτός από ουσιαστικό τότε μπορεί να γίνει η εξαγωγή των δύο συνεχόμενων λέξεων.

Στο επόμενο βήμα ελέγχεται ο σημασιολογικός προσανατολισμός. Για τον σκοπό αυτό χρησιμοποιείται ο αλγόριθμος PMI-RI ο οποίος χρησιμοποιεί κοινές πληροφορίες σαν μέτρο δύναμης των σημασιολογικών σχέσεων των δύο λέξεων. Ο PMI (Pointwise Mutual Information) ορίζεται ως $PMI(word_1, word_2) = \log_2 (pword_1 \ \& \ pword_2 \ / \ pword_1 \ pword_2 \ )$. Όπου $pword_1 \ \& \ pword_2$ είναι οι πιθανότητα να βρεθούν και οι δύο λέξεις μαζί. Η αναλογία αυτή είναι το μέτρο της σημασιολογικής εξάρτησης των λέξεων, ενώ ο λογάριθμος της αναλογίας αυτής είναι η ποσότητα των πληροφοριών που αποκτούμε για την κάθε λέξη όταν παρατηρούμε την άλλη.

Ο σημασιολογικός προσανατολισμός της φράσης (phrase) ορίζεται ως SO = PMI(phrase, “excellent”) - PMI(phrase, “poor”). Οι λέξεις excellent και poor επιλέχθηκαν επειδή στο σύστημα five star η λέξη excellent έχει βάρος πέντε αστέρια ενώ η poor ένα. Έτσι, λοιπόν, ο σημασιολογικός προσανατολισμός είναι θετικός όταν η φράση έχει σχέση με τη λέξη excellent και αρνητικός όταν έχει σχέση με τη λέξη poor.

Ο PMI-RI υπολογίζει το PMI θέτοντας ερωτήματα σε μια μηχανή αναζήτησης και καταγράφοντας τον αριθμό των αποτελεσμάτων και ανάλογα με το αποτέλεσμα, κατατάσσει τις προτάσεις σε προτεινόμενες ή μη προτεινόμενες. Τέλος, βγάζει τον μέσο όρο κάθε πρότασης και ανάλογα με το πρόσημο η κριτική κρίνεται θετική ή αρνητική.

Στον πίνακα 5 βλέπουμε την εφαρμογή του αλγόριθμου σε μία κριτική. Σε αυτή τη φάση έχουν απομονωθεί οι προτάσεις, έχουν εξαχθεί οι κατάλληλες συνεχόμενες λέξεις και έχει υπολογιστεί το PMI για κάθε μία απ’ αυτές. Βλέποντας τον μέσο όρο παρατηρούμε ότι η κριτική είναι θετική, κάτι που θα περιμέναμε γιατί οι περισσότερες προτάσεις έχουν χαρακτηριστεί σαν προτεινόμενες.

Πίνακας 6 Παράδειγμα επεξεργασίας μιας κριτικής όπου έχει ταξινομηθεί σαν *προτεινόμενη*

Εξαγόμενες Λέξεις	Ετικέτες	Σημασιολογικός Προσανατολισμός
online experience	JJ NN	2,253
low fees	JJ NNS	0,333
local branch	JJ NN	0,421
small part	JJ NN	0,053
online service	JJ NN	2,780
printable version	JJ NN	-0,705
direct deposit	JJ NN	1,288
Inconveniently located	RB VBN	-1,541
Μέσος Όρος SO		0,610

Στην αντίθετη περίπτωση του πίνακα 6 οι περισσότερες προτάσεις της κριτικής έχουν χαρακτηριστεί σαν μη προτεινόμενες οπότε και ολόκληρη η κριτική είναι αρνητική.

Πίνακας 7 Παράδειγμα επεξεργασίας μιας κριτικής όπου έχει ταξινομηθεί σαν *μη προτεινόμενη*

Εξαγόμενες Λέξεις	Ετικέτες	Σημασιολογικός Προσανατολισμός
little difference	JJ NN	-1,615
Clever tricks	JJ NNS	-0,040
programs such	NNS JJ	0,117
possible moment	JJ NN	-0,668
unethical practices	JJ NNS	-8,484
low funds	JJ NNS	-6,843

old man	JJ NN	-2,566
other problems	JJ NNS	-2,748
probably wondering	RB VBG	-1,830
Μέσος Όρος SO		-2,4677

Ο αλγόριθμος αυτός έχει πετύχει 74% ακρίβεια σε 410 κριτικές για ταινίες, αυτοκίνητα, τράπεζες και ταξιδιωτικούς προορισμούς απ' την ιστοσελίδα Epinions. Η ακρίβεια φτάνει το 84% για αυτοκίνητα ενώ για ταινίες το 66%.

Ένας άλλος αλγόριθμος που μπορεί να αντικαταστήσει τον PMI-RI είναι αυτός του Χατζηβασιλόγλου και McKeown (1997). Σύμφωνα μ' αυτόν, αφού εξαχθούν όλοι οι συνδυασμοί των επιθέτων απ' το κείμενο, αυτά ελέγχονται για να βρεθούν εκείνα που έχουν τον ίδια σημασιολογικό προσδιορισμό. Το αποτέλεσμα που προκύπτει είναι ένα γράφημα που έχει στους κόμβους του τα επίθετα τα οποία συνδέονται σύμφωνα με την ομοιότητα του σημασιολογικού τους προσανατολισμού. Στη συνέχεια το γράφημα, επεξεργάζεται με τη βοήθεια ενός ειδικού αλγορίθμου και δημιουργεί δύο ομάδες από επίθετα, ανάλογα με τις συνδέσεις τις οποίες υπάρχουν. Τέλος, σύμφωνα με τη θεωρία ότι τα θετικής σημασίας επίθετα χρησιμοποιούνται πιο συχνά, η ομάδα με τη μεγαλύτερη συχνότητα είναι και αυτή με τον μεγαλύτερο σημασιολογικό προσανατολισμό.

Ο αλγόριθμος αυτός μπορεί να πετύχει ακρίβεια από 78% έως 92%, ανάλογα με το μέγεθος των δεδομένων.

2.11 Εφαρμογές

Εκτός από την πολικότητα κειμένου, η sentiment analysis έχει κι άλλες εφαρμογές στο διαδίκτυο.

Εφαρμογές σε Website αξιολόγησης προϊόντων. Πρόκειται για την προσπάθεια εύρεσης θετικών ή αρνητικών αξιολογήσεων και τη στατιστική επεξεργασίας τους από κατάλληλα Websites για την εξαγωγή συμπερασμάτων.

Εφαρμογές σαν μια τεχνολογία υποσυστατικών. Τα συστήματα sentiment analysis και opinion mining έχουν συχνά σημαντικό ρόλο σαν αρχικές τεχνολογίες για άλλα συστήματα. Χρησιμοποιούνται σαν πρόσθετα στα συστήματα συστάσεων

(recommendation systems), ώστε να τα αποτρέψουν να συστήνουν αντικείμενα τα οποία λαμβάνουν πολλές αρνητικές αντιδράσεις. Επίσης, εντοπίζουν την ανταγωνιστική γλώσσα στα e-mail ή σε άλλους τύπους επικοινωνίας και τις ακατάλληλες ή αρνητικές διαφημίσεις απ' τις ιστοσελίδες. Τέλος, χρησιμοποιούνται στην *απάντηση ερωτήσεων* αφού χρησιμοποιούν πολλές πληροφορίες για το πώς αντιμετωπίζονται οι οντότητες.

Εφαρμογές σε Business και Government Intelligence. Η sentiment analysis είναι ένα εξαιρετικό εργαλείο για τη διαχείριση θεμάτων Business Intelligence. Βοηθάει τις εταιρίες να βλέπουν τι άποψη έχουν οι καταναλωτές για τα προϊόντα τους και να βελτιώνονται. Όσο αφορά το Government Intelligence, μπορεί να χρησιμοποιηθεί στον εντοπισμό εχθρικής ή αρνητικής επικοινωνίας.

Εφαρμογές σε διάφορους τομείς. Κατά καιρούς χρησιμοποιούμε sentiment analysis σε διάφορους τομείς, όπως για παράδειγμα την πολιτική. Έχουν γίνει έρευνες για το τι σκέφτονται οι ψηφοφόροι ή και ακόμα τι κάνουν οι πολιτικοί για να βελτιώσουν την ποιότητα των πληροφοριών στις οποίες έχουν πρόσβαση.

3 Συστήματα Ανάλυσης Συναισθήματος

Υπάρχουν πραγματικά πολλά συστήματα για να κάνει κανείς ανάλυση συναισθήματος, τα περισσότερα από τα οποία μπορεί να έχουν εντυπωσιακά αποτελέσματα.

3.1 Το RapidMiner

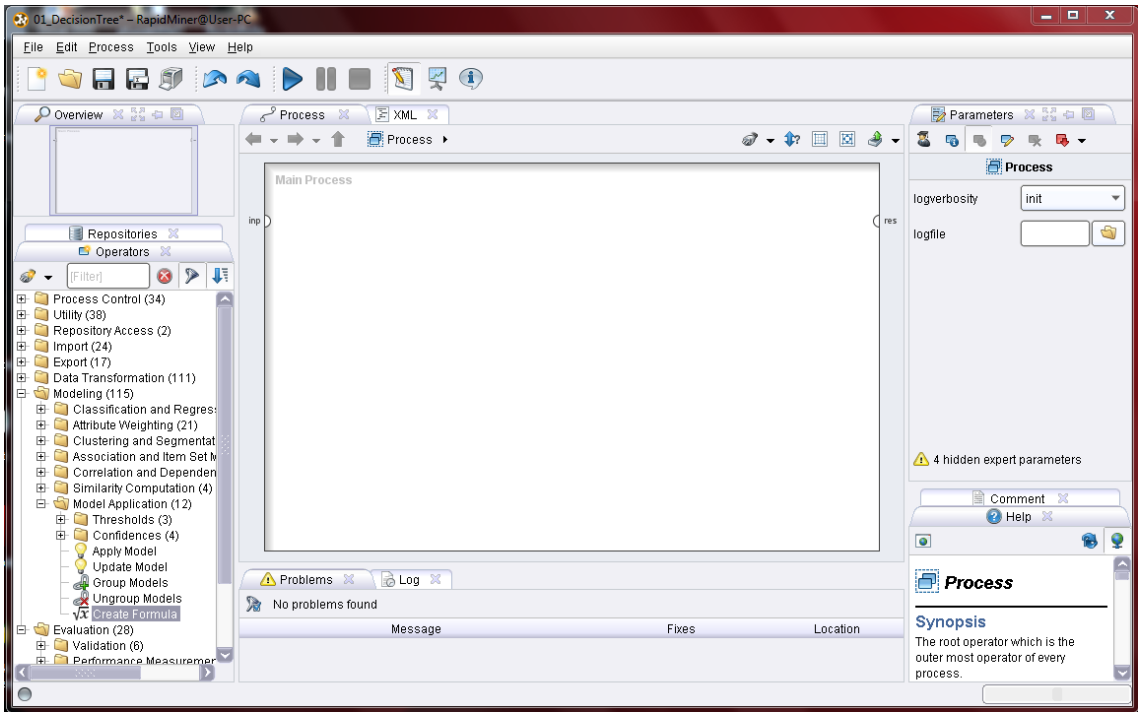
Το RapidMiner είναι ένα open source λογισμικό data mining. Χρησιμοποιείται στην έρευνα, την εκπαίδευση, την ανάπτυξη εφαρμογών και την βιομηχανική ανάπτυξη. Σε μία έρευνα του KDnuggets, μία εφημερίδα για data mining, ψηφίστηκε δεύτερο στα εργαλεία data mining/analytics το 2009 και πρώτο το 2010.

Η εφαρμογή αυτή ξεκίνησε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και Simon Fisher στη μονάδα Τεχνητής Νοημοσύνης του Πανεπιστημίου του Ντορτμουντ. Το 2006 συνεργάστηκαν με την εταιρία Rapid-I η οποία είναι και η κύρια εταιρία για την ανάπτυξή του. Αυτή τη στιγμή υπάρχουν πάνω από 30 άτομα παγκοσμίως που ασχολούνται με την εξέλιξη και τις επεκτάσεις του λογισμικού.

Το RapidMiner είναι γραμμένο σε Java και γι' αυτό μπορεί να τρέχει σε όλα τα δημοφιλή λειτουργικά συστήματα. Ένας μεγάλος αριθμός από χειριστές (operators) καθορίζονται στο RapidMiner και μαζί με τα plugins του καλύπτουν κάθε πλευρά data mining. Γίνεται ένας καθαρός χειρισμός των δεδομένων χωρίς να χρειάζεται να ξέρουμε το είδος των δεδομένων ή τις διαφορετικές όψεις τους. Οι χειριστές αυτοί είναι ευέλικτοι σε δεδομένα εισόδου, εξόδου αφού εξυπηρετούν διάφορα είδη αρχείων όπως excel, SPSS, data sets από βάσεις δεδομένων όπως οι Oracle, mySQL, PostgreSQL, Microsoft SQL Server, Sybase, και dBase. Επίσης αποδέχεται Sparse αρχεία όπως SVMight και mySVM.

Ακολουθεί μια πολυδιάστατη αντίληψη για τα δεδομένα, πράγμα που το καθιστά ικανό να αποθηκεύει διαφορετικές όψεις του ίδιου πίνακα δεδομένων. Γι' αυτό το λόγο διευκολύνει τη διάταξη πολλαπλών όψεων σε επίπεδα στον κεντρικό πίνακα δεδομένων.

Έχει ένα ευέλικτο διαδραστικό σχέδιο που εισάγει τον χρήστη σε εναλλακτικά μεταδεδομένα των διαθεσίμων data sets και του δίνει τη δυνατότητα να κάνει αυτοματοποιημένη αναζήτηση και βελτιστοποιημένη προεργασία τα οποία είναι και τα δύο, απολύτως απαραίτητα για μια αποτελεσματική διεργασία data mining.



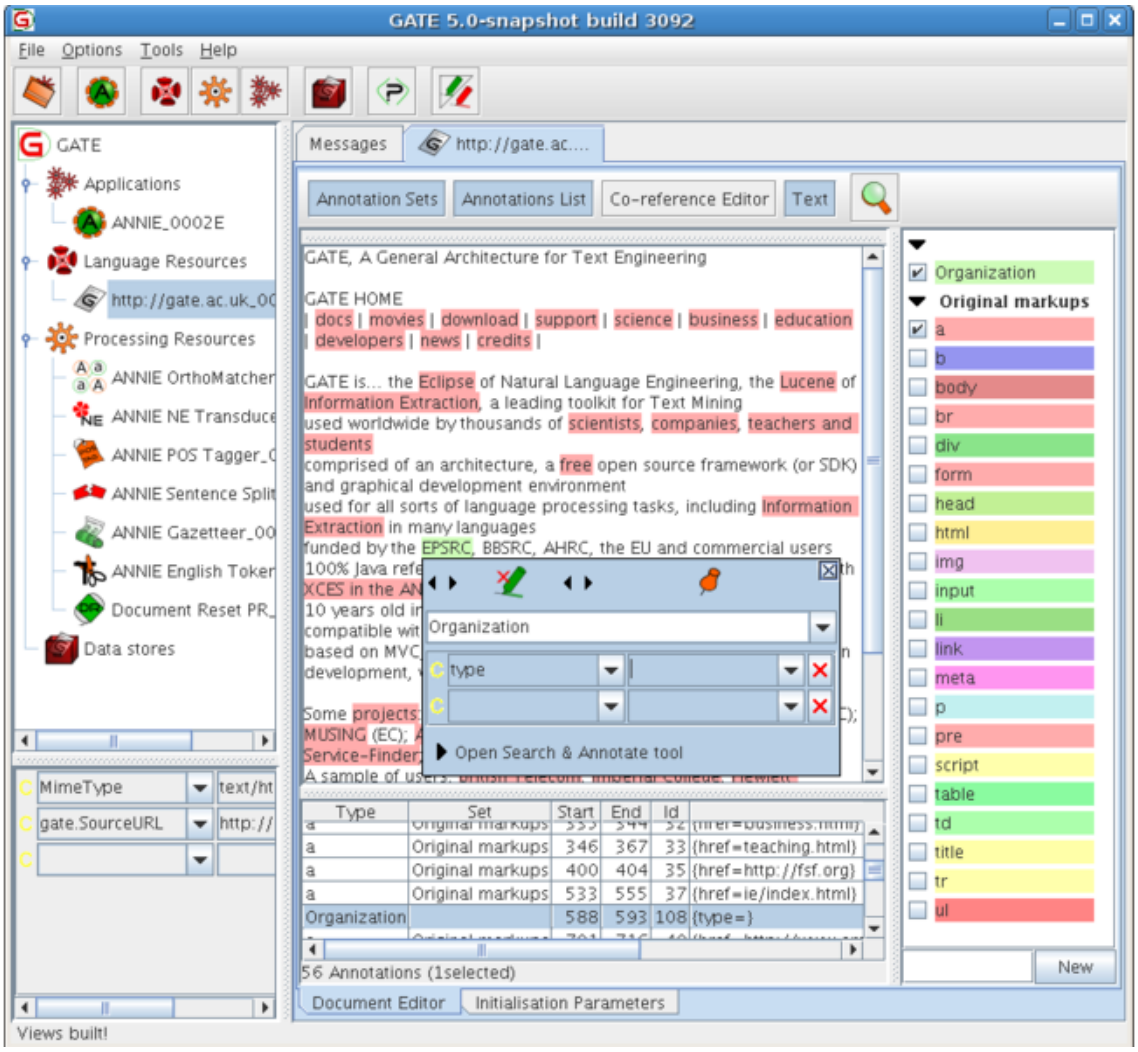
Εικόνα 2 Το γραφικό περιβάλλον του RapidMiner

3.2 Το GATE

Ένα αρκετά διαδεδομένο εργαλείο opinion mining είναι το GATE (General Architecture for Text Engineering). Το GATE είναι ένα ελεύθερο λογισμικό το οποίο άρχισε να αναπτύσσεται στο πανεπιστήμιο του Sheffield το 1995 και τώρα χρησιμοποιείται παγκοσμίως από επιστήμονες, εταιρείες, καθηγητές και φοιτητές. Είναι μια υποδομή ανάπτυξης συστατικών λογισμικού επεξεργασίας της φυσικής γλώσσας και κύριος στόχος του είναι η ανάλυση κειμένου κάθε είδους. Το GATE περιλαμβάνει:

- Το GATE Developer, ένα ενσωματωμένο περιβάλλον ανάπτυξης για επεξεργασία της φυσικής γλώσσας μαζί με ένα σύστημα εξαγωγής πληροφοριών και ένα περιεκτικό σύνολα από άλλα πρόσθετα.
- Το GATE Teamware, ένα περιβάλλον για σχόλια
- Το GATE Embedded, μια βιβλιοθήκη αντικειμένων

- Μια υψηλού επιπέδου εικόνα του πως συντίθεται το λογισμικό επεξεργασίας της γλώσσας
- Μια διαδικασία δημιουργίας υπηρεσιών υποστήριξης



Εικόνα 3 Το γραφικό περιβάλλον του GATE

Πάνω απ' όλα όμως το GATE περιλαμβάνει συστατικά για διάφορα θέματα επεξεργασίας της γλώσσας όπως γραμματική και μορφολογική ανάλυση λέξεων, εργαλεία ανάκτησης πληροφοριών, συστατικά για την εξαγωγή πληροφοριών για διάφορες γλώσσες και πολλά άλλα. Το GATE Developer και το GATE Embedded περιλαμβάνουν ένα σύστημα εξαγωγής πληροφοριών το οποίο ονομάζεται ANNIE και χρησιμοποιείται για να δημιουργεί RDF ή OWL μεταδεδομένα για μη δομημένο περιεχόμενο.

Τα εργαλεία που προσφέρει είναι τα πιο ολοκληρωμένα που υπάρχουν σε παρόμοιο πρόγραμμα. Είναι ένα ανοιχτό λογισμικό το οποίο καταγράφεται και διατηρείται. Μπορεί να επεκταθεί και να επαναχρησιμοποιηθεί αφού ο κώδικάς του υπάρχει σε εφαρμογές περισσότερο από κάθε άλλο παρόμοιο σύστημα. Χαρακτηρίζεται από διαφάνεια, οι χρήστες έχουν μεγαλύτερη επίγνωση του πως δουλεύει το σύστημα και ανθεκτικότητα. Σε αντίθεση με άλλα ερευνητικά λογισμικά, το GATE δημιουργήθηκε για παραγωγή. Ελέγχεται ποιοτικά από ακριβής εφαρμογές ποσοτικών μετρήσεων, οι οποίες εξασφαλίζουν ότι η συμπεριφορά του θα είναι προβλέψιμη.

Το γεγονός, όμως, ότι περιλαμβάνει τόσους πολλούς τομείς (ανάλυση κειμένου, επεξεργασία φυσικής γλώσσας, μηχανική μάθηση, ανάλυση συναισθήματος) και τόσα πολλά εργαλεία το καθιστά ένα αρκετά βαρύ και δύσχρηστο πρόγραμμα.

3.3 LingPipe

Το LingPipe είναι ένα εργαλείο επεξεργασίας κειμένου και επεξεργασίας φυσικής γλώσσας, γραμμένο σε Java που χρησιμοποιεί υπολογιστική γλωσσολογία. Διατίθεται με δύο άδειες εμπορικού λογισμικού και για εκπαιδευτικούς λόγους. Αν θέλουμε να κατεβάσουμε κάποιο demo μπορούμε να επισκεφτούμε την ιστοσελίδα http://alias-i.com/lingpipe/index.html . Δημιουργήθηκε το 1995 μέσα από μια συνεργασία μεταξύ μιας ομάδας φοιτητών του πανεπιστημίου της Πενσυλβάνια με σκοπό τη συμμετοχή τους στο διαγωνισμό DARPA MUC-6 τον οποίο και κέρδισαν. Το βραβείο ήταν ένα συμβόλαιο με την Alias-i . Τον Σεπτέμβριο του 2003 άρχισε να διατίθεται σαν ελεύθερο λογισμικό

Τα θέματα με τα οποία ασχολείται κυρίως είναι η εύρεση ονομάτων ανθρώπων, οργανισμών, τοποθεσιών από διάφορες ειδήσεις, η αυτόματη ταξινόμηση αποτελεσμάτων από αναζητήσεις στο Twitter σε κατηγορίες, η σωστή ορθογραφία διαφόρων ερωτημάτων και φυσικά η ανάλυση συναισθήματος.

Όσο αφορά την ανάλυση συναισθήματος, σκοπός του είναι ο διαχωρισμός των υποκειμενικών και των αντικειμενικών προτάσεων και έπειτα ο διαχωρισμός των θετικών και των αρνητικών κριτικών.

3.4 OpinionFinder

Το OpinionFinder είναι και αυτό ένα open source λογισμικό το οποίο αναπτύχθηκε στο Πανεπιστήμιο του Πίτσμπουργκ με επιρροές απ' το Πανεπιστήμιο του Κόρνελ και το Πανεπιστήμιο της Γιούτα και έκανε την εμφάνιση του το 2006. Χρησιμοποιείται σε θέματα ανάλυσης συναισθήματος κάνοντας χρήση NLP. Αναγνωρίζει τις υποκειμενικές προτάσεις και επισημαίνει διάφορες απόψεις υποκειμενικότητας, συμπεριλαμβανομένου της πηγής της υποκειμενικότητας και λέξεις οι οποίες περιλαμβάνονται σε φράσεις βρίσκοντας το θετικό ή αρνητικό τους συναίσθημα.

3.4.1 Πως Λειτουργεί

Το OpinionFinder λειτουργεί σε δύο βασικά στάδια. Στο πρώτο στάδιο γίνεται η επεξεργασία των δεδομένων. Αρχικά παίρνει το εισερχόμενο κείμενο και αφού αφαιρέσει κάθε HTML ή XML πληροφορία, χωρίζει τις προτάσεις και βάζει όπου χρειάζεται τον δείχτη POS χρησιμοποιώντας το OpenNLP (το OpenNLP είναι μια δομή η οποία χρησιμοποιείται για τον συντονισμό διαφορετικών project που έχουν μια κάποια σχέση με NLP). Στη συνέχεια απομακρύνονται οι αλλοιωμένες λέξεις με τη βοήθεια του προγράμματος SCOL v1K του Steven Abney. Έπειτα με τη βοήθεια κάποιων άλλων προγραμμάτων (SUNDANCE, Sentence UNDerstanding And Concept Extraction και Autoslog-TS) αναγνωρίζονται τα πρότυπα εξαγωγής που χρειάζονται από τα αναγνωριστικά των προτάσεων και το SourceFinder, το οποίο αναγνωρίζει την πηγή του υποκειμενικού περιεχομένου, διαχωρίζοντας τα συναισθήματα του συγγραφέα σε σχετικά ή απλές αναφορές. Μια τελική ανάλυση γίνεται με τα συντακτικά δέντρα τα οποία μετατρέπονται σε δέντρα εξάρτησης για την εύρεση της οντότητας και του υποκειμένου.

Στο δεύτερο στάδιο έχουμε την ανάλυση συναισθήματος και την υποκειμενικότητα. Σ' αυτό το σημείο, ο αλγόριθμος Naive Bayes αναγνωρίζει τις υποκειμενικές προτάσεις. Στη συνέχεια, μια άμεση υποκειμενική έκφραση και ένας ταξινομητής γεγονότων του λόγου επισημαίνουν τις υπόλοιπες υποκειμενικές εκφράσεις και τα γεγονότα που βρίσκονται στο έγγραφο χρησιμοποιώντας το WordNet (ένα αγγλικό λεξικό). Τέλος, εφαρμόζεται η ανάλυση συναισθήματος στις προτάσεις που έχουν προσδιοριστεί σαν υποκειμενικές. Αυτό επιτυγχάνεται με δύο ταξινομητές οι οποίοι έχουν δημιουργηθεί απ' το BoosTexter, ένα πρόγραμμα μηχανικής μάθησης.

Η εύρεση της οντότητας μαζί με τα δέντρα εξάρτησης θα μας βοηθήσουν να φιλτράρουμε το περιεχόμενο και να συμπεριλάβουμε μόνο ότι έχει να κάνει με συναίσθημα και να αποκαλύψουμε σχετικά θέματα που υπάρχουν σε διάφορες συζητήσεις.

Η εύρεση της υποκειμενικότητας και η κατάταξη των γεγονότων του λόγου είναι προκλήσεις που έχουν αναγνωριστεί από πολλές έρευνες στο θέμα της ανάλυσης συναισθήματος Αυτό το σύστημα συνδυάζει μερικές διαδικασίες για την επίτευξη αυτών των στόχων με αποτέλεσμα να μας βοηθάει πραγματικά να μειώσουμε το σχετικό με το συναίσθημα περιεχόμενο ενός θέματος.

3.5 Άλλα συστήματα

3.5.1 Το Attensity Analytics Suite

Το Attensity Analytics Suite δεν είναι ένα open source λογισμικό. Επιτρέπει στον κάθε οργανισμό να ακούσει και να αναλύσει πολλαπλά κανάλια με συνομιλίες των πελατών, να συνδυάζει τη διορατικότητα με την εργασία και δρα σ' αυτές τις συζητήσεις απαντώντας στα ερωτήματα των πελατών γρήγορα. Μ' αυτό τον τρόπο βοηθάει τις επιχειρήσεις να βελτιστοποιήσουν το service τους, να κατανοήσουν την απήχηση που θα έχει η βελτιστοποίηση των προϊόντων στο συναίσθημα των καταναλωτών, να ενημερώνονται έγκαιρα για ανταγωνιστικά θέματα, να ερευνούν πιθανές απάτες και να μετράνε την αποτελεσματικότητα της τελευταίας τους καμπάνιας. Το Attensity Analytics Suite αποτελείτε από πέντε υποσυστήματα, τα οποία χρησιμοποιούν την ανάλυση συναισθήματος και μας βοηθάνε σε ξεχωριστούς κλάδους.

Το Attensity Analyze επιτρέπει στον χρήστη, με τη βοήθεια της ανάλυσης συναισθήματος, να απομονώνει συζητήσεις από e-mail, forum, CRM και διάφορα άλλα και να τα μετατρέπει σε πιθανές προβλέψεις πάντα για το καλό της επιχείρησής του. Μέσα στις δυνατότητές του είναι να προβλέπει αν ο καταναλωτής θέλει να αγοράσει κάποιο προϊόν, αν επιθυμεί να συμβούν κάποια πράγματα κ.τ.λ.

Το Attensity360 for Social Media κάνει παρόμοια δουλειά με το Analyze, αν και είναι ανεξάρτητα μεταξύ τους. Στην ουσία, αυτό που κάνει το Attensity360 είναι να παρακολουθεί εκατομμύρια εγγραφές από τις σελίδες κοινωνικής δικτύωσης, απομονώνει αυτές που έχουν να κάνουν με το συγκεκριμένο θέμα και παίρνει

πληροφορίες που βοηθούν την επιχείρηση να προωθηθεί και να εξυπηρετεί καλύτερα τους πελάτες.

Το επόμενο υποσύστημα είναι το Attensity Opinion Insights, δουλειά του οποίου είναι να αναλύει τις απόψεις των καταναλωτών από πάνω από 10000 σελίδες με κριτικές από καταναλωτές και ειδικούς με αποτέλεσμα να καταφέρνει να ανακαλύπτει τι άποψη έχει ο κόσμος για τη συγκεκριμένη εταιρεία, τα προϊόντα της και τους ανταγωνιστές. Οι πληροφορίες που δίνει έρχονται μέσω ευέλικτων και έγκαιρων αναφορών μαζί με παρουσιάσεις αναλυτών.

Ένα άλλο θέμα που προκύπτει είναι αν μπορεί να υπάρξει κάποιο σύστημα που να μπορεί να εντοπίζει πιθανές απάτες ή ανεπιθύμητες έρευνες από ανταγωνιστές. Αυτό το σύστημα δημιουργήθηκε και ονομάζεται. Attensity Discover.

Εικόνα 4 Ο τρόπος με τον οποίο δουλεύει το Attensity Discovery

Τέλος, το Attensity Intelligence χρησιμοποιεί κάθε είδους κυβερνητικές υπηρεσίες σε καθημερινή βάση με σκοπό την εξαγωγή γεγονότων και σχέσεων από ένα κείμενο, και στη συνέχεια αναζητά και αναλύει τις πληροφορίες του αποτελέσματος.

Το Attensity μας δίνει πραγματικά πολλές δυνατότητες και μέσα απ' αυτές πολλές ευκαιρίες να βελτιώσουμε την επιχείρηση μας, να βρούμε που υπερέχει και που υστερεί, τι γνώμη έχει ο κόσμος γι' αυτήν και τα προϊόντα μας. Το θέμα που δημιουργείται είναι αν μπορεί να βοηθήσει μικρές επιχειρήσεις που το έχουν πραγματικά ανάγκη όπως επίσης αν μπορούν να διαθέσουν τόσα χρήματα για να αγοράσουν ένα αρκετά ακριβό αλλά κατά τα άλλα πολύ ισχυρό εργαλείο.

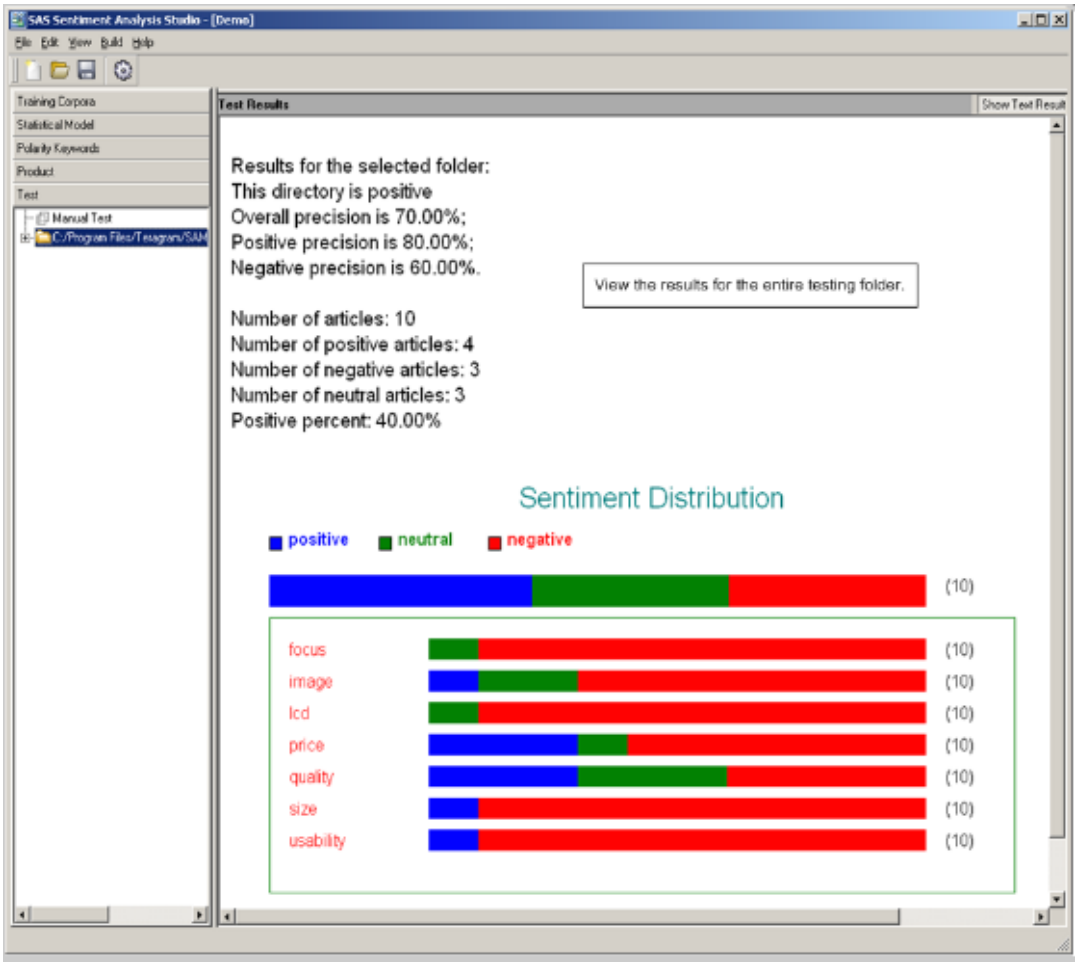
3.5.2 SAS Analytics

Το SAS Analytics είναι κι αυτό ένα εμπορικό λογισμικό το οποίο αναπτύσσεται από την ομώνυμη εταιρεία SAS. Χρησιμοποιείται κυρίως για στατιστικές αναλύσεις, data mining, ανάλυση κειμένου, βελτιστοποίηση, πειραματικό σχεδιασμό και πολλά άλλα. Αυτό όμως που μας ενδιαφέρει περισσότερο είναι ένα συστατικό του, το SAS Sentiment Analysis.

Όπως μπορούμε να καταλάβουμε το SAS Sentiment Analysis ασχολείται με την ανάλυση συναισθήματος. Συλλέγει ψηφιακό περιεχόμενο από διάφορες σελίδες και σελίδες κοινωνικής δικτύωσης κι έπειτα χρησιμοποιεί πολύ δυνατές στατιστικές τεχνικές και γλωσσολογικούς κανόνες για να εξάγει το συναίσθημα που βρίσκεται στα προς εξέταση κείμενα, παρέχοντας περιλήψεις, αναγνωρίζει τις τάσεις και δημιουργεί γραφικές αναφορές, οι οποίες περιγράφουν το συναίσθημα των καταναλωτών, των πελατών και των ανταγωνιστών σε πραγματικό χρόνο.

Μερικά από τα πλεονεκτήματα του είναι ότι ελέγχει τις αλλαγές στην πάροδο του χρόνου, αναγνωρίζει τις περιοχές όπου υπάρχει βελτίωση και μετράει την αποτελεσματικότητα των στρατηγικών που έχουν χρησιμοποιηθεί. Αναγνωρίζει που συζητείται το προς εξέταση θέμα και καθορίζει νέους στόχους και το κυριότερο απ’ όλα κατανοεί το συναίσθημα των καταναλωτών, των συνεργατών, των προμηθευτών και των υπαλλήλων.

Το SAS Sentiment Analysis είναι ένα πολύ ικανό εργαλείο που κάθε αναλυτής θα ήθελε να δουλέψει μαζί του και κάθε επιχείρηση θα ήθελε να το έχει «με το μέρος της», παρόλα αυτά όμως δεν πρέπει να ξεχνάμε το μεγάλο κόστος για την απόκτησή του κάτι που το κάνει απλησίαστο για τις μικρές και τις μεσαίου βεληνεκούς επιχειρήσεις.



Εικόνα 5 Το γραφικό περιβάλλον του SAS Sentiment Analysis

3.6 Οι Μηχανές του Twitter

Αναλύσαμε μερικά αρκετά περίπλοκα συστήματα, παραπάνω, τα οποία χρησιμοποιούνται ως επί το πλείστον από επαγγελματίες. Τι γίνεται όμως με τους υπόλοιπους χρήστες του internet που έχουν μόνο τις βασικές γνώσεις ηλεκτρονικού υπολογιστή; Τη λύση σε αυτό το πρόβλημα δίνει το site κοινωνικής δικτύωσης *Twitter* με μια ομάδα μηχανών ανάλυσης συναισθήματος τα *twitter search, twitter sentiment, tweetfeel, twendz* και *twitrrart.*

3.6.1 Twitter Search

Το *twitter search* είναι ένα υποτυπώδες εργαλείο για ανάλυση συναισθημάτων μπορεί να σας βοηθήσει να δείτε τι σκέφτονται οι άλλοι για εσάς. Περιέχει μια μηχανή αναζήτησης στην οποία γράφετε το όνομά σας μαζί με ένα χαμογελαστό ή λυπημένο

προσωπάκι ή ένα ερωτηματικό (☺, ☹, ?)και αυτό σας επιστρέψει διάφορες εγγραφές που το περιέχουν. Έτσι μπορείτε να καταλάβετε αν τα σχόλια είναι αρνητικά, θετικά ή ουδέτερα. Στην ουσία περιέχει μόνο τις βασικές λειτουργίες, αλλά αυτό μπορεί να το κάνει αρκετά χρήσιμο για τις μικρές επιχειρήσεις. Τη μηχανή αυτή τη βρίσκουμε στο http://search.twitter.com/

Εικόνα 6 Η αρχική σελίδα του twitter search

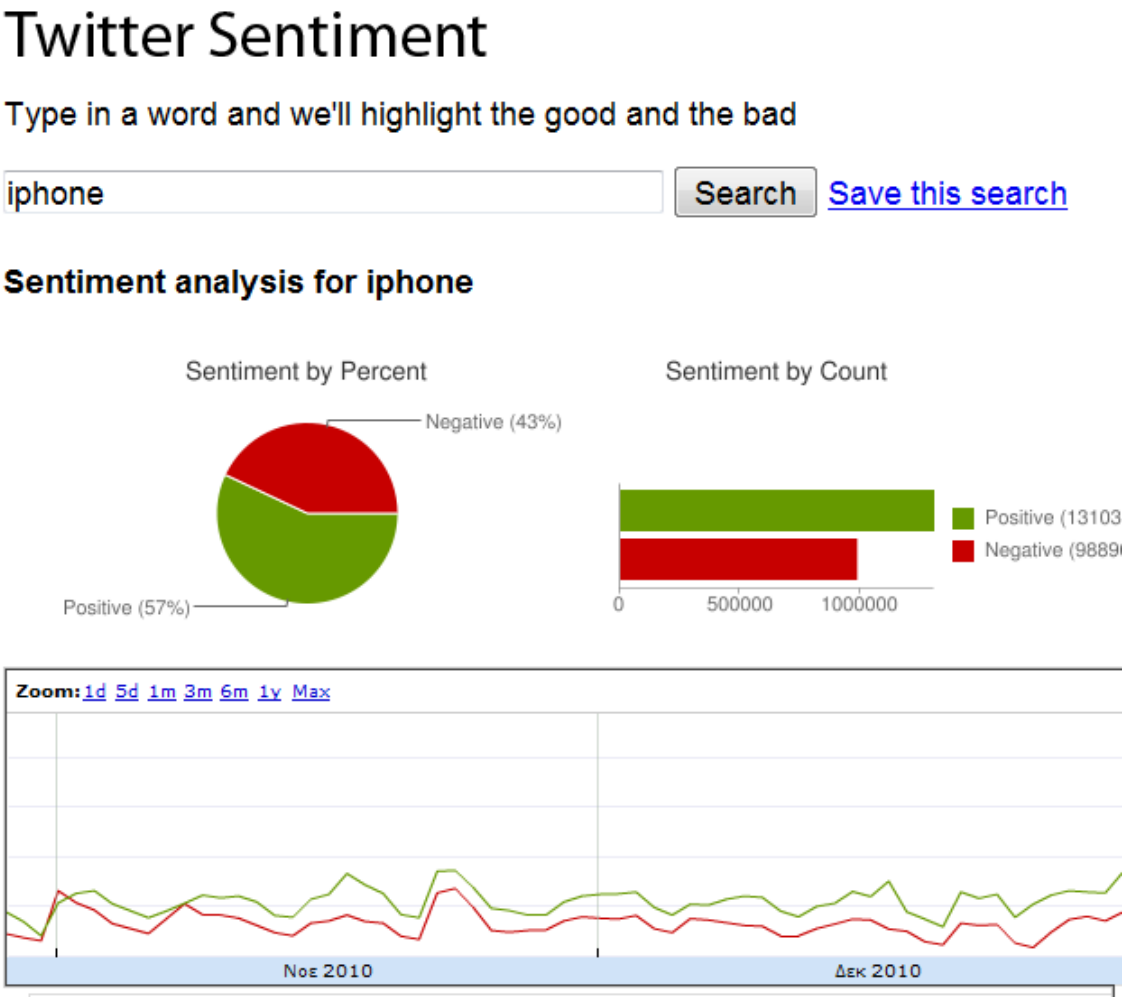
3.6.2 Twitter Sentiment

Πρόκειται για ένα πολύ ισχυρό εργαλείο το οποίο αναπαριστά με διάφορους τρόπους το πώς αισθάνονται οι άλλοι για ένα θέμα. Δημιουργήθηκε στο Πανεπιστήμιο του Stanford από τους Alec Go, Richa Bhayani, και Lei Huang και χρησιμοποιεί αλγόριθμους μηχανικής μάθησης και πιο συγκεκριμένα τον αλγόριθμο της μέγιστης εντροπίας.

Αυτό που πρέπει να κάνουμε είναι να εισάγουμε τον προς αναζήτηση όρο και το εργαλείο θα διαχωρίσει τις θετικές και τις αρνητικές αναφορές που υπάρχουν. Το αποτέλεσμα μπορούμε να το δούμε είτε με το διάγραμμα πίτας που δείχνει τα αποτελέσματα σε ποσοστά, είτε με το γράφημα ραβδώσεων που δείχνει τον αναλυτικό αριθμό των αρνητικών και των θετικών αποτελεσμάτων. Επίσης εμφανίζει και διάγραμμα με την αντίδραση των χρηστών στο συγκεκριμένο θέμα κατά την πάροδο του χρόνου από προηγούμενους μήνες μέχρι και τη στιγμή που κοιτάτε, έτσι μπορούμε να δούμε τις απόψεις του κόσμου για μια συγκεκριμένη ημερομηνία. Τέλος, εμφανίζει στο κάτω μέρος αναλυτικά τις απόψεις των χρηστών οι οποίες έχουν εξεταστεί και

έχουν σημειωθεί με πράσινο φόντο οι θετικές, με κόκκινο οι αρνητικές και με λευκό οι ουδέτερες.

Το twitter sentiment μπορούμε να το βρούμε στη διεύθυνση http://twittersentiment.appspot.com/ .Είναι πραγματικά εντυπωσιακό το τι μπορεί να κάνει. Κι αν δεν εμπιστευόμαστε τα αποτελέσματα μπορούμε να εξετάσουμε τις εγγραφές στο κάτω μέρος και να δούμε αν έχει κάνει λάθος και που.



Εικόνα 7 Μια απλή εικόνα του τι μπορεί να κάνει το twitter sentiment

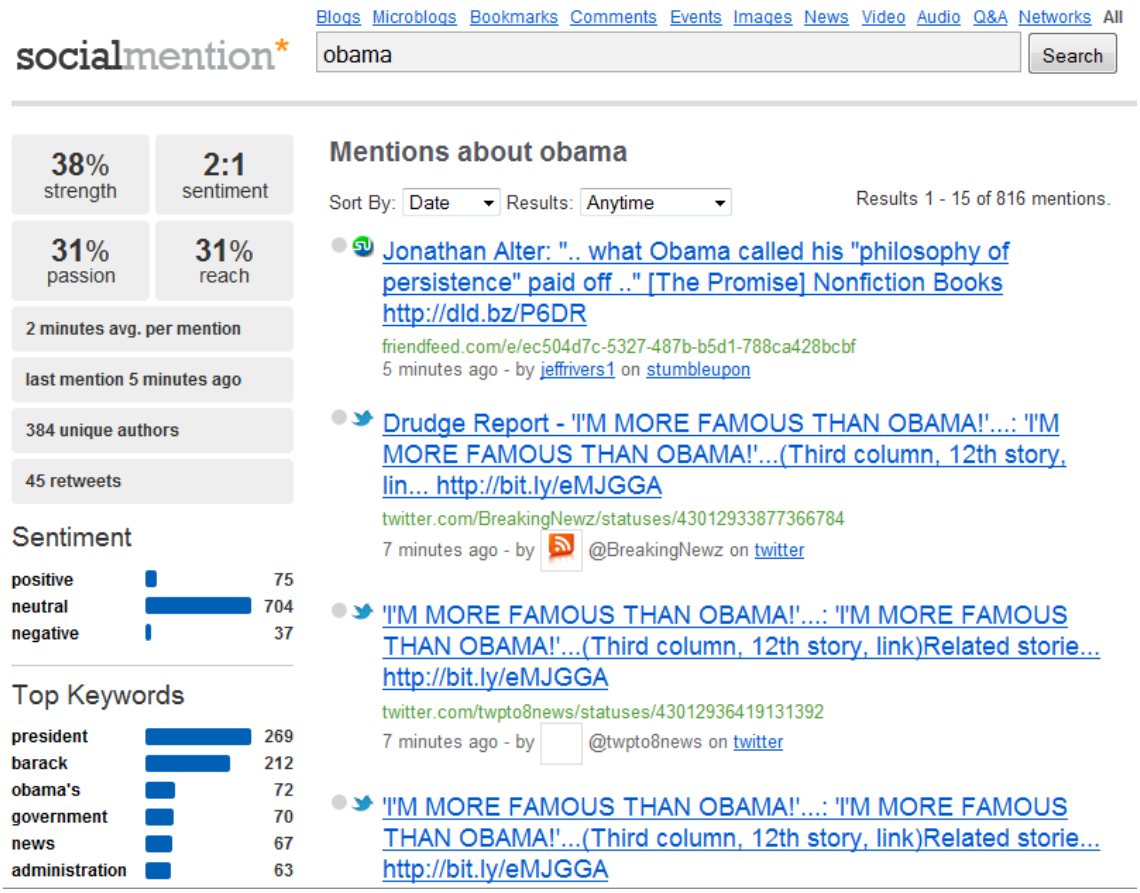
3.6.3 Social Mention

Το social mention είναι το πιο διαδεδομένο σύστημα ανάλυσης συναισθημάτων επειδή είναι χτισμένο πάνω σε ένα εργαλείο εντοπισμού συναισθήματος που ήδη χρησιμοποιούν όλοι οι χρήστες. Δίνει μια αξιόλογη ιδέα του πως συγκεκριμένοι όροι χρησιμοποιούνται απ' τα κοινωνικά δίκτυα.

Οι ενδείξεις που χρησιμοποιεί είναι το strength, δηλαδή η πιθανότητα το προς αναζήτηση αντικείμενο να συζητείται στις σελίδες κοινωνικής δικτύωσης, sentiment, τα

ποσοστά των θετικών και των αρνητικών «χτυπημάτων» (χρησιμοποιώντας τη γλώσσα του twitter), passion, ένα ποσοστό που υποδηλώνει κατά πόσο οι άνθρωποι που συζητούν για το συγκεκριμένο αντικείμενο θα συνεχίσουν να το κάνουν και reach που είναι το μέγεθος του βαθμού επιρροής. Το προς αναζήτηση αντικείμενο μπορεί να είναι ένα προϊόν, μία εταιρεία, κάποιος διάσημος τραγουδιστής ή και ο ίδιος μας ο εαυτός (πολύ θα θέλαμε να ξέρουμε τι συζητούν οι άλλοι πίσω από την πλάτη μας!).

Μπορούμε επίσης να δούμε τα hashtags, που είναι οι πιο συχνοί όροι που σχετίζονται με αυτόν που αναζητούμε, τους χρήστες που σας ανέφεραν περισσότερο, λέξεις κλειδιά που χρησιμοποιούνται συχνά και πολλές άλλες πληροφορίες. Το γεγονός όμως που το κάνει τόσο πολύτιμο είναι ότι η αναζήτηση περιλαμβάνει πολλά διαφορετικά κοινωνικά δίκτυα όπως το facebook, το twitter, το myspace, το yahoo, το google blog, το youtube…Η σελίδα στην οποία μπορούμε να το βρούμε είναι η http://www.socialmention.com/



Εικόνα 8 Εδώ αναζητούμε την άποψη των χρηστών για τον Obama

3.6.4 Twendz

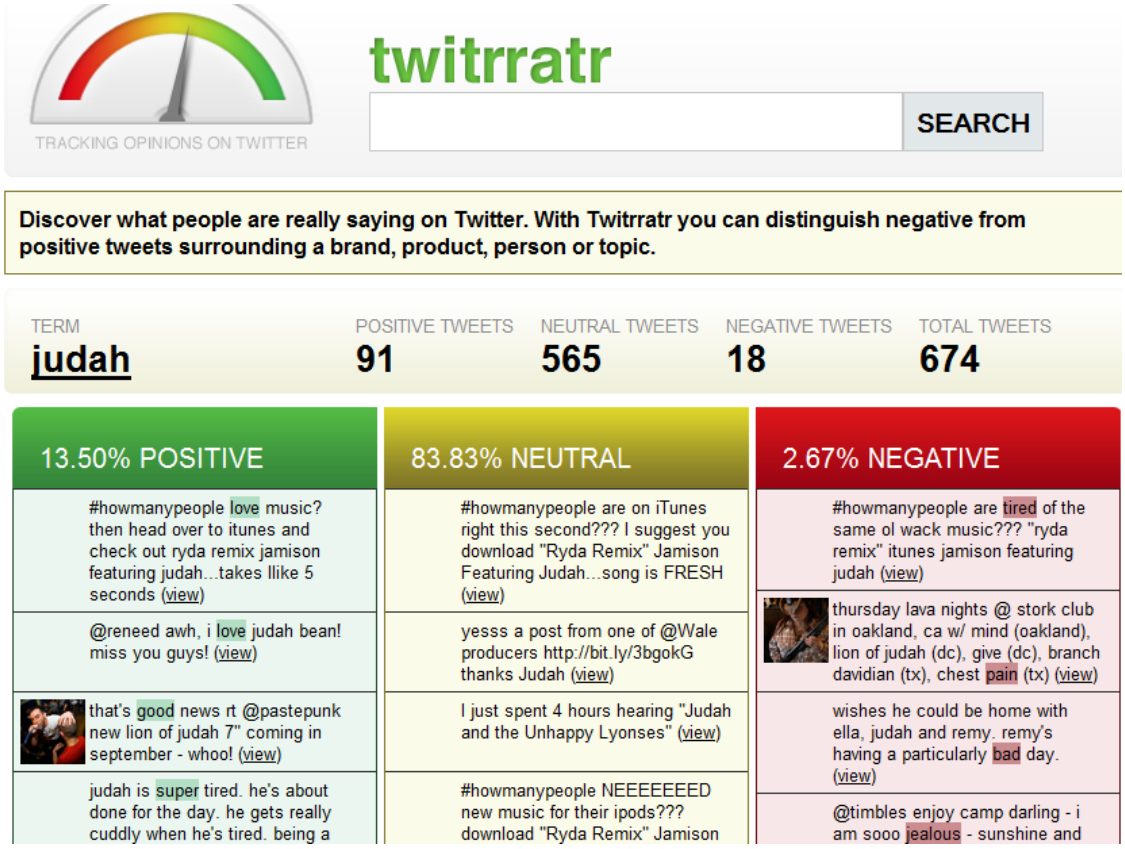
Αν και το twendz δεν έχει ολοκληρωθεί ακόμα και γι' αυτό ορισμένες φορές υπολειτουργεί, αυτό που έχει σημασία είναι η ανάλυση σε πολλά επίπεδα που κάνει. Χρησιμοποιεί έναν συνδυασμό από λέξεις κλειδιά και σύμβολα για να συγκρίνει και να διασταυρώσει απόψεις ούτως ώστε να μαντέψει, στην ουσία, το συναίσθημα των δημοσιεύσεων. Όχι μόνο προσπαθεί να κατατάξει το σύνολο των δημοσιεύσεων σε θετικές αρνητικές ή ουδέτερες αλλά επιλέγει λέξεις κλειδιά που χρησιμοποιούνται σε μεγάλο βαθμό και βρίσκει το συναίσθημα σύμφωνα με αυτές. Μπορεί να βρίσκεται ακόμα σε αρχικό στάδιο παρόλα αυτά, όμως, δίνει αρκετά ενδιαφέροντα αποτελέσματα και έχει και πιο ευχάριστο interface από τα υπόλοιπα. Το twendz υπάρχει στη σελίδα http://twendz.waggeneredstrom.com/

Εικόνα 9 Το twendz δεν στέκεται μόνο στην αναζήτηση του συναισθήματος του συγκεκριμένου όρου αλλά ψάχνει και το συναίσθημα των λέξεων κλειδιών

3.6.5 Twitrratr

Το Twitrratr είναι μια σελίδα αφιερωμένη στην εύρεση απόψεων στο twitter. Περιλαμβάνει μια λίστα από θετικές λέξεις κλειδιά και μια λίστα από αρνητικές λέξεις κλειδιά οι οποίες συγκρίνονται με όποιον όρο αναζητήσετε. Βάση αυτού, δημιουργεί τρεις στήλες για τις αρνητικές, τις θετικές και τις ουδέτερες δημοσιεύσεις. Είναι ένας διαφορετικός τρόπος εμφάνισης συναισθήματος καθώς τώρα πια οι δημοσιεύσεις είναι

ομαδοποιημένες σύμφωνα με στην κατηγορία στην οποία ανήκουν. Το twitrratr μπορούμε να το βρούμε σε αυτή την τοποθεσία http://twitrratr.com/



Εικόνα 10 Μια αναζήτηση του χρήστη judah μας δίνει τα παρακάτω αποτελέσματα

3.6.6 TweetFeel

Το TweetFeel είναι μια απλή και αρκετά εύχρηστη μηχανή αναζήτησης συναισθήματος. Όταν βρει τις δημοσιεύσεις για τον προς αναζήτηση όρο, βρίσκει το ποσοστό θετικότητας ή αρνητικότητας και το εμφανίζει στο πάνω μέρος της οθόνης. Επίσης, εμφανίζει τις θετικές δημοσιεύσεις κάτω από ένα χαμογελαστό πρόσωπο και τις αρνητικές κάτω από ένα λυπημένο. Το σημαντικό είναι ότι χρησιμοποιεί το πράσινο χρώμα όπου υπάρχει θετικό συναίσθημα και το κόκκινο όπου υπάρχει αρνητικό, έτσι μπορούμε να ξέρουμε το αποτέλεσμα πριν ακόμα δούμε τους αριθμούς. Το TweetFeel μπορούμε να το βρούμε στη διεύθυνση http://www.tweetfeel.com/

Εικόνα 11 Ένα πολύ ευχάριστο περιβάλλον για την αναζήτηση συναισθήματος

4 Η Ανάλυση Συναισθήματος στην Πράξη

Σ' αυτό το κεφάλαιο θα κάνουμε μια προσπάθεια να κατανοήσουμε την ανάλυση κειμένου μέσα από ένα πείραμα. Με τη βοήθεια του RapidMiner που αναλύσαμε παραπάνω θα πάρουμε ένα σύνολο δεδομένων (datasets) και θα το κατατάξουμε σε θετικό ή αρνητικό με την ανάλυση συναισθήματος.

4.1 Σύνολα Δεδομένων

Τα *σύνολα δεδομένων* ή αλλιώς datasets, είναι μια συλλογή από δεδομένα που συνήθως αναπαριστούνται με τη μορφή πίνακα. Κάθε στήλη αναπαριστά μια συγκεκριμένη μεταβλητή, ενώ κάθε γραμμή αντιστοιχεί σε ένα μέλος του dataset που εξετάζουμε. Ένα dataset μπορεί να έχει διάφορα χαρακτηριστικά τα οποία καθορίζουν την δομή και τις ιδιότητές του. Αυτά περιλαμβάνουν τον αριθμό των μεταβλητών καθώς και τις στατιστικές μετρήσεις που τους έχουν εφαρμοστεί.

Οι μεταβλητές μπορεί να έχουν τη μορφή αριθμών, όπως για παράδειγμα το ύψος σε εκατοστά, ή κειμένου, όπως η εθνικότητα. Σε γενικές γραμμές, οι μεταβλητές μπορεί να είναι οτιδήποτε μπορεί να μετρηθεί, πάντως σε κάθε dataset πρέπει να είναι όλες το ίδιο είδος. Παρόλα αυτά, μερικές φορές μπορεί να υπάρξουν οι λεγόμενες χαμένες μεταβλητές (missing values) τις οποίες θα πρέπει με κάποιο τρόπο να διαχειριστούμε και να τις εμφανίσουμε.

Στη στατιστική ανάλυση τα datasets προκύπτουν συνήθως από την δειγματοληψία και κάθε γραμμή αντιστοιχεί σε ένα στοιχείο του πληθυσμού, ενώ υπάρχουν ειδικοί αλγόριθμοι οι οποίοι μπορούν να δημιουργήσουν datasets πολλές φορές για την αξιολόγηση ειδικών λογισμικών.

Στο πείραμα που ακολουθεί θα χρησιμοποιήσουμε έτοιμα datasets, polarity datasets v2.0 που βρίσκονται στη σελίδα του Bo Pang και της Lillian Lee http://www.cs.cornell.edu/People/pabo/movie-review-data/ . Το συγκεκριμένο dataset προέρχεται απ' τη σελίδα imdb,η οποία είναι στην ουσία μια βάση δεδομένων για

ταινίες, ηθοποιούς κ.τ.λ. Είναι γραμμένο στα αγγλικά και περιέχει 1000 αρνητικές και 1000 θετικές κριτικές για ταινίες κατηγοριοποιημένες από τους συγγραφείς. Τα συγκεκριμένα δεδομένα είναι σε μορφή κειμένου (txt) και έχουν δημιουργηθεί ειδικά γι' αυτόν τον σκοπό, δηλαδή για την ανάλυση της απόδοσης τεχνικών ανάλυσης συναισθήματος. Κάθε γραμμή μιας κριτικής αντιστοιχεί σε μία πρόταση όπως έχει οριστεί από τον Adwait Ratnaparkhi στον ανιχνευτή ορίου γραμμής MXTERMINATOR.

4.2 Επεξήγηση Πειράματος

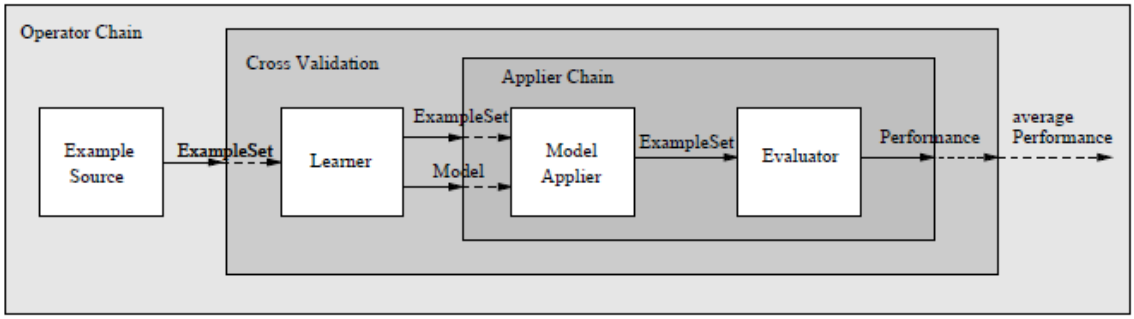
Σκοπός του πειράματος αυτού είναι να εφαρμόσουμε πάνω στα δεδομένα μια σειρά από αλγορίθμους, που πρώτα θα τα επεξεργαστούν για να τα φέρουν σε μία επιθυμητή μορφή, έπειτα θα τα κατηγοριοποιήσουν σε υποκειμενικά και αντικειμενικά, τα υποκειμενικά θα αναλυθούν περεταίρω σε θετικά και αρνητικά και τέλος θα έχουμε το αποτέλεσμα, αν, δηλαδή, στο σύνολό τους είναι θετικά ή αρνητικά. Κατά τη διάρκεια του πειράματος, εφαρμόζουμε ότι έχει ειπωθεί παραπάνω, στην προσπάθειά μας να κάνουμε μια όσο το δυνατόν πιο λεπτομερή και ακριβής ανάλυση και να δώσουμε στον αναγνώστη να καταλάβει και εμπράκτως όλες τις πτυχές της ανάλυσης συναισθήματος.

Το εργαλείο που θα χρησιμοποιηθεί εδώ είναι το RapidMiner καθώς είναι απ' τα πιο ολοκληρωμένα συστήματα data mining που υπάρχουν και κυρίως μπορούμε να το αποκτήσουμε και να το χρησιμοποιήσουμε χωρίς κόστος. Περιέχει πρόσθετα για αρκετά εξειδικευμένα θέματα όπως είναι η ανάλυση κειμένου και η ανάλυση συνεχόμενων δεδομένων. Επίσης οι δυνατότητές του όσο αφορά την ανάλυση κειμένου και την ανάλυση συναισθήματος είναι τεράστιες. Παρέχει αμέτρητες επιλογές από χειριστές και επιπλέον μια βιβλιοθήκη για data mining από το εργαλείο WEKA. Ένας άλλος λόγος που επέλεξα το συγκεκριμένο εργαλείο είναι γιατί, παρόλη τη δυσκολία του, έχει ένα πολύ καλό περιβάλλον βοήθειας που εξηγεί αναλυτικά τι είναι το κάθε αντικείμενο που χρησιμοποιούμε, έχει ένα πολύ εξυπηρετικό forum συζητήσεων και σε σύγκριση με άλλα συστήματα είναι πιο ξεκάθαρο.

Το πείραμα θα επεξηγείται αναλυτικά βήμα βήμα, ούτως ώστε να μπορεί ο αναγνώστης να καταλαβαίνει κάθε κίνηση και θα συνοδεύεται από διάφορα screen shots και γραφήματα για να βλέπει τι ακριβώς γίνεται και στην πράξη.

4.3 Το Πείραμα για το RapidMiner

Όπως αναφέραμε και παραπάνω το RapidMiner χρησιμοποιεί operators. Ένας operator είναι μια ακολουθία ή ένας συνδυασμός από επεξεργασμένα δεδομένα και μεθόδους μηχανικής μάθησης. Μια αλυσίδα από τέτοιους operators ονομάζεται operator chain (αλυσίδα χειριστών).Μια operator chain είναι ένας καινούριος operator έχει κάθε ιδιότητα και κάθε λειτουργία που έχουν και αυτοί από τους οποίους απαρτίζεται. Οι operators οι οποίοι εσωκλείουν άλλους operators ή operator chains ονομάζονται wrappers. Στην παρακάτω εικόνα μπορούμε να δούμε ένα παράδειγμα operator chain.



Εικόνα 12 Ένα απλό παράδειγμα αλυσίδας χειριστών

Το RapidMiner έχει πάρα πολλούς χειριστές μερικοί απ' τους οποίους είναι οι Read CVS, Read Excel, Read Access και άλλοι για εισαγωγή δεδομένων,Write CVS, Write Excel, Write Access κ.τ.λ. για εξαγωγή δεδομένων,Tokenize, Stem (Snowball), Read Document και Write Document για ανάλυση κειμένου κ.τ.λ.

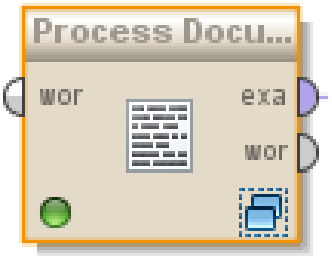
Ξεκινώντας, Θα πρέπει να κατεβάσουμε τους ειδικούς operators για ανάλυση κειμένου και να τους αποθηκεύσουμε στο lib\plugins. Τα πρόσθετα αυτά περιέχουν θέματα ειδικά σχεδιασμένα για να βοηθήσουν στην προετοιμασία εγγράφων, σε μορφή κειμένου, για κάθε είδους μέθοδο επεξεργασίας δεδομένων όπως το tokenization, η απομάκρυνση των ασήμαντων λέξεων και το stemming. Μ' αυτόν τον τρόπο μπορούμε να διαβάζουμε και να επεξεργαζόμαστε ολόκληρα κείμενα και να έχουμε όλες τις λειτουργίες της ανάλυσης κειμένου, συμπεριλαμβανομένου και της ανάλυσης συναισθήματος. Ένας δεύτερος τρόπος να το κάνουμε αυτό είναι να κάνουμε update στο πρόγραμμα και να κατεβάσουμε τις ενημερώσεις για ανάλυση κειμένου.

4.3.1 Εισαγωγή Δεδομένων

Όπως είπαμε παραπάνω, χρησιμοποιούμε έτοιμα δεδομένα τα οποία είναι σε μορφή κειμένου (txt) και περιέχουν κριτικές ταινιών. Τα δεδομένα αυτά βρίσκονται σε δύο

φακέλους, τον φάκελο neg και τον φάκελο pos. Για να εισάγουμε δεδομένα τα οποία βρίσκονται σε φάκελο χρησιμοποιούμε τον Process Documents from Files operator. Ο συγκεκριμένος χειριστής δημιουργεί διανύσματα λέξεων από μια συλλογή κειμένων τα οποία είναι αποθηκευμένα σε πολλαπλά αρχεία. Έπειτα εισάγουμε τους φακέλους που θέλουμε να επεξεργαστούμε μέσα στον χειριστή ενώ παράλληλα τους δίνουμε και ένα όνομα (μια ετικέτα). Στην περίπτωσή μας είναι αυτονόητο να τους δώσουμε τα ονόματα neg και pos ούτως ώστε να α ταιριάξουμε με τους φακέλους από τους οποίους προήλθαν.

Παρακάτω βλέπουμε μια εικόνα του χειριστή. Παρατηρούμε ότι από τη αριστερή πλευρά μπορούμε να εισάγουμε τα δεδομένα, αν παραδείγματος χάρη χρησιμοποιήσουμε μια βάση δεδομένων (σ' αυτή την περίπτωση τα διαβάζουμε με τον Read Database operator και στη συνέχεια τον συνδέουμε με τον Process Documents from Files). Από τη δεξιά πλευρά επιλέγουμε τον τρόπο με τον οποίο θα εξάγουμε τα δεδομένα. Αν θέλουμε να είναι σε μορφή κειμένου επιλέγουμε το exa, ενώ αν θέλουμε να εξάγουμε τις λέξεις μετά την επεξεργασία επιλέγουμε το wor.

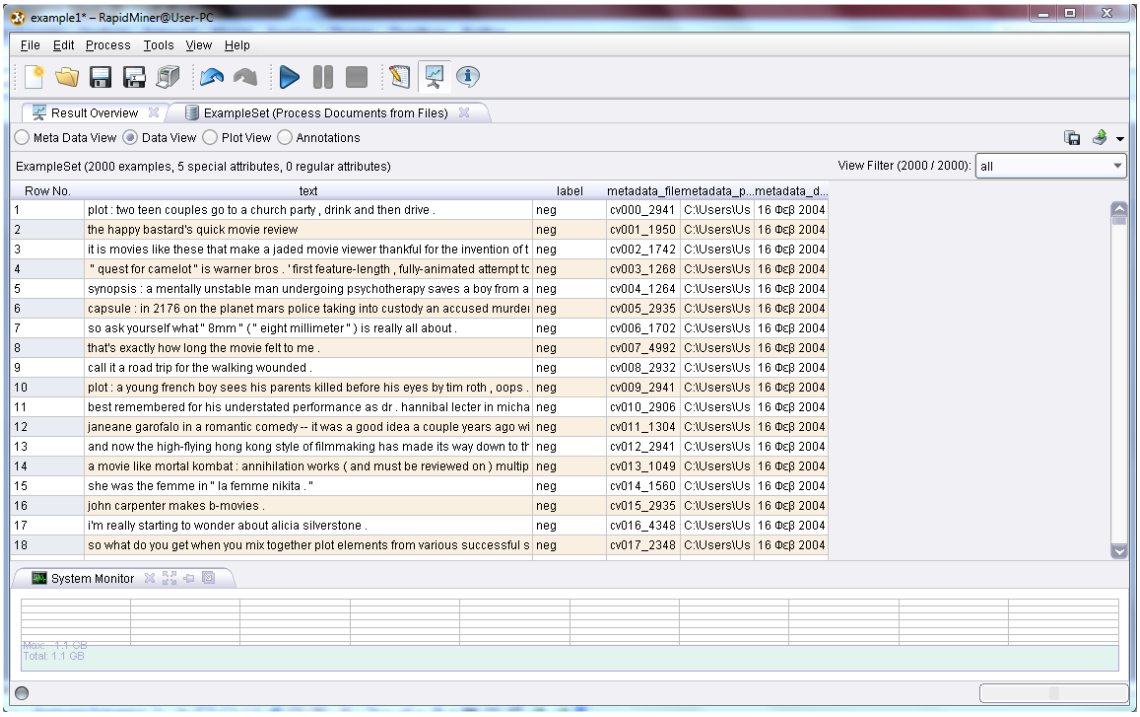


Εικόνα 13 Ο Process Documents from Files operator είναι υπεύθυνος για την επεξεργασία δεδομένων που βρίσκονται μέσα σε φακέλους.

Υπάρχουν κι άλλοι τρόποι για να εισάγουμε δεδομένα στο RapidMiner όπως για παράδειγμα, με τη βοήθεια του Create Document operator με τον οποίο μπορούμε να δημιουργήσουμε εκείνη τη στιγμή κείμενο, τον Read Document operator με τον οποίο εισάγουμε δεδομένα από κάποιο αρχείο, ή τον Read Excel για να εισάγουμε αποκλειστικά δεδομένα τύπου excel (αντιστοίχως άλλους χειριστές για κάθε τύπου αρχεία). Στην δική μας περίπτωση ο πιο κατάλληλος φαίνεται να είναι ο Process Documents from Files operator.

Ας δούμε όμως πως παρουσιάζονται τα δεδομένα στο RapidMiner. Καταρχάς, το RapidMiner δημιουργεί έναν πίνακα για τα δεδομένα που στην πρώτη στήλη έχει έναν αύξοντα αριθμό για την αρίθμηση τους, στη δεύτερη τα δεδομένα, στην τρίτη το όνομα της ετικέτας της λίστας στην οποία ανήκουν, στην τέταρτη τα μεταδεδομένα, στην

πέμπτη την τοποθεσία όπου έχουν αποθηκευτεί τα μεταδεδομένα και στην τελευταία στήλη, πληροφορίες γι' αυτά.



Εικόνα 14 Εισαγωγή δεδομένων στο RapidMiner

4.3.2 Επεξεργασία Δεδομένων

Στη συνέχεια θα προσπαθήσουμε να επεξεργαστούμε τα δεδομένα χρησιμοποιώντας κάποιες τεχνικές όπως tokenization, απομάκρυνση των κοινών λέξεων με μικρή σημασία (stop words) και stemming. Πριν όμως προχωρήσουμε θα πρέπει να εξηγήσουμε αυτές τις τεχνικές.

Η τεχνική tokenization βρίσκει τις λέξεις του εγγράφου, δηλαδή το χωρίζει σε λέξεις και έπειτα απομακρύνει τις μικρές που συνήθως δεν έχουν μεγάλη σημασία (δεν δίνουν καμιά ιδιαίτερη ουσία στο κείμενο). Το μέγεθος αυτών των λέξεων το καθορίζουμε εμείς. Στην δική μας περίπτωση καθορίζουμε το φίλτρο να δέχεται από δύο χαρακτήρες και πάνω καθώς ένα διαφορετικό φίλτρο θα μπορούσε να αφαιρέσει σημαντικές λέξεις όπως no, ok.

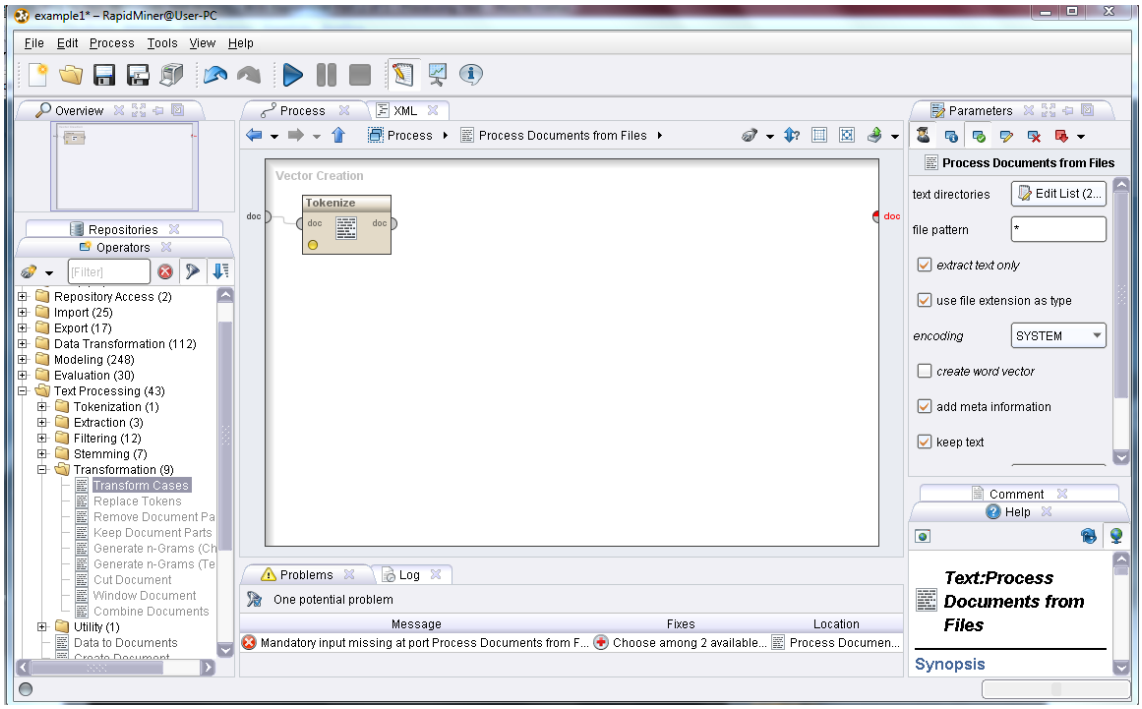
Για την απομάκρυνση των κοινών λέξεων (stop words) συνήθως χρησιμοποιούμε ένα αρχείο με λίστες τις οποίες εμείς έχουμε επιλέξει και θεωρούμε ότι δεν έχουν μεγάλη σημασία. Το RapidMiner όμως χρησιμοποιεί έναν ειδικό χειριστή γι' αυτή τη δουλεία, οποίος περιέχει μια δικιά του λίστα με πιθανά stop words όπως a, about,

above, after, ourselves, out, over, own, same, shan't, she, she'd, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't και άλλες πολλές.

Το stemming είναι μια τεχνική η οποία μειώνει τις λέξεις στην κοινή τους ρίζα. Για παράδειγμα, οι λέξεις experience, experimental, experienced έχουν κοινή ρίζα το exper-. Δουλειά αυτής της μεθόδου, λοιπόν, είναι να βρίσκει αυτές τις λέξεις, να αναγνωρίζει τη ρίζα τους και να κρατάει μόνο αυτό. Υπάρχει μια διαφωνία σχετικά με το αν το stemming και γενικά το φιλτράρισμα των λέξεων μπορούν να αποδώσουν καλά ένα πείραμα, παρόλα αυτά η τεχνική λέξη διάνυσμα είναι πολύ μικρότερη απ' την αρχική και επεξεργάζεται πιο εύκολα και γρήγορα από την αρχική.

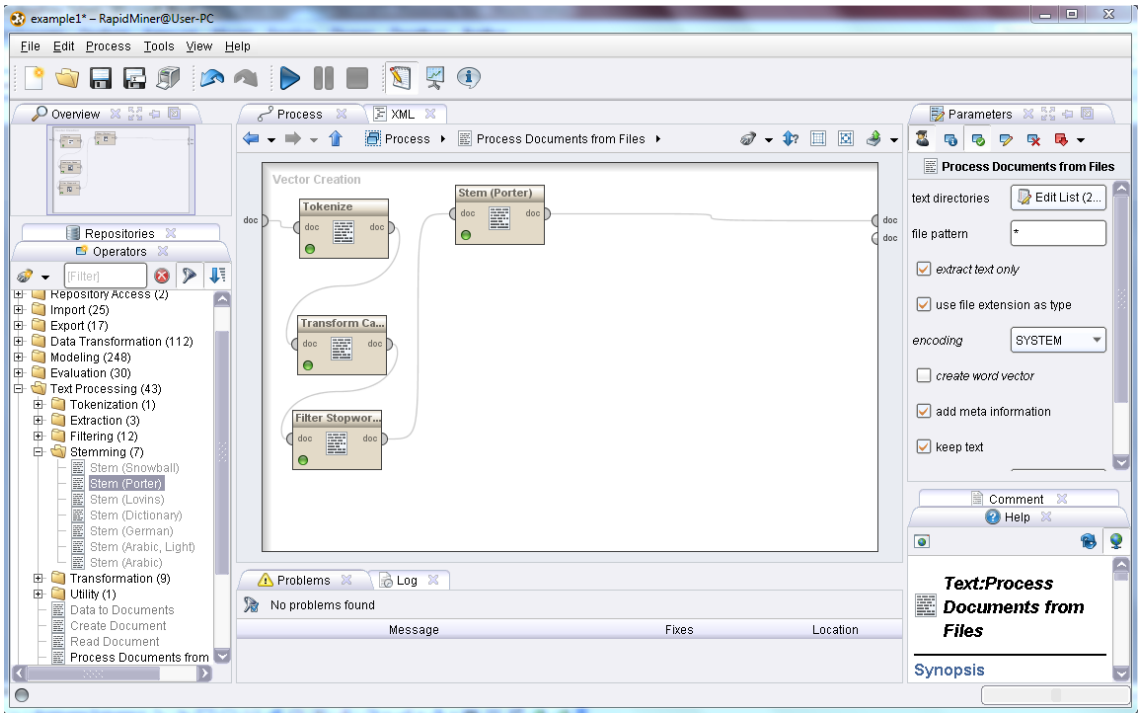
Τέλος έχουμε την n-gram tokenizer τεχνική, η οποία δημιουργεί φράσεις με τις λέξεις οι οποίες προκύπτουν απ' το κείμενο. Ένα 2-gram ή bigram tokenizer δημιουργεί φράσεις που αποτελούνται από δύο λέξεις με κάθε πιθανό συνδυασμό. Τα n-gram έχουν τη δυνατότητα να δίνουν περισσότερες πληροφορίες σχετικά με την πολικότητα, όπως για παράδειγμα οι φράσεις not nice γίνεται not_nice και μπορεί πια να χρησιμοποιηθεί σαν ενιαίο αντικείμενο από τους ταξινομητές.

Στη συνέχεια του πειράματος θα χρειαστεί να βάλουμε μέσα στον Process Documents from Files operator τους κατάλληλους χειριστές για να επεξεργαστούν τα δεδομένα. Ξεκινάμε με τον Tokenize operator ο οποίος θα κάνει tokenization. Αφού πατήσουμε διπλό κλικ πάνω στον Documents from Files operator για να μπούμε στο εσωτερικό του, βρίσκουμε τον Tokenize operator στην βιβλιοθήκη Text Processing και τον σέρνουμε μέσα.



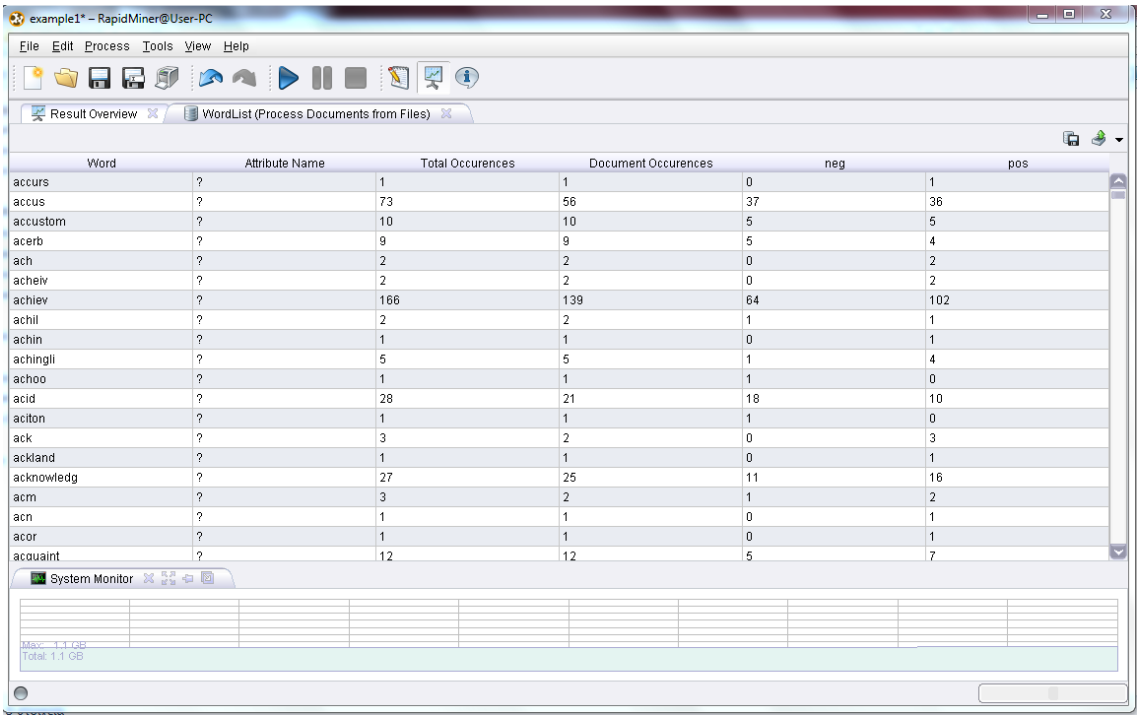
Εικόνα 15 Προσθέτοντας χειριστές

Έπειτα μια καλή ιδέα είναι να χρησιμοποιήσουμε τον Transform Cases operator συνδέοντας τον με τον προηγούμενο ούτως ώστε να μετατραπούν όλα τα κεφαλαία σε μικρά. Μετά θα προσθέσουμε τον Filter Stopwords (English) operator ο οποίος θα φιλτράρει και θα αφαιρέσει όλα τα stop words που υπάρχουν στο κείμενο. Η επόμενή μας δουλειά είναι να κάνουμε stemming. Για τον σκοπό αυτό θα χρησιμοποιήσουμε τον Stemming Porter operator και θα τον συνδέσουμε με τον προηγούμενο. Η εικόνα του RapidMiner με όλους τους χειριστές που έχουμε χρησιμοποιήσει έχει ως εξής:



Εικόνα 16 Χρησιμοποιώντας τους χειριστές Tokenize , Transform Cases , Filter Stopwords, Stemming Porter

Μέχρι αυτή τη στιγμή έχουμε «κομματιάσει» τα δεδομένα σε λέξεις, τις οποίες έχουμε περιορίσει από 2 έως 999 χαρακτήρες, έχουμε κάνει όλα τα γράμματα μικρά, έχουμε αφαιρέσει τις κοινές λέξεις (αν παρατηρήσουμε, λέξεις όπως η and δεν υπάρχουν) και τέλος έχουμε περιορίσει τις λέξη με κοινή ρίζα. Στην παρακάτω εικόνα μπορούμε να δούμε τα αποτελέσματα αυτά βάση ενός πίνακα. Ο πίνακας αυτός στην πρώτη στήλη έχει τις λέξεις των κειμένων μετά την επεξεργασία, στη δεύτερη τα ονόματα των χαρακτηριστικών αν υπάρχουν (στην περίπτωσή μας δεν υπάρχουν). Στην τρίτη στήλη παρουσιάζει το πόσες φορές εμφανίζεται μία λέξη συνολικά, στην τέταρτη στο έγγραφο και στις δύο τελευταίες στήλες βλέπουμε στους κάθε φακέλους.



Word	Attribute Name	Total Occurences	Document Occurences	neg	pos
accurs	?	1	1	0	1
accus	?	73	56	37	36
accustom	?	10	10	5	5
acerb	?	9	9	5	4
ach	?	2	2	0	2
acheiv	?	2	2	0	2
achiev	?	166	139	64	102
achil	?	2	2	1	1
achin	?	1	1	0	1
achingli	?	5	5	1	4
achoo	?	1	1	1	0
acid	?	28	21	18	10
aciton	?	1	1	1	0
ack	?	3	2	0	3
ackland	?	1	1	0	1
acknowledg	?	27	25	11	16
acm	?	3	2	1	2
acn	?	1	1	0	1
acor	?	1	1	0	1
acquaint	?	12	12	5	7

Εικόνα 17 Τα αποτελέσματα της επεξεργασίας.

4.3.3 Δημιουργία Λέξεων Διανυσμάτων και Κανόνες Συσχέτισης

Σε μια περαιτέρω διερεύνηση του θέματος θα μπορούσαμε προσπαθήσουμε να δημιουργήσουμε λέξεις διανύσματα, να έχουμε μια πρώτη επαφή με τη συχνότητα των αντικειμένων, να βρούμε και να εφαρμόσουμε κάποιους κανόνες συσχέτισης. Με αυτό τον τρόπο θα μπορέσουμε να παρατηρήσουμε τη συχνότητα εμφάνισης των λέξεων και τις σχέσεις μεταξύ τους. Έτσι με ένα απλό εγχείρημα text mining θα καταλάβουμε ακόμα περισσότερο τον τρόπο που δουλεύει και η ανάλυση συναισθήματος αλλά και το RapidMiner.

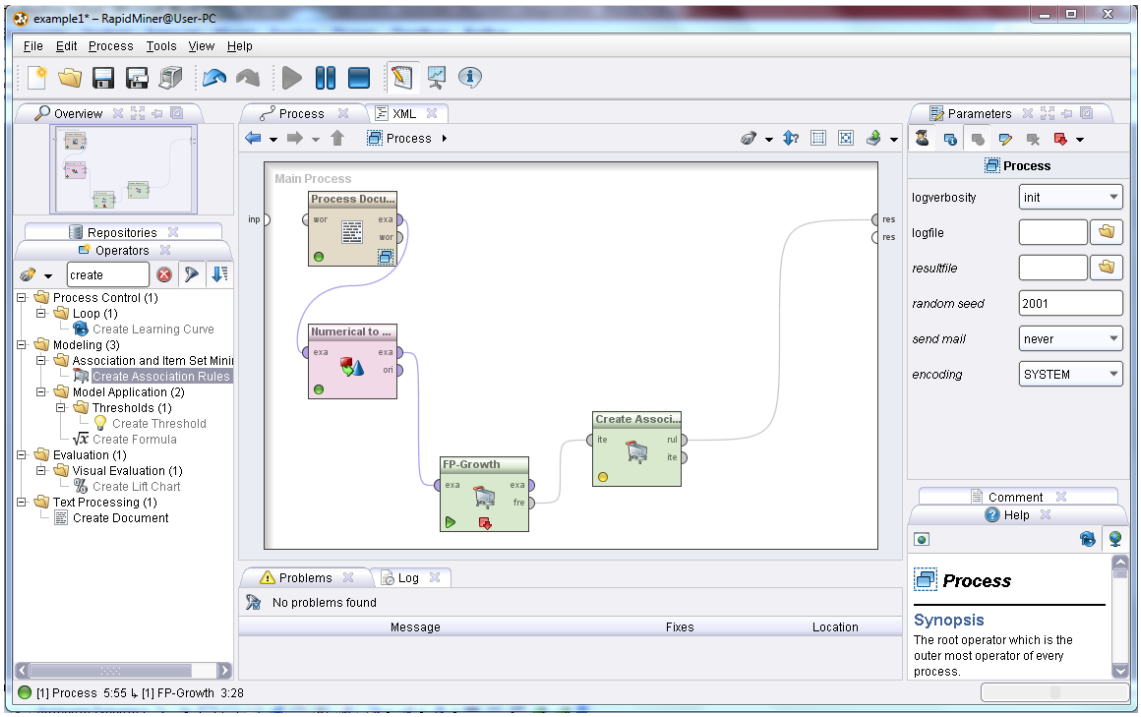
Η δημιουργία λέξεων διανυσμάτων κατά την επεξεργασία των δεδομένων δίνει τη δυνατότητα δημιουργίας μιας καινούριας στήλη στον πίνακά μας η οποία δίνει τις τιμές 0 ή 1 σε κάθε λέξη. Την τιμή 0 παίρνουν οι λέξεις οι οποίες δεν υπάρχουν στο συγκεκριμένο κείμενο ενώ αντίστοιχα την τιμή 1 παίρνουν οι λέξεις τις οποίες συναντούμε έστω και μία φορά.

Σε αυτή τη φάση θα χρησιμοποιήσουμε τον χειριστή Numerical to Binominal operator και θα τον συνδέσουμε με τον Process Documents from Files operator. Ο συγκεκριμένος χειριστής θα μετατρέψει όλες τις αριθμητικές τιμές σε δυαδικές. Δηλαδή, Στον πίνακα που μας δείχνει τα στατιστικά αποτελέσματα ξαφνικά θα εμφανιστούν οι λέξεις false και true. Όπου βλέπουμε false σημαίνει πως η λέξη δεν

εμφανίζεται στο συγκεκριμένο έγγραφο και true εμφανίζεται. Επίσης, μπορούμε να ανατρέξουμε στα μεταδεδομένα, εκεί θα δούμε πόσα false (δηλαδή σε πόσα έγγραφα δεν εμφανίζεται) και πόσα true (σε πόσα έγγραφα εμφανίζεται) έχει.

Μετά από τον Numerical to Binominal operator θα τοποθετήσουμε τον FP-Growth operator. Αυτός ο χειριστής είναι ένας αλγόριθμος ο οποίος μας δείχνει το ποσοστά εμφάνισης κάθε λέξεις αλλά και ποσοστό εμφάνισης περισσότερων λέξεων μαζί. Όταν μια λέξη εμφανίζεται πολλές φορές, ο FP-Growth επιλέγει κι άλλες λέξεις με τις εμφανίζονται συχνά μαζί δείχνει το ποσοστό τους.

Στο τέλος αυτού του σταδίου θα προσπαθήσουμε να δημιουργήσουμε κανόνες συσχέτισης με τον Create Association Rules operator. Ο συγκεκριμένος χειριστής βρίσκει τις σχέσεις και προσπαθεί να τις συνδυάσει με τέτοιο τρόπο ώστε οι δυνατότερες λέξεις (αυτές που εμφανίζονται συχνότερα) να παίρνουν το νόημά τους απ' αυτές με τις οποίες σχετίζονται. Όπως για παράδειγμα η λέξη movie καθορίζεται από τις λέξεις actor και play.



Εικόνα 18 Βρίσκοντας τη συχνότητα των λέξεων και δημιουργώντας κανόνες συσχέτισης.

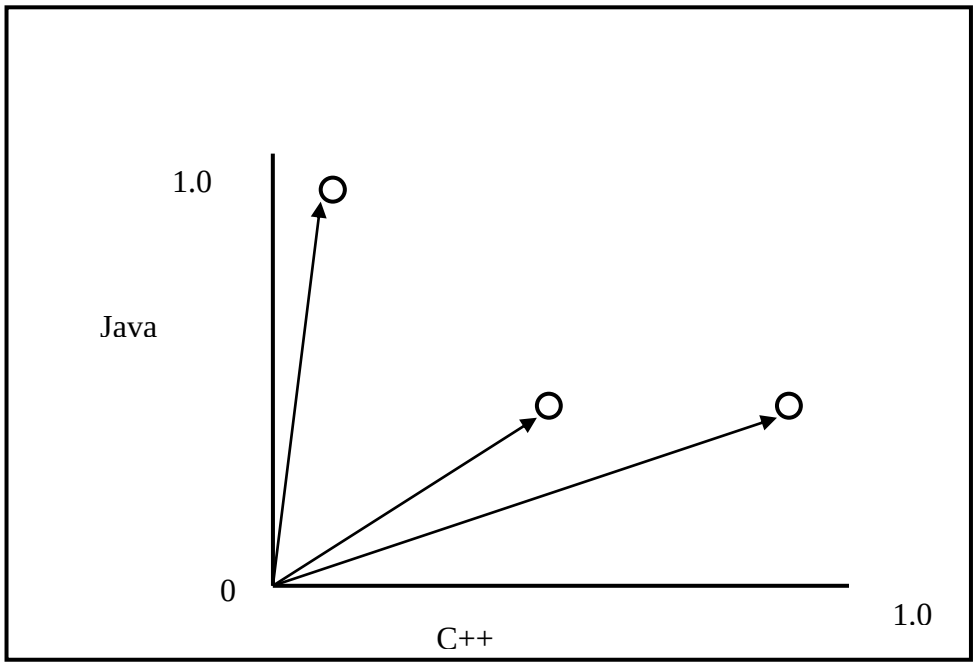
4.3.4 Ομοιότητα μεταξύ των Εγγράφων

Αν και δεν είναι απαραίτητο καλό θα ήταν να εξετάσουμε το TF-IDF (term frequency-inverse document frequency) ανάμεσα στα έγγραφα, να υπολογίσουμε την ομοιότητα μεταξύ των εγγράφων και να ενώσουμε έγγραφα.

Το TF-IDF είναι ένα μέγεθος χρησιμοποιείται για να κάνει στατιστικές μετρήσεις για να ανακαλύψει πόσο σημαντική είναι μια λέξη σε ένα κείμενο, σε ένα έγγραφο ή σε μια συλλογή εγγράφων. Αφού πρώτα ορίσουμε την term frequency του εγγράφου βρίσκουμε την inverse document frequency η οποία είναι ο συνολικός αριθμός των εγγράφων δια τον αριθμό των εγγράφων που περιέχουν την λέξη συν ένα και τα πολλαπλασιάζουμε (tf*idf). Έτσι, αν σε ένα κείμενο περιέχεται συχνά μια λέξη τότε θα έχει μεγάλο TF-IDF αφού και η συχνότητα της λέξης αυξάνεται. Αν όμως η λέξη περιέχεται σε πολλά έγγραφα το TF-IDF θα μειωθεί αφού η συχνότητα ενός όρου και ο αριθμός των εγγράφων στα οποία εμφανίζεται είναι αντιστρόφως ανάλογα ποσά.

Στο RapidMiner το TF-IDF μπορούμε να το εξετάσουμε, αν από τις ιδιότητες του Process Document from Files operator επιλέξουμε απ' το vector creation το TF-IDF. Παρόμοια εξετάζουμε και το term frequency.

Ας δούμε τώρα πως μπορούμε να εκφράσουμε την ομοιότητα κάποιων εγγράφων. Αν, για παράδειγμα, θέλουμε να συγκρίνουμε κάποια έγγραφα με βάση τις λέξεις Java και C++ θα δημιουργήσουμε ένα γράφημα όπου θα αναπαριστούμε τις λέξεις στους δύο κάθετους άξονες και θα μετράμε τη συχνότητα εμφάνισης τους.



Οι κύκλοι στο παραπάνω σχήμα αντιστοιχούν σε έγγραφα. Παρατηρούμε ότι στο πρώτο έγγραφο εμφανίζεται περισσότερες φορές η λέξη Java και ελάχιστες φορές η C++. Στο δεύτερο έγγραφο εμφανίζονται σχεδόν το ίδιο ενώ στο τρίτο έγγραφο βλέπουμε ότι η λέξη C++ εμφανίζεται πιο πολύ ενώ η λέξη Java έχει μέτρια συχνότητα εμφάνισης. Αν το εξετάσουμε συνολικά ο «νικητής» θα βρεθεί από τις λεπτομέρειες.

Παρόλα αυτά η ομοιότητα των εγγράφων θα βρεθεί από τις γωνίες που σχηματίζουν τα βέλη. Όπως βλέπουμε στο σχήμα όσο πιο μικρή είναι η γωνία τόσο πιο όμοια είναι τα έγγραφα μεταξύ τους.

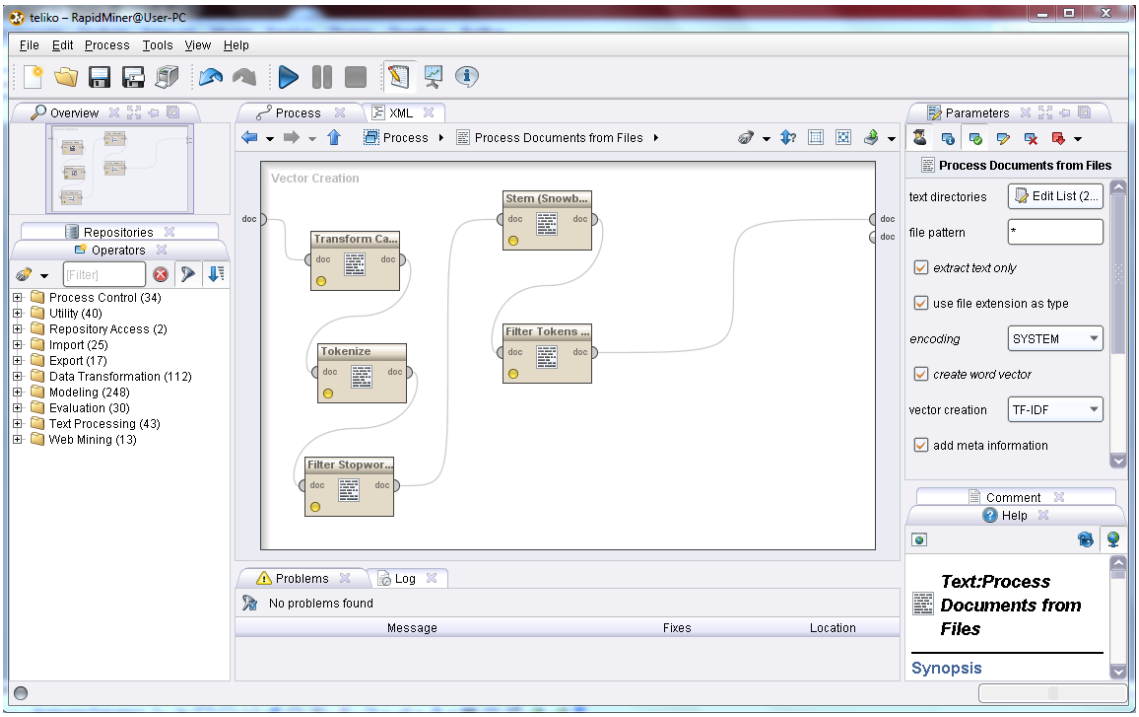
Για να διαπιστώσουμε αν δύο έγγραφα είναι όμοια μεταξύ τους χρησιμοποιούμε τον Data to Similarity operator και τον συνδέουμε με τον Process Document from Files operator. Ο Data to Similarity operator θα ελέγξει όλα τα κείμενα, περίπου όπως εξηγήσαμε παραπάνω, και θα δημιουργήσει έναν πίνακα. Στις δύο πρώτες στήλες του πίνακα θα υπάρχουν τα έγγραφα τα όμοια μοιάζουν μεταξύ τους και σε μία τρίτη στήλη θα υπάρχει ο βαθμός ομοιότητάς τους.

Στη συνέχεια θα προσπαθήσουμε να κάνουμε clustering στα έγγραφά μας χρησιμοποιώντας τον k-Means ή αλλιώς Clustering operator. Το clustering εξετάζει τα έγγραφα και προσπαθεί να εντοπίσει αυτά που μοιάζουν πιο πολύ μεταξύ τους και τα ομαδοποιεί.

4.3.5 Κατηγοριοποίηση και Πολικότητα Κειμένου

Μπαίνοντας στην ουσία του πράγματος, θα προσπαθήσουμε με το RapidMiner να κάνουμε αυτό για το για το οποίο μιλάει ολόκληρη η εργασία, δηλαδή, θα προσπαθήσουμε να βρούμε την πολικότητα των κειμένων. Αν οι κριτικές, εν τέλει, είναι θετικές ή αρνητικές. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε ανάλυση συναισθήματος.

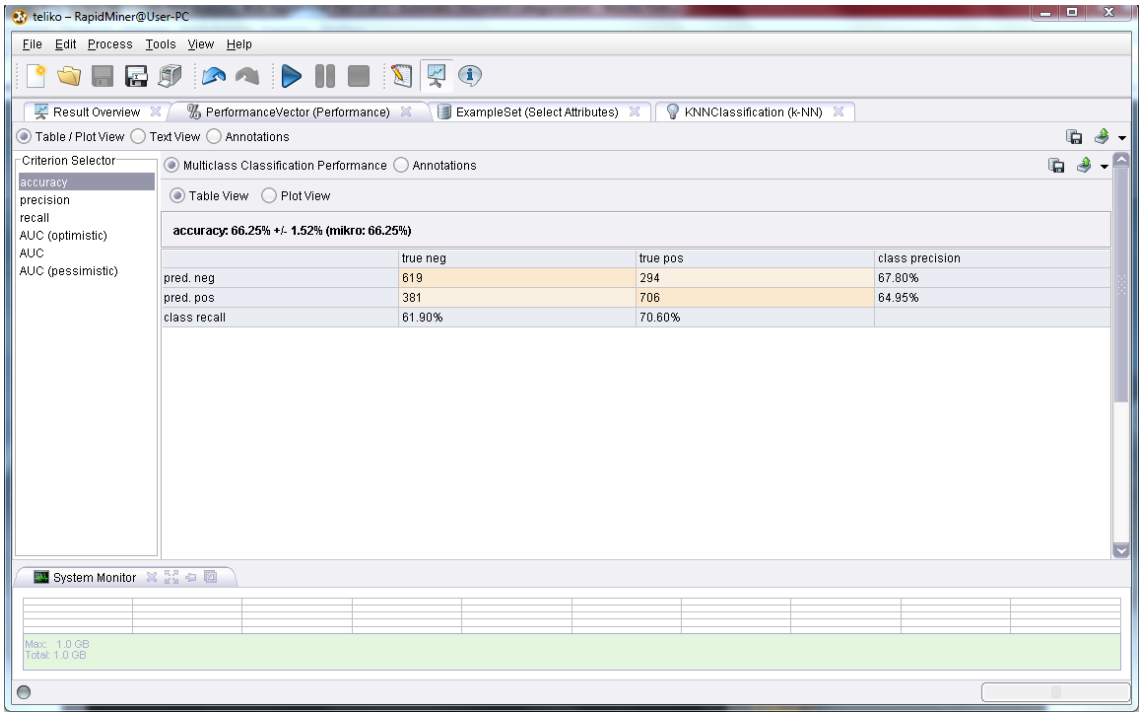
Αρχικά πρέπει να θυμηθούμε ποιους χειριστές περιέχει ο Process Documents from Files. Αυτοί είναι οι Transform Case operator που κάνει όλα τα γράμματα μικρά ή κεφαλαία (στη δική μας περίπτωση μικρά), Tokenize operator που χωρίζει το κείμενο σε λέξεις. Filter StopWords operator που απομακρύνει τα stop words, Stem operator που βρίσκει τις ρίζες των λέξεων και Filter Tokens (By Length) που απομακρύνει τις μικρές λέξεις (στην περίπτωσή μας επιλέγουμε να είναι αυτές του ενός χαρακτήρα). Ποιο αναλυτικά μπορούμε να δούμε τους χειριστές και τον τρόπο που συνδέονται στην παρακάτω εικόνα.



Εικόνα 19 Οι χειριστές που έχουμε επιλέξει για το σύστημα για την επεξεργασία δεδομένων

Για να προχωρήσουμε όμως στην κατηγοριοποίηση κειμένου θα πρέπει να προσθέσουμε έναν ακόμα χειριστή που θα κάνει αυτή τη δουλειά, τον Validation operator. Ο Validation operator στην ουσία κάνει διασταύρωση απόψεων, δηλαδή παίρνει τα δεδομένα και τα χωρίζει σε ομάδες τεκμηρίωσης (στην περίπτωσή μας επιλέγουμε 5) για να μπορέσει διασταυρώνοντας τα μεταξύ τους να τα επεξεργαστεί. Την επεξεργασία αυτή την πετυχαίνει με διάφορους αλγόριθμους που θα προσθέσουμε εμείς ανάλογα με το αποτέλεσμα που θέλουμε να έχουμε.

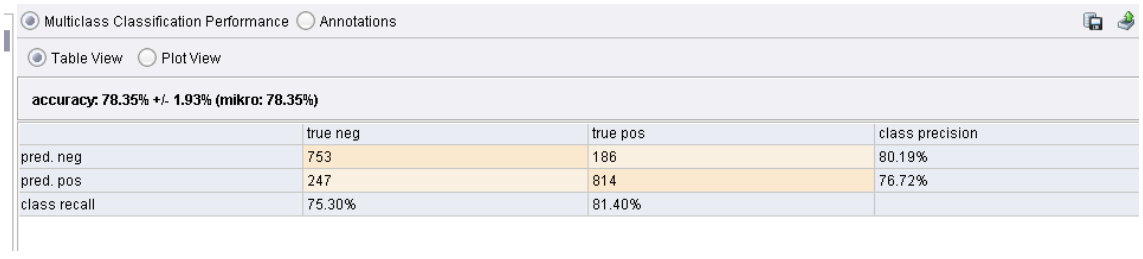
Για να βρούμε την πολικότητα των κριτικών των ταινιών, αρχικά, θα χρησιμοποιήσουμε έναν αλγόριθμο μηχανικής μάθησης, τον Naive Bayes και θα τον προσθέσουμε μέσα στον Validation στο training μέρος. Από δίπλα, δηλαδή στο testing μέρος θα προσθέσουμε άλλους δύο χειριστές για να πάρουμε περαιτέρω πληροφορίες για τα δεδομένα και για να μετρήσουμε την απόδοση, τον Apply Model operator και τον Performance operator. Αν τρέξουμε το πρόγραμμα θα πάρουμε τα εξής αποτελέσματα:



Εικόνα 20Τα πρώτα αποτελέσματα για τον καθορισμό της πολικότητας κειμένου

Απ' ότι μπορούμε να δούμε, οι αρνητικές κριτικές φτάνουν το 61,9% ενώ οι θετικές 70,6% με ακρίβεια κάθε κλάσης 67,8% για τον φάκελο neg με τις αρνητικές κριτικές και 64,95% για τον φάκελο pos δηλαδή τις θετικές κριτικές και ακρίβεια του συστήματος 66,25%. Αν και μπορούμε να πάρουμε μια πρώτη γεύση του τι συμβαίνει, δηλαδή, ότι οι θετικές κριτικές υπερτερούν, παρόλα αυτά δεν μπορούμε να βασιστούμε απόλυτα γιατί καταρχάς η απόδοση του Naive Bayes αλγορίθμου δεν είναι αυτή που περιμέναμε, η απόδοση του συστήματος στη συγκεκριμένη περίπτωση δεν είναι ικανοποιητική, και η ακρίβεια είναι πολύ μικρή. Γι' αυτό το λόγο θα τρέξουμε το πείραμά μας, για δεύτερη φορά με έναν άλλο αλγόριθμο.

Επιλέγουμε άλλον ένα αλγόριθμο μηχανικής μάθησης, τον SVM. Η χρήση του γίνεται με τον SVM operator. Συνδέοντας τον όπως προηγουμένως και τρέχοντας το πρόγραμμα παρατηρούμε κάτι πολύ ευχάριστο.



Εικόνα 21 Τα αποτελέσματα του SVM

Παρατηρούμε ότι η ακρίβεια έχει αυξηθεί αρκετά σε βαθμό που να μπορούμε πολύ εύκολα να χρησιμοποιήσουμε τα αποτελέσματα. Έχει φτάσει στο 78,35% για την ακρίβεια του συστήματος, 80,19% για την ακρίβεια του φακέλου neg και 76,72% για τον φάκελο pos. Έτσι τα αποτελέσματα, αν και επαληθεύουν το προηγούμενο αλγόριθμο, τώρα γίνονται πιο αξιόπιστα. Παρατηρούμε, ότι τώρα έχουμε 75,3% εμφάνιση αρνητικών σχολίων και 81,4% εμφάνιση θετικών. Οι αρνητικές κριτικές στον φάκελο neg είναι 753 και οι θετικές 247, ενώ οι αρνητικές κριτικές στον φάκελο pos είναι 186 και οι θετικές 814.

Χωρίς δισταγμό, λοιπόν, μπορούμε να πούμε ότι η ταινία για την οποία μιλούν οι κριτικές 2000 ανθρώπων, είναι καλή! Μπορούμε να πάμε να τη δούμε!

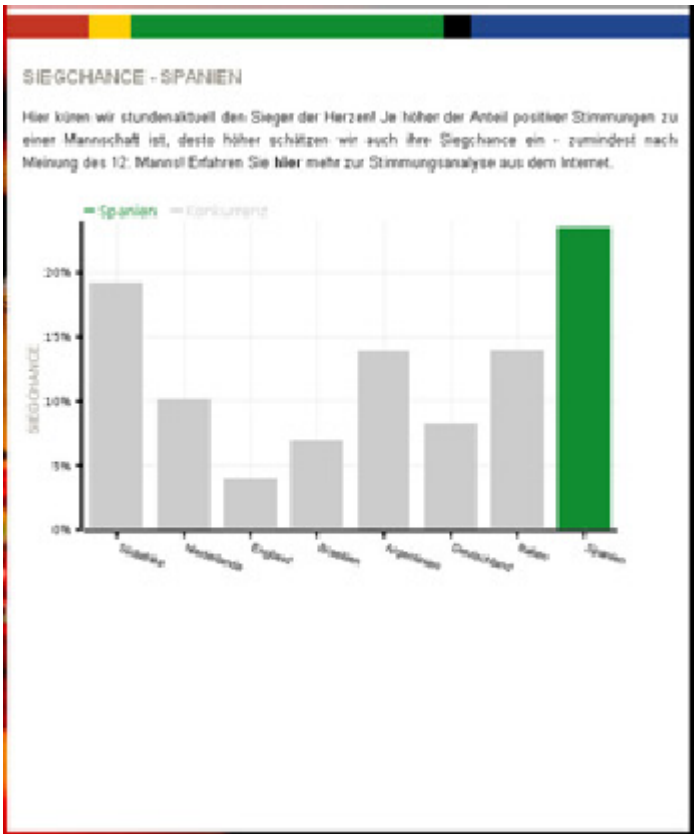
5 Επιτυχίες της Ανάλυσης Συναισθήματος

Κατά καιρούς έχουν γίνει διάφορες επιτυχημένες προβλέψεις με τη βοήθεια της ανάλυσης συναισθήματος. Οι προβλέψεις αυτές βασίζονται στο πως βλέπει ο κόσμος ένα συγκεκριμένο θέμα, τι πιστεύει γι' αυτό συμπεριλαμβάνοντας βέβαια και κάποιους εξωτερικούς παράγοντες.

5.1 Η Ανάλυση Συναισθήματος Προέβλεψε τον Νικητή του Mundial του 2010

Το καλοκαίρι του 2010 δεν ήταν όπως τα άλλα κι αυτό γιατί ο περισσότερος κόσμος αντί να βγει έξω και να χαρεί τον ήλιο και τη θάλασσα καθόταν μες στο σπίτι και έβλεπε αγώνες ποδοσφαίρου του Mundial του 2010. Ποιος όμως θα μπορούσε να γνωρίζει το αποτέλεσμα; Μα φυσικά η ανάλυση συναισθήματος η οποία κατάφερε και ερμήνευσε σωστά τις προσδοκίες του κόσμου.

Η εταιρία Rapid-I εξήγαγε μια έρευνα η οποίο ονομάστηκε "Mannschaft der Herzen" ("Η ομάδα της καρδιάς μας") με σκοπό την εύρεση του νικητή. Η λειτουργία του ήταν να ελέγχει το συναίσθημα στα κείμενα ειδήσεων, στα blogs , στα forum, σε διάφορες δημοσιεύσεις και να συγκεντρώνει τα αποτελέσματα. Η πορεία της έρευνας ήταν να συνδεθεί ο server του RapidMiner, RapidAnalytics, με 1000 online κανάλια και να αναλύει χιλιάδες κείμενα το λεπτό. Το πιο συγκλονιστικό αποτέλεσμα ήταν η πρόβλεψη της Ισπανίας σαν νικητή ακόμα πριν αρχίσει το γεγονός.



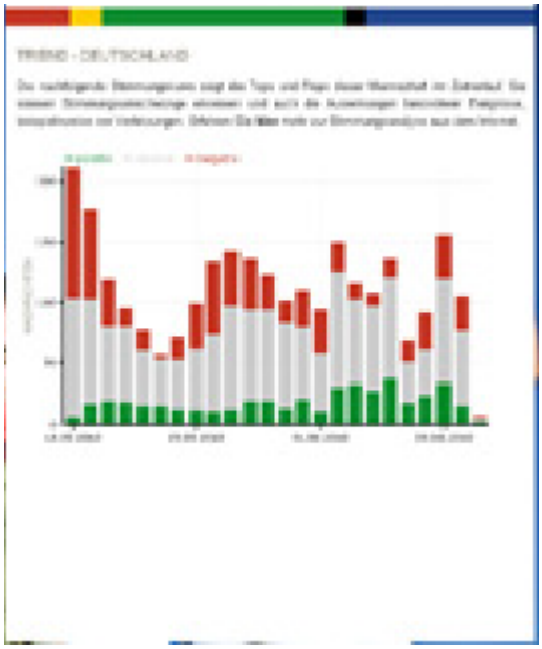
Εικόνα 22 Στην παραπάνω εικόνα βλέπουμε τα ποσοστά σε γράφημα στις 4 Ιουνίου, μια εβδομάδα πριν ξεκινήσει το τουρνουά.

Οι μπάρες δείχνουν την ποσότητα των θετικών δεδομένων έναντι των αρνητικών και των ουδέτερων. Όσο μεγαλύτερη είναι η αναλογία τόσο πιθανότερο είναι να κερδίσει η συγκεκριμένη ομάδα. Η Ισπανία είναι η πράσινη μπάρα στα δεξιά και παρατηρούμε ότι η αναλογία που έχει σε θετικό συναίσθημα είναι σαφώς μεγαλύτερη απ' τις υπόλοιπες ομάδες.

Ένα κομμάτι της έρευνας που παρουσιάζει αρκετό ενδιαφέρον είναι οι εναλλαγές του συναισθήματος στην πάροδο του χρόνου. Κατά τη διάρκεια του τουρνουά ,τα θετικά σχόλια για την ομάδα της Γερμανίας γίνονταν όλο και περισσότερα .Τελικά η Γερμανία αν και είχε τις περισσότερες πιθανότητες να κερδίσει αντικαταστάθηκε και πάλι από τη Ισπανία μέχρι τους τελικούς καθώς, προφανώς, οι οπαδοί βλέποντας την απόδοση της ομάδας άλλαξαν γνώμη.

Επιπλέον οι εξωτερικοί παράγοντες αποδείχτηκαν εξαιρετικά χρήσιμοι. Ήταν εύκολο να διαπιστώσουμε το γεγονός ότι ο αρχηγός της Αγγλίας Ferdinand τραυματίστηκε θα επηρέαζε αρκετά την γνώμη του κόσμου, συνεπώς και τα αποτελέσματα της ανάλυσης συναισθήματος. Οι παρακάτω εικόνες δείχνουν τις εναλλαγές του συναισθήματος στην πάροδο του χρόνου και τους εξωτερικούς

παράγοντες σαν tag cloud που εμφανίζουν το συναίσθημα του θέματος με χρώμα και την σημαντικότητα με το μέγεθος.



Εικόνα 23Οι εναλλαγές του συναισθήματος.

Εικόνα 24Οι εξωτερικοί παράγοντες.

5.2 Η Ανάλυση Συναισθήματος Προβλέπει τους Χρηματιστηριακούς Δείκτες

Οι διαθέσεις και τα συναισθήματα που εκφράζουν οι χρήστες του twitter και των ιστολογιών για τη χρηματιστηριακή αγορά αλλά και γενικότερα, ενδέχεται να δίνουν τη δυνατότητα να προβλέψουμε την κίνηση των μετοχών, σύμφωνα με αμερικανική

έρευνα. Και το πιο σημαντικό είναι ότι η "πρόβλεψη" είναι δυνατό να γίνει έως και μία εβδομάδα νωρίτερα.

Σύμφωνα με όσα γράφει η εφημερίδα "Daily Mail", οι υπολογισμοί των ερευνητών, υπό τον καθηγητή πληροφορικής του πανεπιστημίου της Ιντιάνα Γιόχαν Μπόλεν, που δημοσιεύτηκαν στο ηλεκτρονικό περιοδικό ανοικτής πρόσβασης arXiv, έδειξαν ότι η συσχέτιση ανάμεσα στην κίνηση του βασικού χρηματιστηριακού δείκτη Ντάου Τζόουνς και στη συλλογική διάθεση του κοινού, όπως αυτή αποκαλύπτεται μέσω του twitter, μπορεί να προβλέψει με ακρίβεια σχεδόν 90% πώς θα κινηθεί η χρηματιστηριακή αγορά.

Με τη βοήθεια του OpinionFinder μετρήθηκε η διαχρονική μεταβολή των συλλογικών συναισθημάτων. Οι ερευνητές ανέλυσαν περισσότερα από 9,8 εκατομμύρια σύντομα ηλεκτρονικά μηνύματα (tweets) από 2,7 εκατ. χρήστες στη διάρκεια ενός δεκάμηνου. Κκατέγραψαν ιδιαίτερα εκείνα τα μηνύματα που περιείχαν στοιχεία σχετικά με τα αρνητικά ή θετικά συναισθήματα των χρηστών όχι ειδικά για το χρηματιστήριο, αλλά γενικότερα για τη ζωή τους και την κοινωνία.

Οι διακυμάνσεις αυτής της διάθεσης (που αποτελούν ένα "βαρόμετρο" του κοινού αισθήματος), στη συνέχεια συσχετίστηκαν με τα "σκαμπανεβάσματα" του χρηματιστηρίου. Κάπως έτσι, προέκυψε ότι οι διακυμάνσεις της ψυχικής διάθεσης των χρηστών του twitter "προέβλεψαν" σε ποσοστό 87,6% το ημερήσιο κλείσιμο του δείκτη Ντάου Τζόουνς σε βάθος χρόνου. Η πιθανότητα να έχει επιτευχθεί από καθαρή τύχη αυτό το ποσοστό ακρίβειας, υπολογίστηκε σε μόλις 3,4%.

Το συμπέρασμα της έρευνας ανοίγει το δρόμο μελλοντικά οι επενδυτές να μπορούν να "παίζουν" στο χρηματιστήριο με βάση τα συλλογικά αισθήματα που εκφράζονται μέσω κοινωνικών δικτύων και ιστολογιών.

Για μια πραγματική ανακάλυψη τύπου "Εύρηκα" έκανε λόγο ο Μπόλεν, αν και παραδέχτηκε ότι δεν μπορεί να εξηγήσει αυτή την απρόσμενη συσχέτιση. Ειδικότερα, διαπιστώθηκε ότι ένα συγκεκριμένο συναίσθημα των ανθρώπων, η ηρεμία, φαίνεται να προβλέπει περισσότερο από κάθε άλλο προς ποια κατεύθυνση θα κατευθυνθούν οι μετοχές τις επόμενες μέρες.

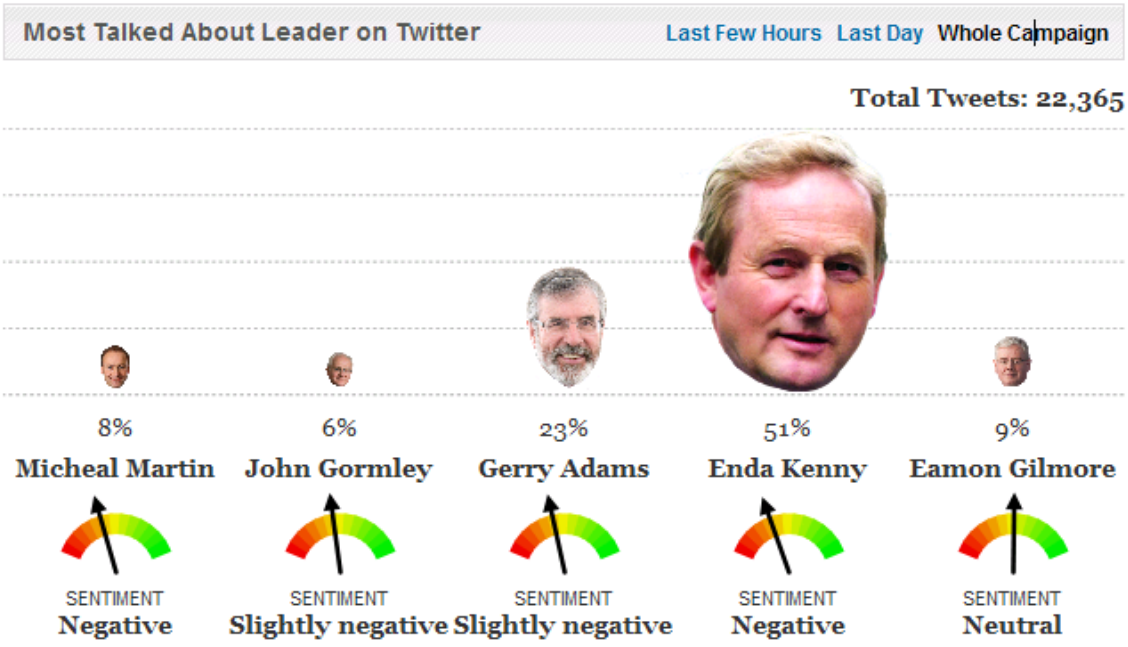
Οι ερευνητές δημιούργησαν ένα επιμέρους "δείκτη ηρεμίας" (με βάση τα σχετικά αισθήματα που αποκαλυφτήκαν μέσω του twitter) και βρήκαν ότι αυτός μπορεί να προβλέψει από δύο έως έξι μέρες νωρίτερα αν το χρηματιστήριο θα πέσει ή θα ανέβει.

Προηγούμενες έρευνες είχαν δείξει ότι και τα blogs, που έχουν πλέον πολλαπλασιαστεί εντυπωσιακά διεθνώς, είναι σε θέση να "αποκαλύψουν" τη διάθεση της κοινής γνώμης.

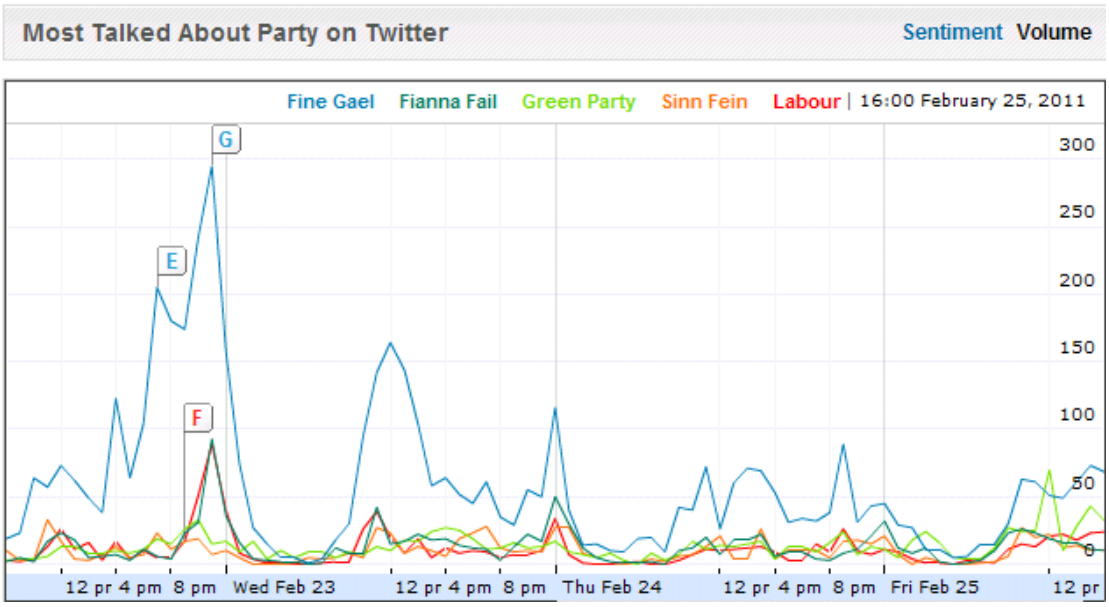
5.3 Η Ανάλυση Συναισθήματος Προβλέπει το Αποτέλεσμα των Εκλογών του 2011 στην Ιρλανδία.

Μέχρι τώρα το κλίμα των εκλογών χαρακτηρίζονταν από πάμπολα γραφήματα των λεγόμενων Exit Polls, που εμφανίζονταν σε κάθε μέσο ενημέρωσης και προσπαθούσαν να πληροφορήσουν τον κόσμο για το ποια θα είναι τα αποτελέσματα. Πολύχρωμα, φανταχτερά διαγράμματα που τα περισσότερα όμως είχαν μεταξύ τους διαφορετικά αποτελέσματα. Πόσος κόπος και πόσα χρήματα χρειάζονταν κάθε φορά για να δημιουργηθούν! Και εδώ έρχεται να δώσει τη λύση η ανάλυση συναισθήματος.

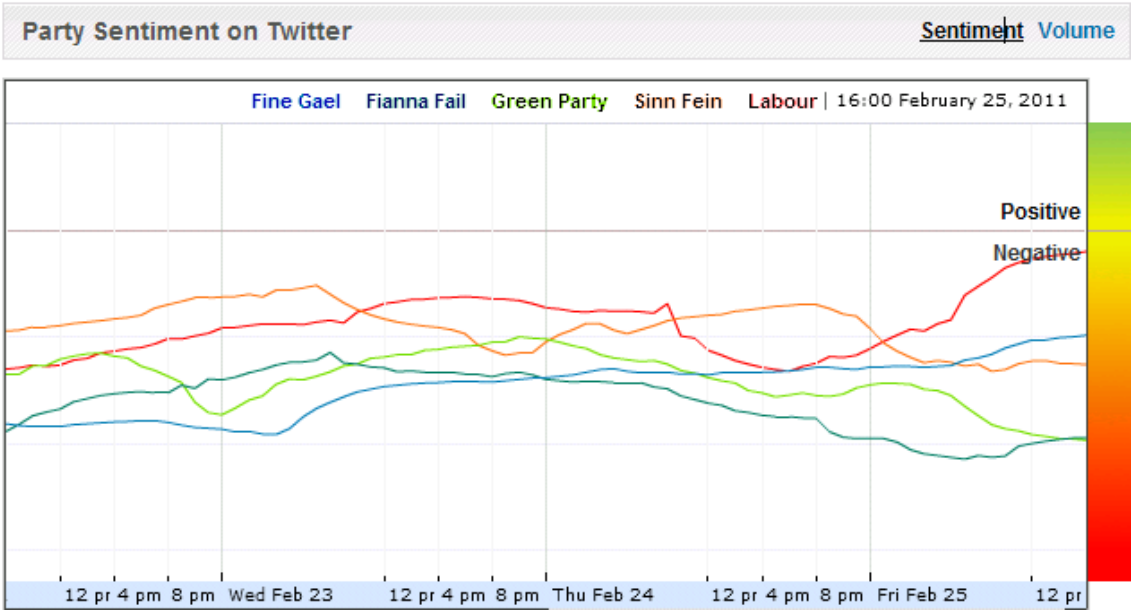
Παρακάτω πρόκειται να παρουσιάσουμε μια έρευνα που έγινε για τις εκλογές του 2011 στην Ιρλανδία λίγο πριν αυτές ξεκινήσουν. Σκοπός της ήταν να προβλέψουν τα αποτελέσματα με όσο το δυνατόν μεγαλύτερη ακρίβεια. Το πείραμα αυτό διεξήγαγε το διαδικτυακό site ειδήσεων The Journal που με τη βοήθεια του Twitter χρησιμοποίησαν τις συζητήσεις των χρηστών για να προβλέψουν τον νικητή. Στις 25 Φεβρουαρίου, λοιπόν, η εικόνα είχε ως εξής:



Εικόνα 25 Τα αποτελέσματα του Twitter για τις εκλογές του 2011 στην Ιρλανδία



Εικόνα 26 Εδώ μπορούμε να δούμε τον όγκο των συζητήσεων που αφορούν τις εκλογές της Ιρλανδίας για το 2011



Εικόνα 27 Εδώ φαίνεται το συναίσθημα που έχει εξαχθεί απ' τις ίδιες συζητήσεις

Όπως βλέπουμε απ' τις δύο πρώτες εικόνες το κόμμα Fine Gael και ο αρχηγός του Enda Kenny υπερέχουν κατά πολύ στον αριθμό των συζητήσεων, από τα άλλα κόμματα. Παρόλα αυτά όμως αρκετές απ' αυτές τις συζητήσεις είναι αρνητικές. Αυτό το στοιχείο δεν θα πρέπει να περάσει απαρατήρητο. Για να μπορέσουμε, όμως, να καταλάβουμε περισσότερα για τη φύση των συζητήσεων που εξετάζουμε θα πρέπει να προσθέσουμε σε όλα αυτά και το επίπεδο του συναισθήματος. Τα αποτελέσματα μπορούμε να τα δούμε στην εικόνα 17.

Αμέσως μετά τη διεξαγωγή της έρευνας βγήκαν και τα αποτελέσματα των εκλογών. Πρωθυπουργός της Ιρλανδίας ήταν πια ο Enda Kenny. Όπως ακριβώς είχε προβλέψει το twitter με τη βοήθεια της ανάλυσης συναισθήματος. Έτσι, λοιπόν, χωρίς κάθε φορά να τραβολογάμε τους ανθρώπους να μας πουν την άποψή τους, χωρίς να τους ενοχλούμε και κυρίως χωρίς κόστος μπορούμε να βγάλουμε αποτελέσματα και κυρίως αξιόπιστα.

5.4 Το Twitter Προβλέπει τις Επιτυχίες του Box Office

Ερευνητές της εταιρίας HP (Hewlett Packard) ανέλυσαν τρία εκατομμύρια μηνύματα (τα γνωστά tweets) για περίπου 25 ταινίες. Ανακάλυψαν ότι η ταχύτητα κατά την οποία αναπαράγονται αυτά τα μηνύματα μπορεί να χρησιμοποιηθεί για να προβλέψει τι κέρδη θα έχει μια ταινία πριν καν αρχίσει να προβάλλεται. Όπως δήλωσε και ο Bernardo Huberman, αρχηγός του εργαστηρίου κοινωνικών υπολογισμών της HP, στις ειδήσεις του BBC, ¨Οι προβλέψεις μας ήταν απίστευτα κοντά!¨. Για παράδειγμα, είπε ότι το σύστημα προέβλεψε ότι η ταινία The Crazies θα έφτανε τα 16,8 εκατομμύρια δολάρια και στην πραγματικότητα έφτασε τα 16,06 εκατομμύρια. Επίσης προέβλεψε ότι το ποσό που θα έφτανε το αισθηματικό δράμα Dear John ήταν 30,71 εκατομμύρια και στην πραγματικότητα έπιασε 30,46 εκατομμύρια δολάρια.

Η ομάδα κατάφερε να κάνει τις προβλέψεις για τα έσοδα της πρώτης εβδομάδας αναλύοντας τα μηνύματα που είχαν σχέση με τη συγκεκριμένη ταινία μέχρι την κυκλοφορία της. Για την ανάλυση αυτή αναπτύχθηκαν ειδικοί αλγόριθμοι που μετρούσαν την ταχύτητα με την οποία παράγονταν τα μηνύματα. ¨Όσο γρηγορότερα δημοσιεύουν μηνύματα οι άνθρωποι τόσο πιο πιθανό είναι να πάνε και να δούνε την ταινία¨.

Έπειτα η ομάδα μπορούσε να προβλέψει την τρέχουσα επιτυχία της ταινίας, συμπεριλαμβανομένων, και των εσόδων της δεύτερης εβδομάδας με τη βοήθεια της ανάλυσης συναισθήματος. Ανέλυαν τα μηνύματα και στη συνέχεια αποφάσιζαν αν είναι θετικά, αρνητικά ή ουδέτερα.

Το εργαλείο που χρησιμοποίησαν ήταν το Mechanical Turk από την Amazon, το οποίο πληρώνει ανθρώπους για να κάνουν μικρές δουλειές που όμως οι υπολογιστές δυσκολεύονται. Οι άνθρωποι κατηγοριοποιούσαν τα μηνύματα και έπειτα έπιανε δουλειά η ανάλυση συναισθήματος. Μετά, το σύστημα εντόπιζε τα θετικά των ταινιών

και τα χρησιμοποιούσε σε άλλα προγνωστικά προγράμματα όπως το Hollywood Stock Exchange. Για παράδειγμα, η ανάλυση έδειξε μία έκρηξη σε θετικά συναισθήματα για την ταινία The Blind Side που μάλιστα προτάθηκε και για Oscar αλλά έδειξε τα αντίθετα αποτελέσματα για το New Moon το οποίο αν και ξεκίνησε καλά, στη συνέχεια έχασε τους θεατές του.

Ο Dr Huberman πιστεύει ότι η δημογραφικότητα του Twitter μπορεί να περιορίσει τη χρήση άλλων συστημάτων ανάλυσης, όπως αυτά που προορίζονται για παιδιά. Παρόλα αυτά, μπορεί να φανεί πιο χρήσιμη σε άλλα θέματα, όπως το πόσο θα πουλήσει ένα προϊόν.

5.5 Το RapidMiner Αποκαλύπτει ότι ένα Καινούριο Απορρυπαντικό Βρωμάει

Το 2008 ένας παραγωγός απορρυπαντικών ήθελε να ανακαλύψει τι πιστεύουν οι καταναλωτές για ένα καινούριο προϊόν του. Αντί να χάσει χρόνο και χρήμα με τις ακριβές μελέτες που θα έκανε παραδοσιακά μια εταιρεία marketing, αποφάσισε να αφήσει το RapidMiner να κάνει αυτή τη δουλειά, με σκοπό να του δώσει μια ιδέα για την κατάσταση της αγοράς, φτηνά και γρήγορα.

Μετά από μια μικρή περίοδο που το RapidMiner συνέλλεξε δεδομένα που είχαν να κάνουν με το συγκεκριμένο προϊόν, από σελίδες σε ολόκληρο το διαδίκτυο, επέλεξε τις πιο ολοκληρωμένες και σχετικές απόψεις και ανέλυσε το συναίσθημά τους. Πολλοί καταναλωτές είχαν εκφράσει ελεύθερα την άποψή τους για διάφορα προϊόντα σε πολλές σελίδες του διαδικτύου, forums, blogs. Δεν υπάρχει καλύτερος τρόπος για τη συλλογή απόψεων απ' αυτόν, χωρίς να ενοχλεί κανείς τον κόσμο με ερωτήσεις, φθηνά και γρήγορα.

Ο παραγωγός ανακάλυψε ότι το απορρυπαντικό του δεν το προτιμούσαν λόγω της δυνατής και άσχημης μυρωδιάς που έβγαζε! Έτσι, λοιπόν, η προηγούμενη διαφημιστική εκστρατεία που είχε ως στόχο να εκθειάσει την καθαριστική δράση του απορρυπαντικού απέτυχε γιατί πολλοί καταναλωτές πίστευαν ότι «βρωμούσε»! Από κει κι ύστερα ο παραγωγός μπορούσε να επέμβει και να αλλάξει αυτό που έκανε το προϊόν του τόσο λίγο δημοφιλή, τη μυρωδιά. Αμέσως μετά, η εταιρία ξεκίνησε μια καινούρια διαφημιστική καμπάνια με νέο στόχο τη φρεσκάδα και την υπέροχη μυρωδιά που έβγαζε το απορρυπαντικό και έτσι να ξανακερδίσουν τους πελάτες που είχαν χάσει και να προσελκύσουν και καινούριους.

Με το λογισμικό RapidMiner, την ειλικρινή γνώμη των καταναλωτών και την ανάλυση συναισθήματος, η εταιρία γλύτωσε από ένα πολύ μεγάλο φιάσκο, χρήματα και τη δύσκολη θέση που θα βρισκόταν αν τελικά απέσυρε το προϊόν της από την αγορά λόγω των χαμηλών πωλήσεων.

6 Συμπεράσματα

Η εύρεση πολικότητας κειμένου ή διαφορετικά η ανάλυση συναισθήματος είναι ένας αρκετά περίπλοκος και δύσκολος θα λέγαμε κλάδος. Καταρχάς ακόμα και οι άνθρωποι δυσκολεύονται ορισμένες φορές να καταλάβουν τι λέει κάποιος άλλος και το συναίσθημα των λεγόμενων του. Όσο για τους υπολογιστές, το να βρούνε τον τόνο και το νόημα σε ένα συναίσθημα δεν είναι καθόλου εύκολη υπόθεση καθώς οι άνθρωποι εκφράζονται με διαφορετικούς τρόπους ο καθένας Πολλές εφαρμογές προσπαθούν να κατανοήσουν το συναίσθημα χρησιμοποιώντας λέξεις κλειδιά. Για παράδειγμα, αν η λέξη «χαρούμενος» χρησιμοποιείται σε μία πρόταση τότε η πρόταση είναι θετική. Τι γίνεται όμως όταν έχουμε την πρόταση ‘Θα ήμουν πάρα πολύ χαρούμενος αν δεν ξανάβλεπα αυτή την ταινία’.

Πέρα απ’ τις δυσκολίες, τίθενται και ορισμένα άλλα θέματα όπως η ιδιωτικότητα και η χειραγώγηση που μπορεί να υπάρξει, που κάνουν τον κλάδο ακόμα πιο περίπλοκο. Οι εφαρμογές που χρησιμοποιούν δεδομένα για τις προτιμήσεις των χρηστών μπορεί να φέρουν διάφορες ανησυχίες για την παραβίαση της ιδιωτικής τους ζωής. Για παράδειγμα, το να ερευνά μια εταιρία καφέ τη άποψη ενός χρήστη για τα προϊόντα της από ένα blog δεν είναι τόσο μεμπτό, το να ελέγχονται όμως οι συζητήσεις του κινητού του τηλεφώνου για την άποψη που έχει για μια χώρα από κυβερνητικούς της αξιωματούχους μπορεί να αποφέρει πολλά προβλήματα.

Σε κλάδους όπου εξυπηρετούνται οικονομικά συμφέρονται πολλές φορές τίθεται το θέμα της χειραγώγησης. Οι εταιρείες ήδη συμμετέχουν στις online συζητήσεις μεταξύ των καταναλωτών σας μέρος των δημοσίων σχέσεών τους αφού, όπως λένε, ‘οι εταιρίες δεν μπορούν να ελέγξουν το περιεχόμενο των απόψεων των καταναλωτών’. Μπορούν όμως να δώσουν προσοχή σ’ αυτό και σε πολλές περιπτώσεις να το επηρεάσουν. Σε μία εργασία που εξήγαγε ο Aberdeen [12] περισσότερο από τις διπλάσιες εταιρείες που χρησιμοποιούν σελίδες κοινωνικής δικτύωσης συμμετέχουν ενεργά στις συζητήσεις των καταναλωτών έναντι αυτών που μένουν παθητικοί παρατηρητές (67% vs. 33%). Πάνω από το ένα τρίτο συμμετέχει καθημερινά σε αυτές τις συζητήσεις (39%) και αλληλεπιδρά με τους καταναλωτές σε μια προσπάθεια να τους κατευθύνει τη γνώμη, να

διορθώσει λάθος πληροφορίες, να ψαρέψει πελατεία, να ανταμείψει την αφοσίωση, να δοκιμάσει καινούριες ιδέες ή για πολλούς άλλους λόγους.

Παρόλο που είναι ένα δύσκολο πεδίο και σε ορισμένες περιπτώσεις και η τεχνολογία αλλά και οι άνθρωποι δυσκολεύονται να το αντιμετωπίσουν και παρόλους τους ηθικούς φραγμούς που είδαμε παραπάνω, στο μόνο πράγμα που συμφωνούν όλοι είναι ότι πρέπει να το κατανοήσουμε. Τα πλεονεκτήματά του τα συναντήσαμε κατά τη διάρκεια αυτής της εργασίας και όπως διαπιστώσαμε υπερτερούν κατά πολύ από τα μειονεκτήματα στα οποία γίνονται συνεχώς προσπάθειες να βρεθούν λύσεις με την εξέλιξη του κλάδου. Βοηθάει τις εταιρίες να κατανοήσουν τι σκέφτονται οι χρήστες για τα προϊόντα, τις υπηρεσίες, την εμπειρία τους, την εξυπηρέτηση πελατών ακόμα και τον ανταγωνισμό. Βοηθάει ακόμα και οργανισμούς να αναγνωρίσουν θέματα έκτακτης ανάγκης και μηνύματα που ζητούν βοήθεια.

Συνεισφέρει τόσο πολύ στις εταιρίες που στις 26 Ιανουαρίου 2011 η εταιρία Google δεν δίστασε να πληρώσει δέκα εκατομμύρια δολάρια για την απόκτηση ενός εργαλείου ανάλυσης συναισθήματος. Το μόνο που μένει είναι να περιμένουμε μέχρι να είναι έτοιμο για να δούμε και αυτό τι δυνατότητες μπορεί να μας προσφέρει.

Κατά τη διάρκεια αυτής της εργασίας συνάντησα πολλές δυσκολίες. Αρχικά, η πολικότητα κειμένου είναι ένα θέμα που αν δεν έχει κάποιος τις βασικές γνώσεις τουλάχιστον από data mining δεν μπορεί εύκολα να κατανοήσει. Χρειάζεται επίσης περεταίρω διερεύνηση σε διαφόρους άλλους κλάδους όπως η ανάλυση κειμένου, η επεξεργασία φυσικής γλώσσας, η υπολογιστική γλωσσολογία, η μηχανική μάθηση και πολλά άλλα. Παρόλα αυτά η βιβλιογραφία είναι αρκετά μεγάλη. Υπάρχουν πάρα πολλά άρθρα που καλύπτουν κάθε πτυχή του συγκεκριμένου θέματος με διάφορα παραδείγματα, αναλύσεις, παρουσιάσεις και πιο εξειδικευμένες έρευνες που με τη βοήθειά τους και φυσικά αρκετή μελέτη μπορεί να λυθεί οποιαδήποτε απορία μας έχει παρουσιαστεί. Δυστυχώς, όμως, ο χρόνος που είχα στη διάθεσή μου ήταν πολύ λίγος για να καλύψω όλο αυτό τον όγκο των αρχείων, παρόλα αυτά είμαι πεπεισμένη ότι έχω δώσει βάση στα σημαντικότερα σημεία και τα έχω συντάξει έτσι ώστε να δημιουργηθούν στον αναγνώστη όσο το δυνατόν λιγότερες απορίες.

Μετά λύπης μου, όμως, συνειδητοποίησα ότι πολύ λίγη, έως και ελάχιστη έρευνα έχει γίνει περί του θέματος στην Ελλάδα. Υπάρχουν βέβαια κάποια πολύ καλά άρθρα που έχουν δημοσιευτεί, τα οποία και συμπεριλαμβάνω στη βιβλιογραφία αυτής της εργασίας, αλλά για ένα τόσο μεγάλο και συναρπαστικό θέμα, όπως είναι η ανάλυση

συναισθήματος, θα περίμενε κανείς πολύ πιο μεγάλο ενδιαφέρον. Βέβαια, πρέπει να λάβουμε υπ' όψιν μας ότι οι απαίτησης στην Ελλάδα γι' αυτόν τον κλάδο δεν είναι ακόμα πολύ μεγάλες, καθώς δεν υπάρχουν πολλές εταιρίες αυτού του βεληνεκούς που να χρειάζονται και να μπορούν να χρησιμοποιήσουν ανάλυση συναισθήματος. Έπειτα, το κόστος για μια τέτοιου είδους ολοκληρωμένη έρευνα είναι πολύ μεγάλο και τα δεδομένα στον ελληνικό ιστό πολύ λίγα. Δεν χρειάζεται και δεν μπορεί να χρησιμοποιήσει το ίδιο την ανάλυση συναισθήματος μια πολυεθνική εταιρία με πελάτες σε όλα τα μήκη και τα πλάτη της γης με μια μικρή επιχείρηση που το αγοραστικό της κοινό περιορίζεται στην πόλη την οποία βρίσκεται.

Μία άλλη δυσκολία την οποία συνάντησα είναι τα αρκετά πολύπλοκα συστήματα που υπάρχουν. Τα περισσότερα απ' αυτά δεν εξειδικεύονται στην ανάλυση συναισθήματος αλλά καλύπτουν ένα ευρύ φάσμα ανάλυσης κειμένου, ή ανάλυσης πινάκων, επεξεργασία φυσικής γλώσσας κ.τ.λ. Αυτό τα καθιστά αρκετά δύσκολα στην κατανόησή τους και δύσχρηστα στη λειτουργία τους. Το μόνο πράγμα που βρήκα αρκετά εύκολα, μπορώ να πω, είναι τα δεδομένα τα οποία χρησιμοποίησα, καθώς βρίσκονται παντού σε forum, blogs, merchandise sites και πολλές άλλες σελίδες. Το μόνο που μένει να κάνουμε είναι απλά να αποφασίσουμε το θέμα το οποίο θα αναλύσουμε!

Μελετώντας όμως για το συγκεκριμένο θέμα και παρόλες τις δυσκολίες που αντιμετώπισα άρχισα να ανακαλύπτω όλες τις δυνατότητες που μπορεί να προσφέρει και δεν μπορώ να κρύψω ότι εντυπωσιάστηκα. Ήξερα, βέβαια, ότι η τεχνολογία έχει προχωρήσει σημαντικά και ο κλάδος της τεχνητής νοημοσύνης έχει εξελιχθεί κατά πολύ αλλά η ανακάλυψη της ανάλυσης συναισθήματος με εξέπληξε ευχάριστα. Αν σκεφτούμε, ειδικά, ότι στην ουσία ο υπολογιστής, μια μηχανή δηλαδή, μπορεί να διαβάσει τη σκέψη των ανθρώπων, κάτι που, πρακτικά, οι άνθρωποι δεν μπορούν να κάνουν.

Έτσι, λοιπόν, αποφάσισα η σχέση μου με το συγκεκριμένο θέμα να μη σταματήσει εδώ. Εκτός από την περαιτέρω έρευνα και μελέτη που θα κάνω για την ανάλυση συναισθήματος καθώς σκέφτομαι ότι θα μπορούσε να αποτελέσει το αντικείμενο δουλειάς μου στο μέλλον, θα ήθελα πολύ να ασχοληθώ πιο εξειδικευμένα με τη δημιουργία ενός νέου συστήματος που θα ασχολείται αποκλειστικά και μόνο με ανάλυση συναισθήματος και ίσως αν τα καταφέρω να συνεχίσω κάνοντάς το τα δεδομένα που θα δέχεται σαν είσοδο να είναι στην ελληνική γλώσσα. Το σχέδιο μου

είναι αρκετά φιλόδοξο και είμαι σίγουρη ότι με την πάροδο του χρόνου θα τα καταφέρω.

Η ανάλυση συναισθήματος είναι ένας πολύ ενδιαφέρον τομέας με πολλές δυνατότητες αλλά και πολλές προκλήσεις. Δυστυχώς ή ευτυχώς έχει μπει πλέον στην καθημερινότητά μας με εντυπωσιακά αποτελέσματα και μας υπόσχεται ότι στο μέλλον θα δούμε ακόμα περισσότερα. Εύχομαι σε αυτούς που δεν γνώριζαν το θέμα να τους άρεσε η παρουσίαση και το κοιτάξουν περαιτέρω και σε αυτούς που το γνώριζαν να το ψάξουν ακόμα περισσότερο γιατί απλά αξίζει τον κόπο.

Βιβλιογραφία

[1] Karanikas H. , Theodoulidis B. ," Knowledge Discovery in Text and Text Mining Software". Centre for Research in Information Management: November 2002

[2] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Institute for Information Technology National Research Council of Canada Ottawa

[3] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques". Department of Computer Science Cornell University Ithaca, IBM Almaden Research Center

[4] Ειρήνη Καλδέλη, , "Εκπαίδευση ταξινομητών κειμένου για το χαρακτηρισμό άποψης".

[5] Ιωάννης Θ. Νασίκας, "Text Mining: Μια νέα προτεινόμενη μέθοδος με χρήση κανόνων συσχέτισης". Πανεπιστήμιο Πατρών, Ιούνιος 2006

[6] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, , " Sentiment Strength Detection in Short Informal Text ". Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton.

[7] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis". Foundations and TrendsR in Information Retrieval.

[8] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pp. 483–490, 2006.

[9] Farah Benamara, Carmine Cesarano και Diego Reforgiato, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone"

[10] Zhe Xu, "A Sentiment Analysis Model Integrating Multiple Algorithms and Diverse Features", Ohio State University, 2010.

[11] Bing Liu, "Sentiment Analysis and Subjectivity", Department of Computer Science, University of Illinois at Chicago

[12] Aberdeen , "more than 250 enterprises using social media monitoring and analysis solutions in a diverse set of enterprises"

Χρήσιμα Links

[1] http://www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf

[2] http://aibook.csd.auth.gr/include/slides/Chap18.pdf

[3] http://aibook.csd.auth.gr/include/ch18.pdf

[4] http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf

[5] http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

[6] http://www.puffinwarellc.com/index.php/news-and-articles/articles/33.html?start=1

[7] http://delab.csd.auth.gr/courses/c_ir/lsa.pdf

[8] http://www.gelbukh.com/clbook/Computational-Linguistics.htm#_Toc86751631

[9] http://blog.typeslashcode.com/voxpop/2010/02/OpinionFinder -open-source-sentiment-analysis-toolkit/

[10] http://rapid-i.com/content/view/181/190/lang,en/

[11] http://smallbiztrends.com/2010/03/tracking-twitter-sentiment.html

[12] http://socialmouths.com/blog/2010/03/31/6-tools-for-twitter-sentiment-tracking/

[13]http://www.readwriteweb.com/archives/sentiment_analysis_is_ramping_up_in_2009.php

[14] http://www.attensity.com/applications-and-solutions/attensity-analytics-suite/

[15] http://www.sas.com/text-analytics/sentiment-analysis/index.html#section=1

[16] http://blog.typeslashcode.com/voxpop/2010/02/OpinionFinder -open-source-sentiment-analysis-toolkit/

[17] http://alias-i.com/lingpipe/index.html

[18] http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it

[19] http://news.bbc.co.uk/2/hi/technology/8612292.stm

[20] http://brenocon.com/blog/2008/12/facebook-sentiment-mining-predicts-presidential-polls/

[21] http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon

[22] https://www.ibm.com/developerworks/mydeveloperworks/blogs/business-analytics/entry/can_twitter_sentiment_analysis_predict_outcomes_like_the_irish_election?lang=en_us

[23] http://kmandcomputing.blogspot.com/2008/06/opinion-mining-with-rapidminer-quick.html

[24] http://vancouverdata.blogspot.com/2010/11/text-analytics-with-rapidminer-part-4.html

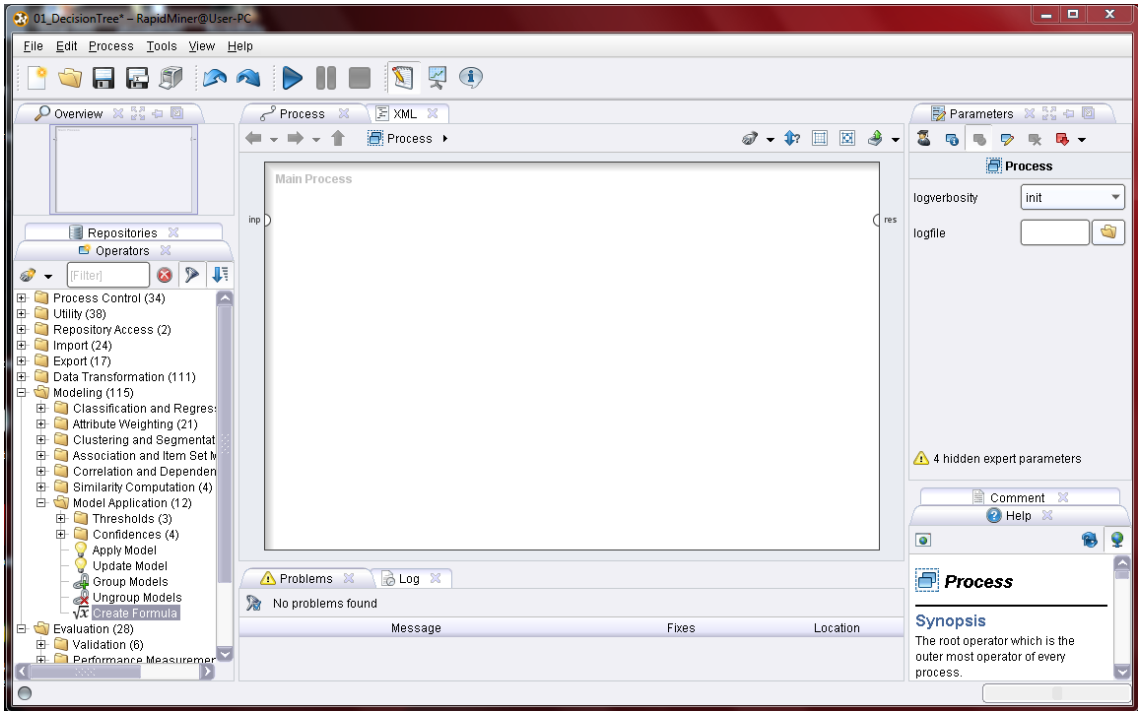
[25] http://www.kdnuggets.com/

[26] http://rapid-i.com/rapidforum/index.php?action=search2

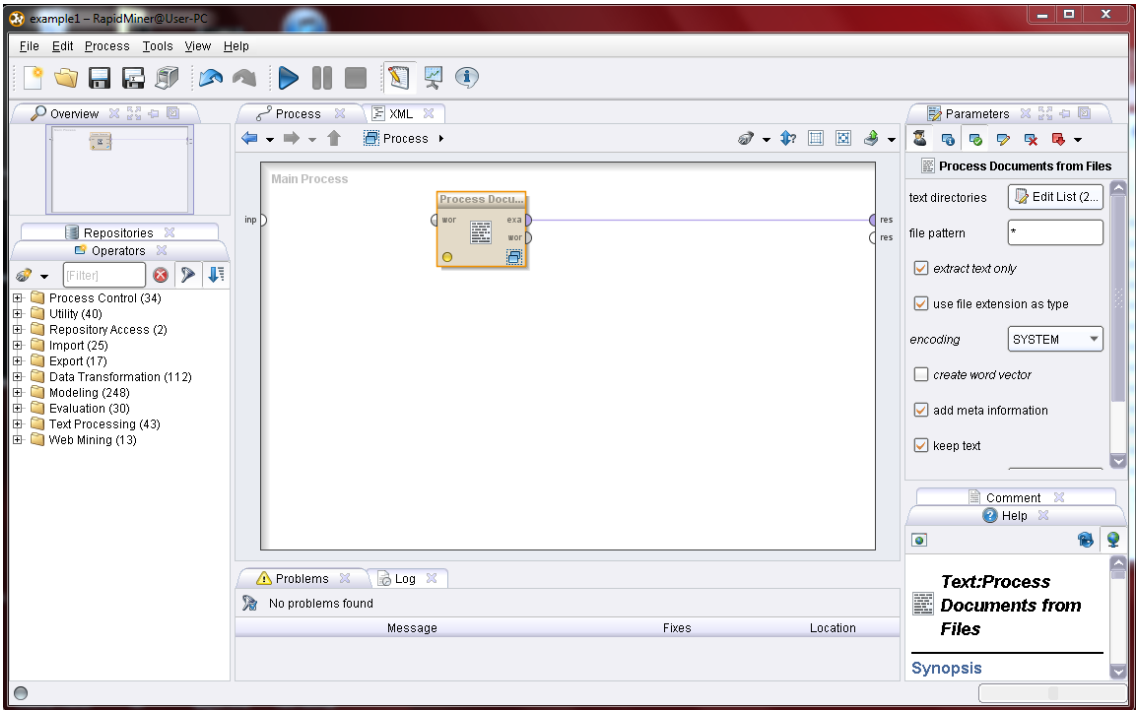
[27] http://nlp.stanford.edu/courses/cs224n/2010/reports/ssoriajr-kanej.pdf

[28] http://www.livepedia.gr/index.php/%CE%95%CE%BB%CE%BB%CE%B7%CE%BD%CE%B9%CE%BA%CE%AE_%CE%95%CE%BB%CE%B5%CF%8D%CE%B8%CE%B5%CF%81%CE%B7_%CE%95%CE%B3%CE%BA%CF%85%CE%BA%CE%BB%CE%BF%CF%80%CE%B1%CE%AF%CE%B4%CE%B5%CE%B9%CE%B1

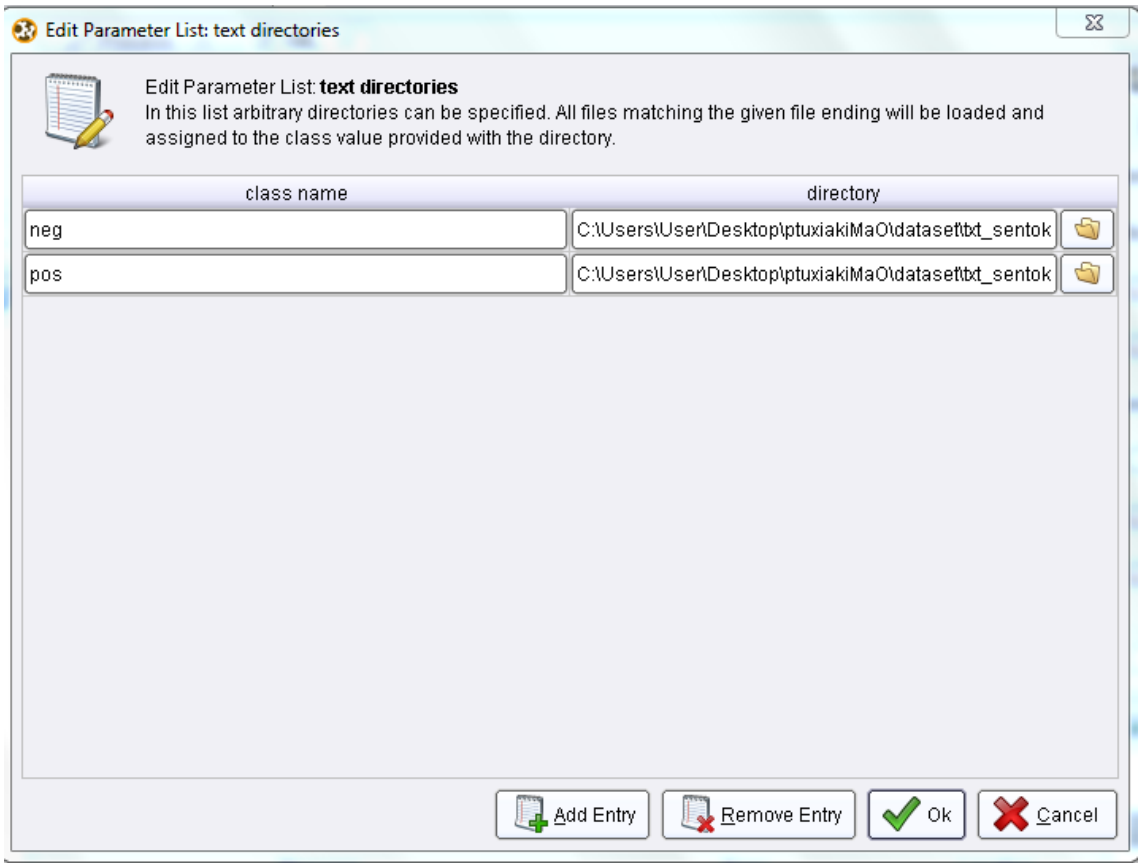
Παράρτημα



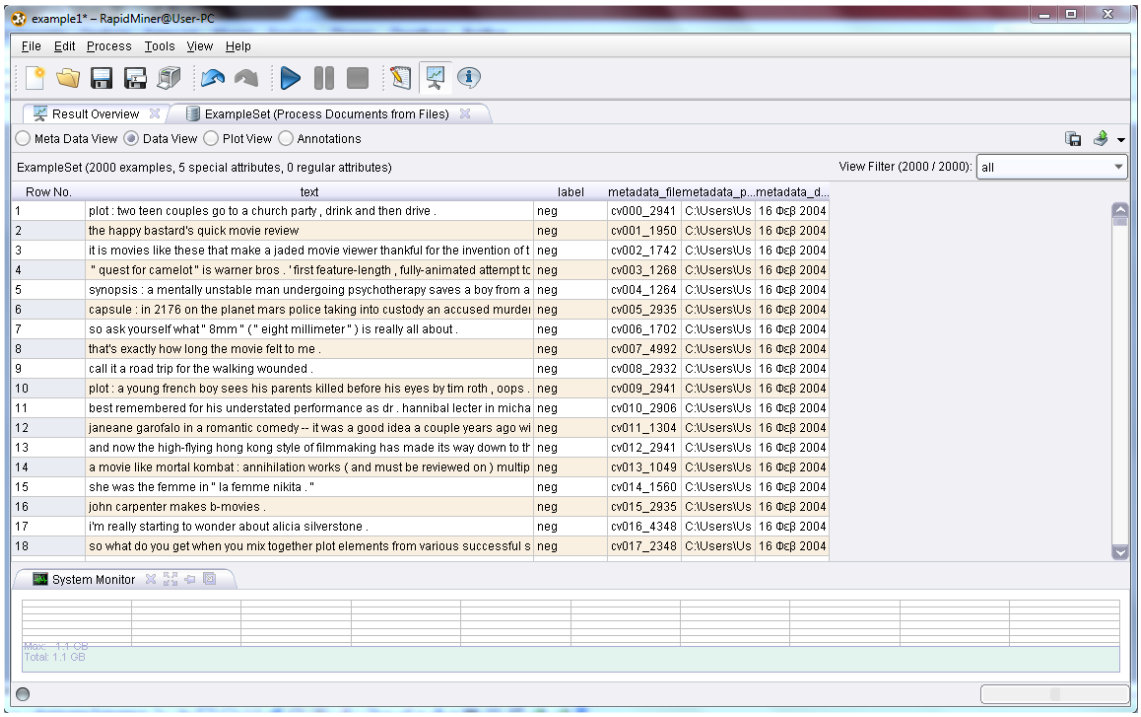
Εικόνα 28 Το RapidMiner



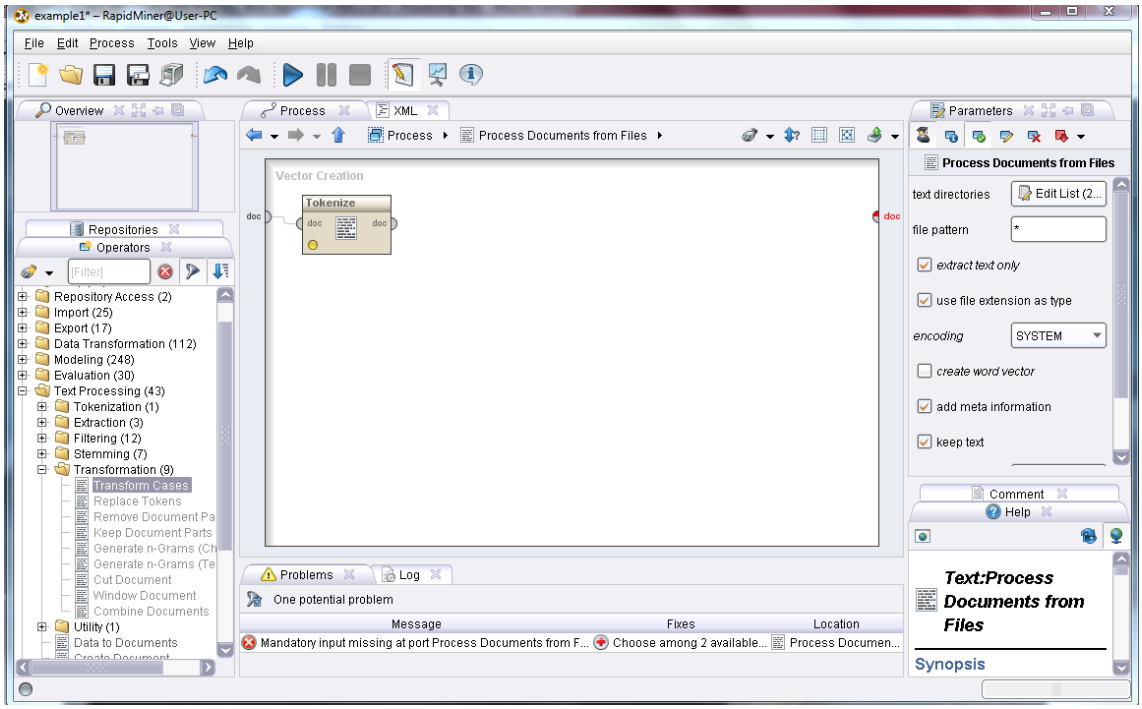
Εικόνα 29 Εισάγοντας τον Process Documents from Files operator για να βάλουμε τα δεδομένα.



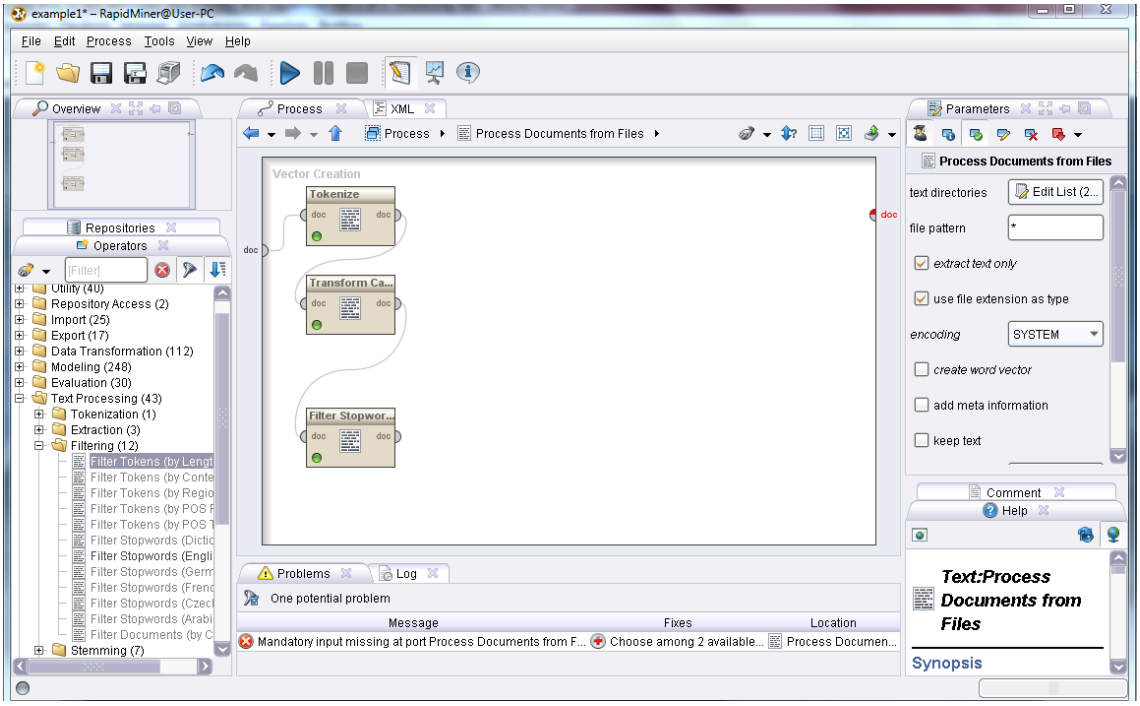
Εικόνα 30 Εισαγωγή των δύο φακέλων(neg και pos) με τα δεδομένα



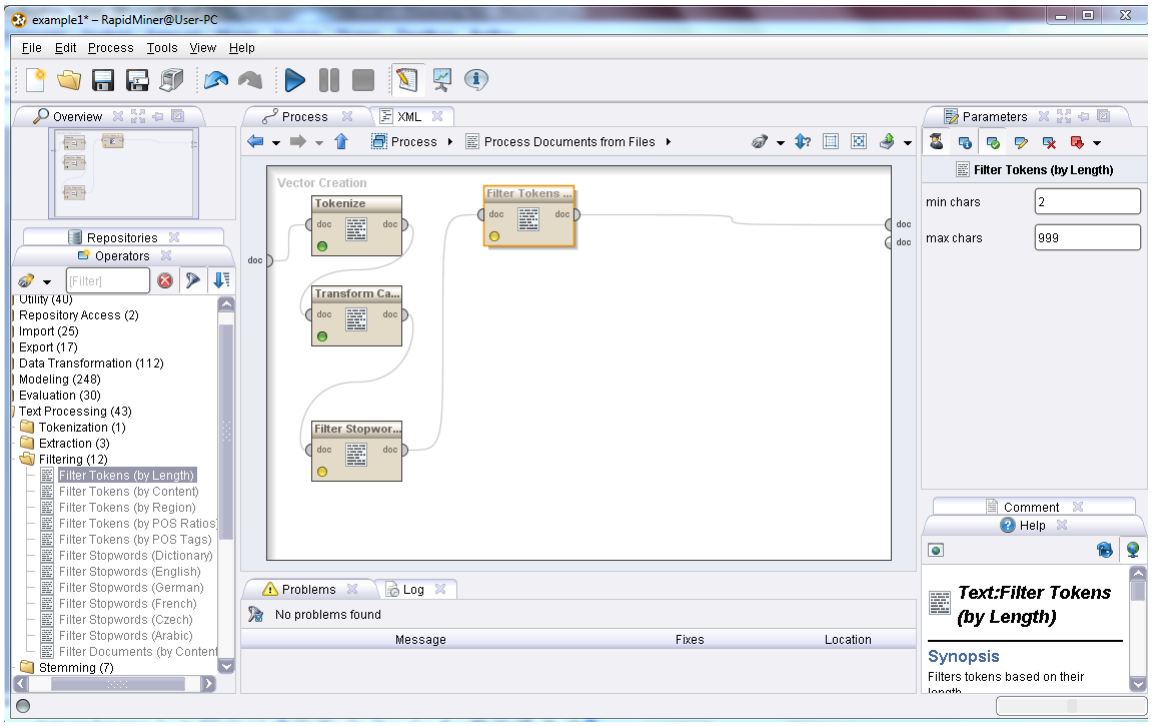
Εικόνα 31 Ο τρόπος με τον οποίο το RapidMiner παρουσιάζει τα δεδομένα. Με αριθμητική κατάταξη δείχνει μέρος του κειμένου, το όνομα της λίστας στην οποία ανήκει, τα μεταδεδομένα, την τοποθεσία των μεταδεδομένων και την ημερομηνία τροποποίησής τους.



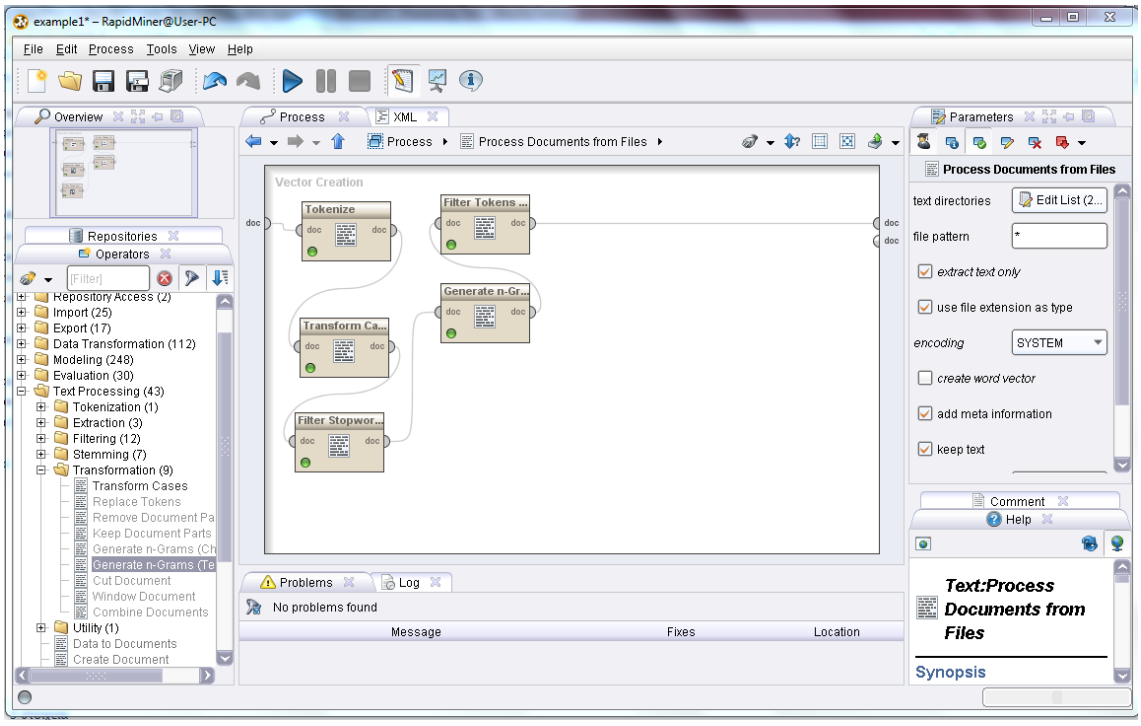
Εικόνα 32 Η εισαγωγή χειριστή για την δημιουργία λέξεων



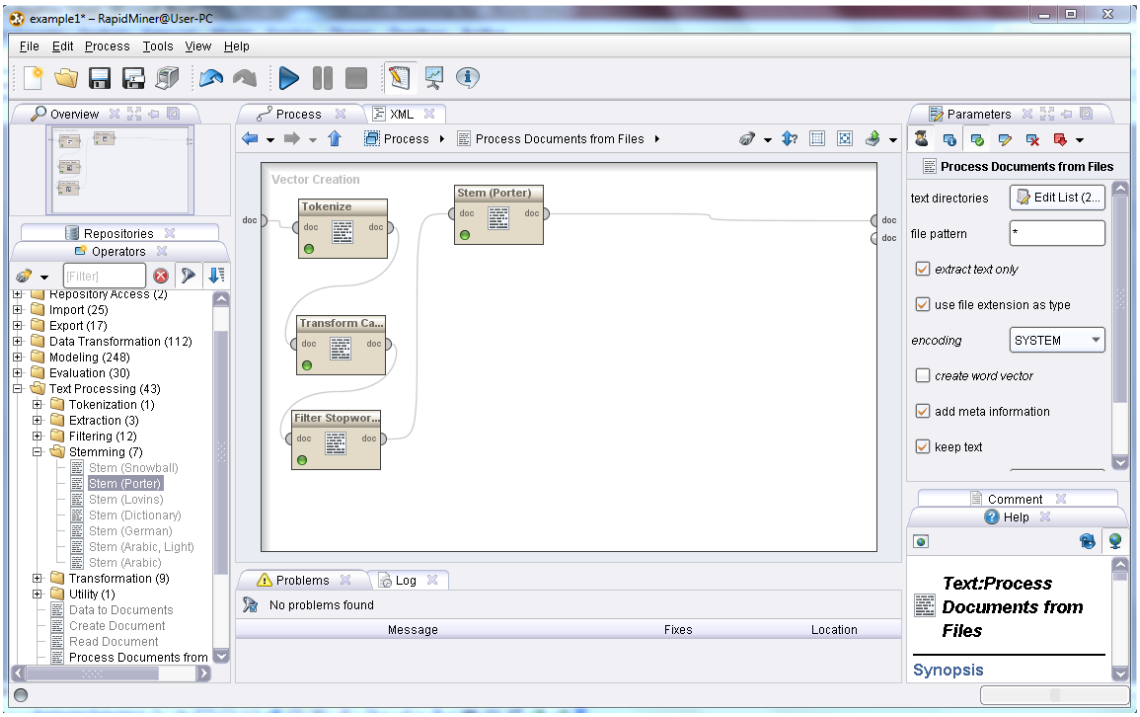
Εικόνα 33 Εισαγωγή χειριστών για μετατροπή όλων των γραμμάτων σε μικρά και η απομάκρυνση των stop words



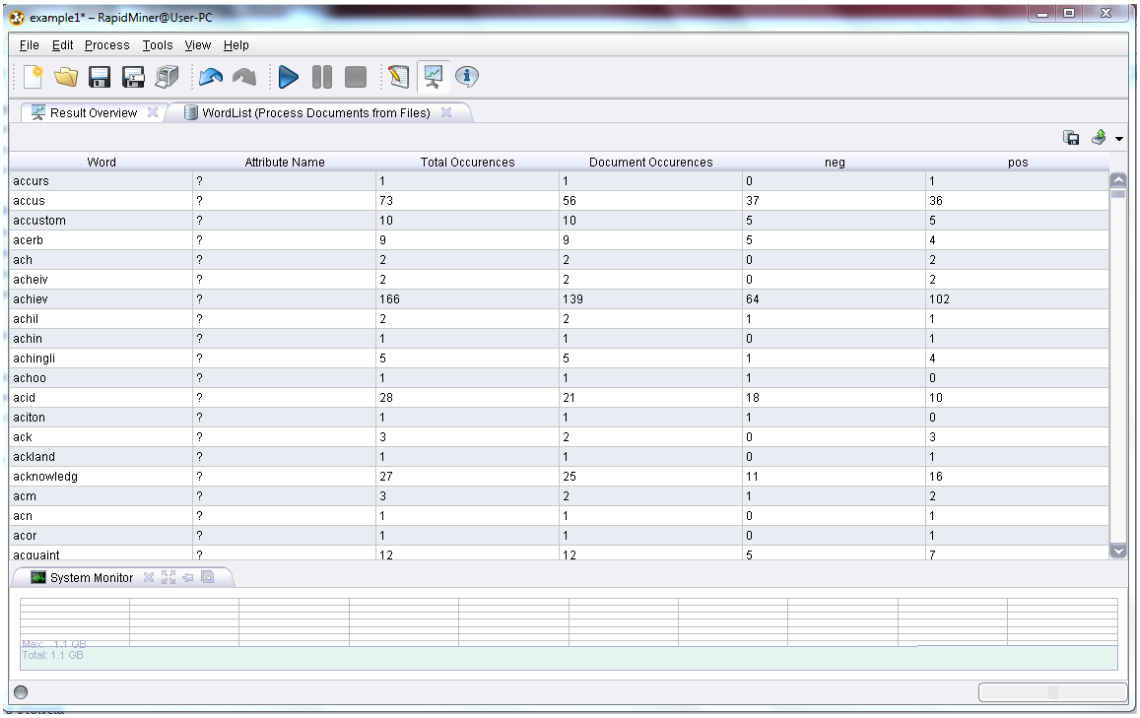
Εικόνα 34 Εισαγωγή χειριστή για απομάκρυνση λέξεων με βάση το μήκος τους που το καθορίζουμε εμείς



Εικόνα 35 Εισαγωγή χειριστή για τη δημιουργία n-grams

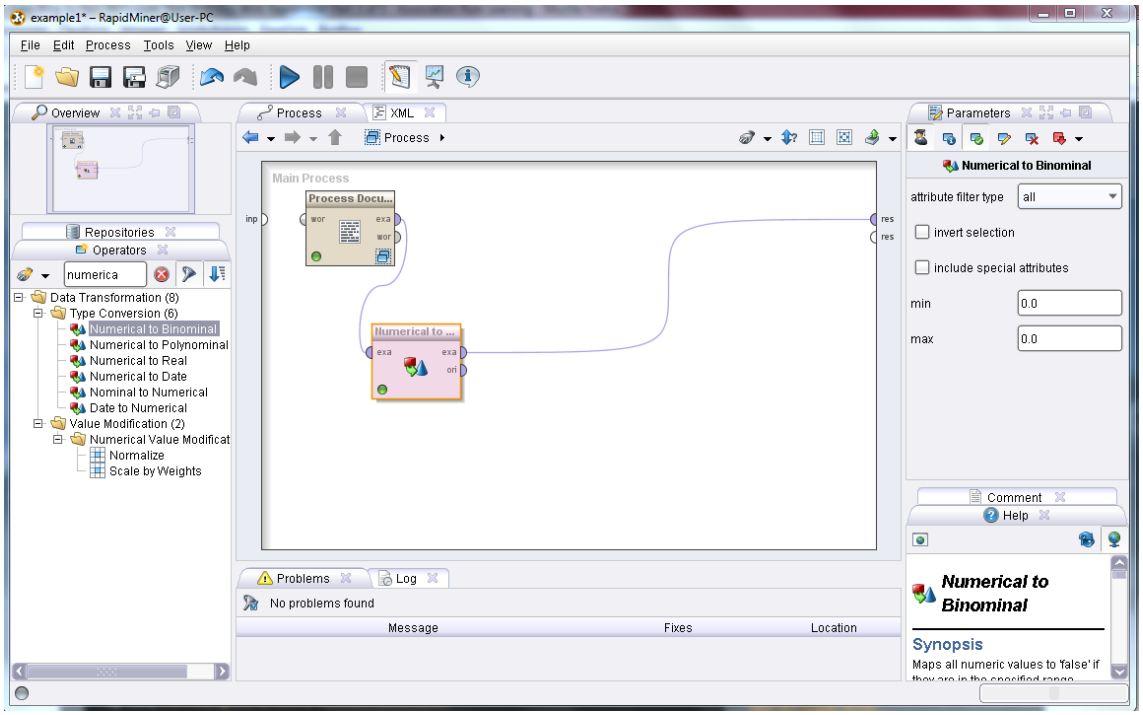


Εικόνα 36 Σ' αυτήν την περίπτωση έχουμε αφαιρέσει τους τελευταίους δύο χειριστές (Filter Tokens by Length operator και Generate n-Grams operator) και στη θέση τους βάλαμε τον χειριστή που είναι υπεύθυνος για την εύρεση της ρίζας των λέξεων

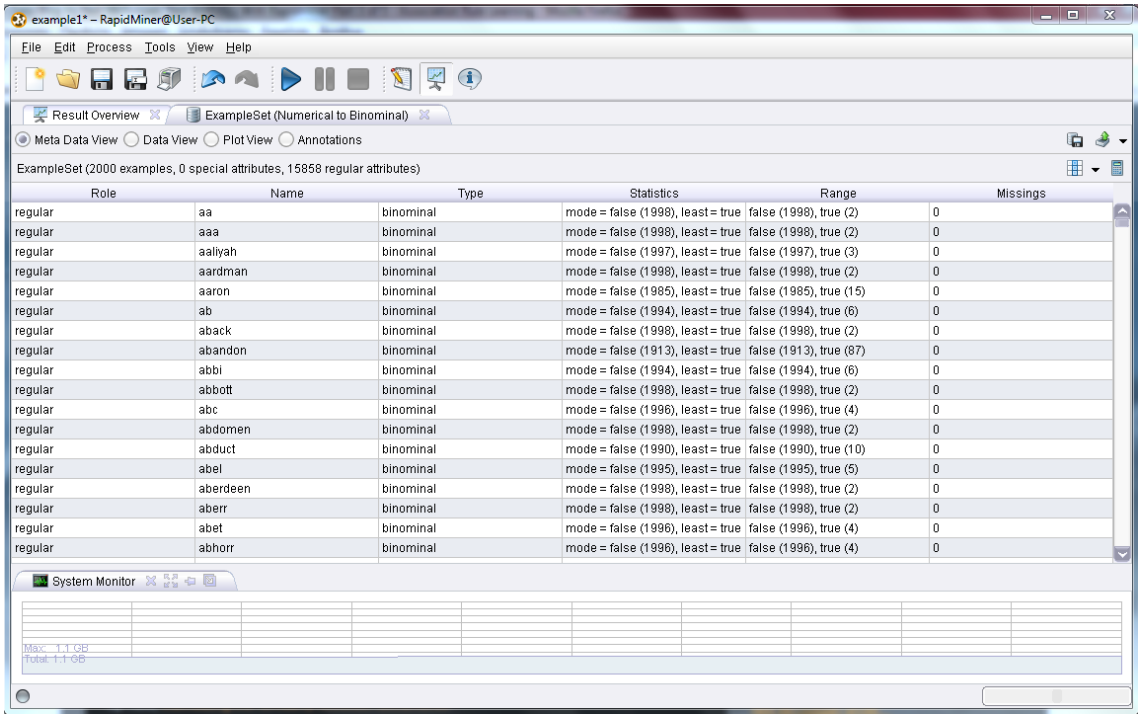


Word	Attribute Name	Total Occurences	Document Occurences	neg	pos
accurs	?	1	1	0	1
accus	?	73	56	37	36
accustom	?	10	10	5	5
acerb	?	9	9	5	4
ach	?	2	2	0	2
acheiv	?	2	2	0	2
achiev	?	166	139	64	102
achil	?	2	2	1	1
achin	?	1	1	0	1
achingli	?	5	5	1	4
achoo	?	1	1	1	0
acid	?	28	21	18	10
aciton	?	1	1	1	0
ack	?	3	2	0	3
ackland	?	1	1	0	1
acknowledg	?	27	25	11	16
acm	?	3	2	1	2
acn	?	1	1	0	1
acor	?	1	1	0	1
acquaint	?	12	12	5	7

Εικόνα 37 Εδώ μπορούμε να δούμε πόσες φορές εμφανίζεται κάθε λέξη σε κάθε κείμενο, στους φακέλους, στο έγγραφο κτλ

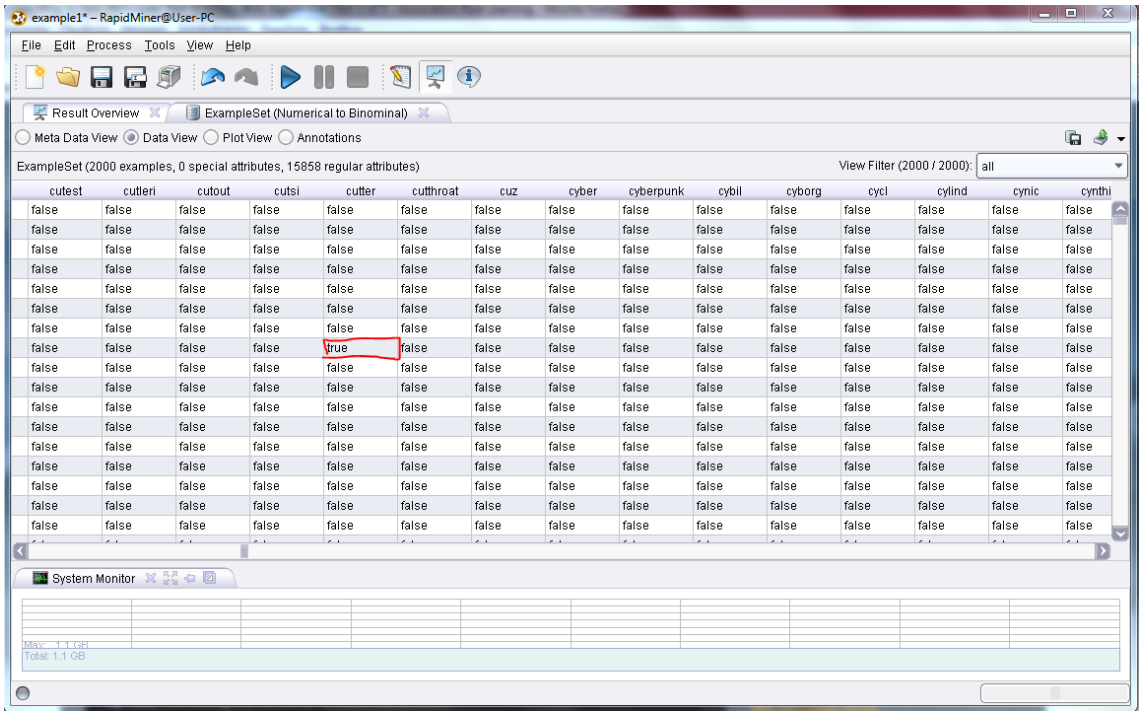


Εικόνα 38 Μετατρέποντας τις αριθμητικές τιμές σε δυαδικές

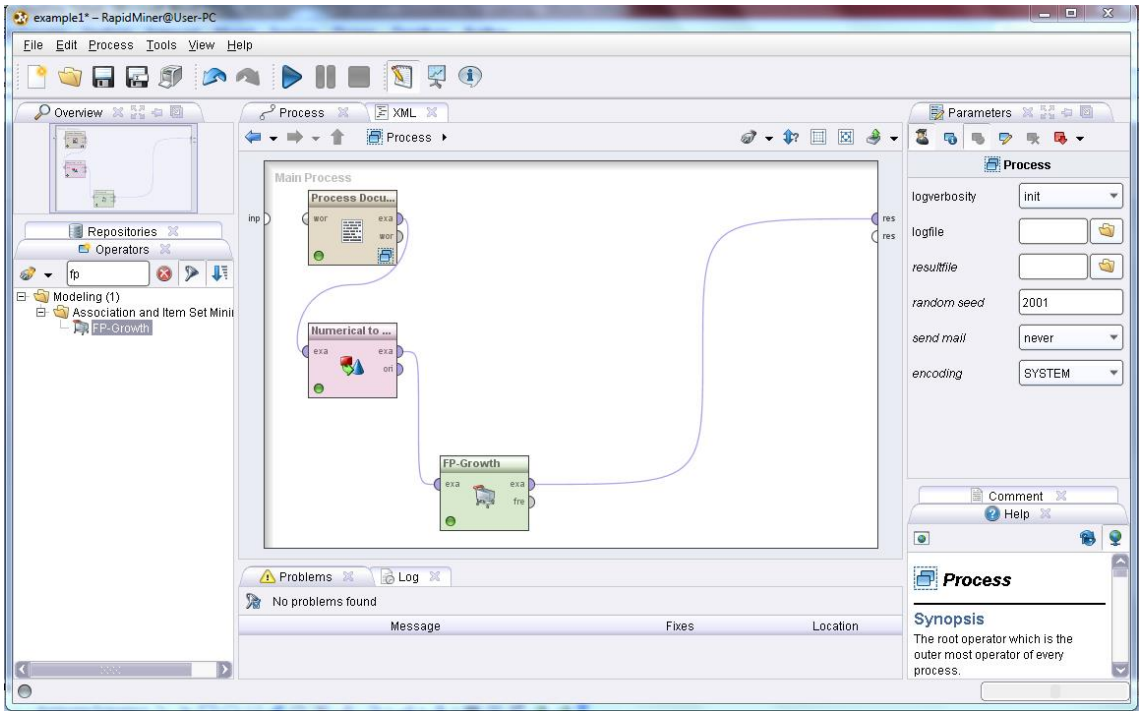


Role	Name	Type	Statistics	Range	Missings
regular	aa	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	aaa	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	aaliyah	binominal	mode = false (1997), least = true	false (1997), true (3)	0
regular	aardman	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	aaron	binominal	mode = false (1985), least = true	false (1985), true (15)	0
regular	ab	binominal	mode = false (1994), least = true	false (1994), true (6)	0
regular	aback	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	abandon	binominal	mode = false (1913), least = true	false (1913), true (87)	0
regular	abbi	binominal	mode = false (1994), least = true	false (1994), true (6)	0
regular	abbott	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	abc	binominal	mode = false (1996), least = true	false (1996), true (4)	0
regular	abdomen	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	abduct	binominal	mode = false (1990), least = true	false (1990), true (10)	0
regular	abel	binominal	mode = false (1995), least = true	false (1995), true (5)	0
regular	aberdeen	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	aberr	binominal	mode = false (1998), least = true	false (1998), true (2)	0
regular	abet	binominal	mode = false (1996), least = true	false (1996), true (4)	0
regular	abhorr	binominal	mode = false (1996), least = true	false (1996), true (4)	0

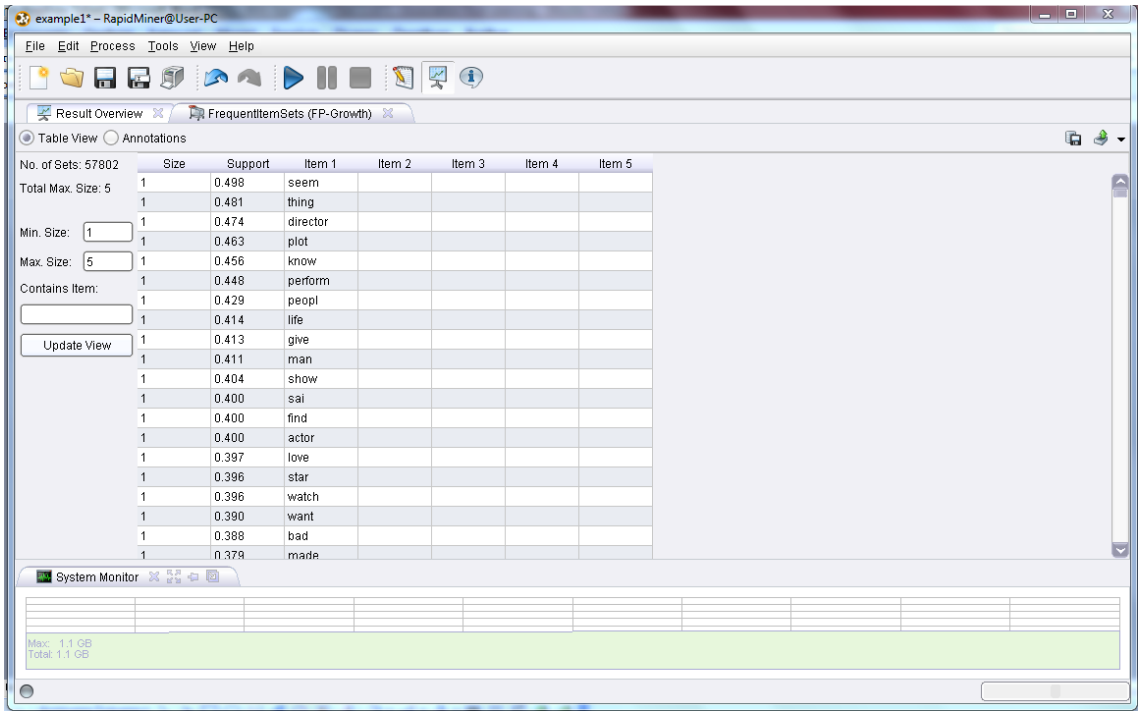
Εικόνα 39 Μετατρέποντας τις αριθμητικές τιμές σε δυαδικές (αποτέλεσμα 1)



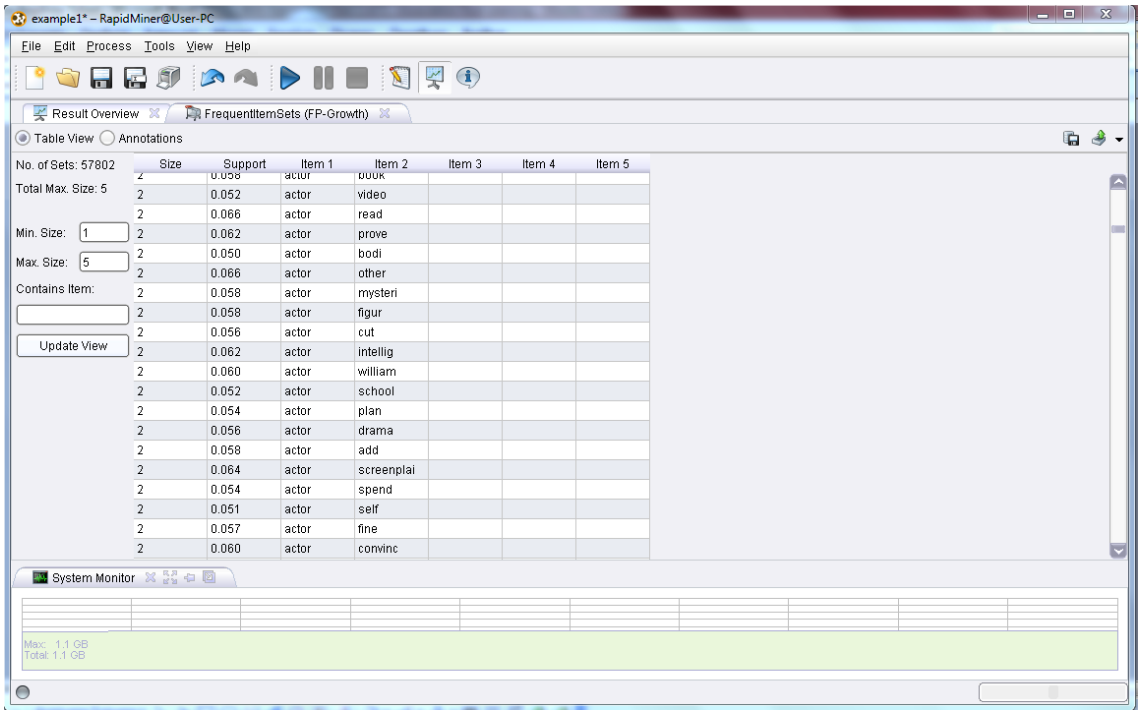
Εικόνα 40 Μετατρέποντας τις αριθμητικές τιμές σε δυαδικές (αποτέλεσμα 2)



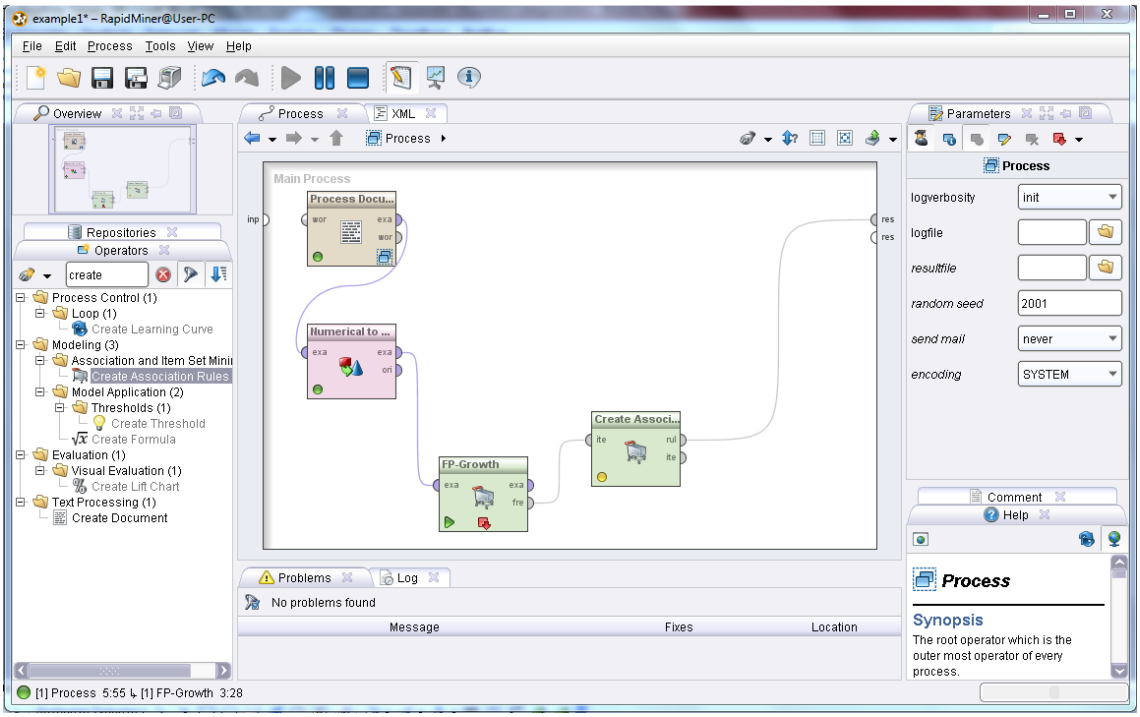
Εικόνα 41 Χρησιμοποιώντας τον αλγόριθμο Fp-Gowth για την εμφάνιση των συχνοτήτων των λέξεων



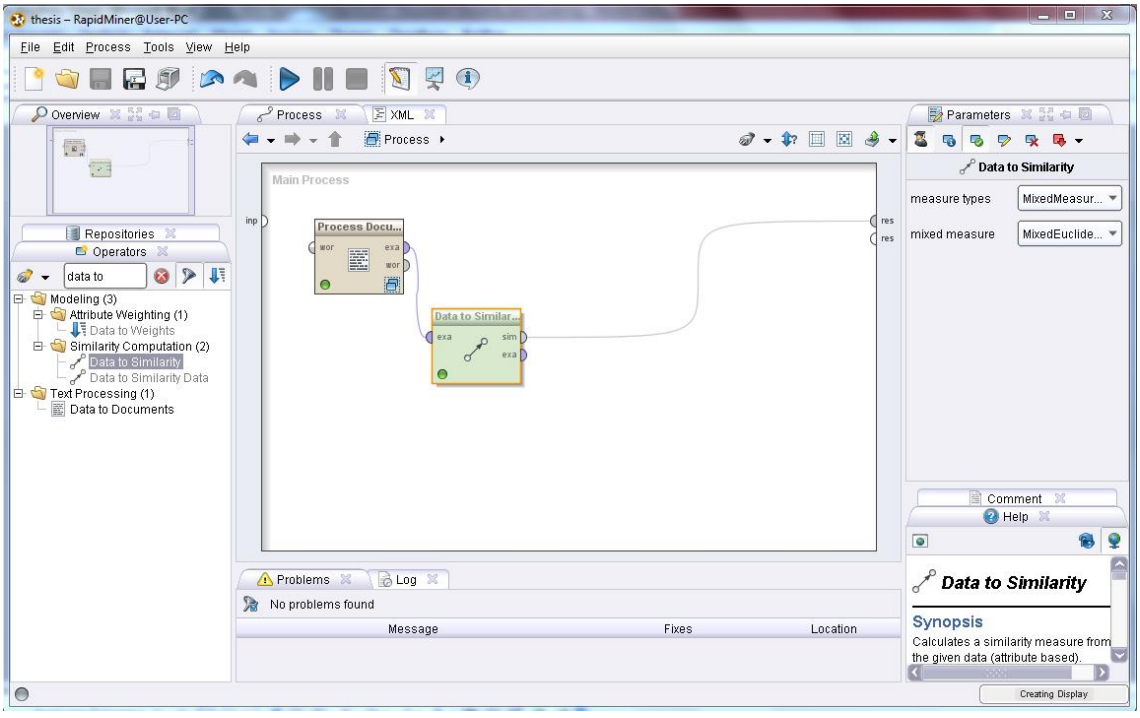
Εικόνα 42 Η συχνότητα της κάθε λέξης



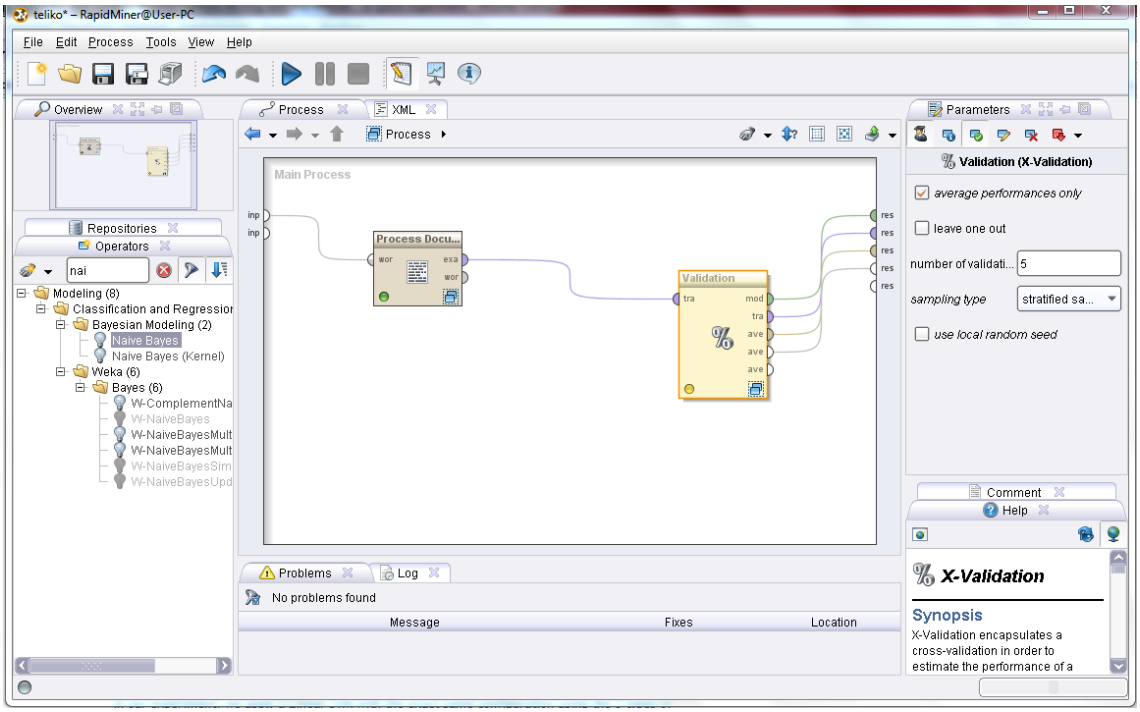
Εικόνα 43 Οι συχνότητες λέξεων που συσχετίζονται



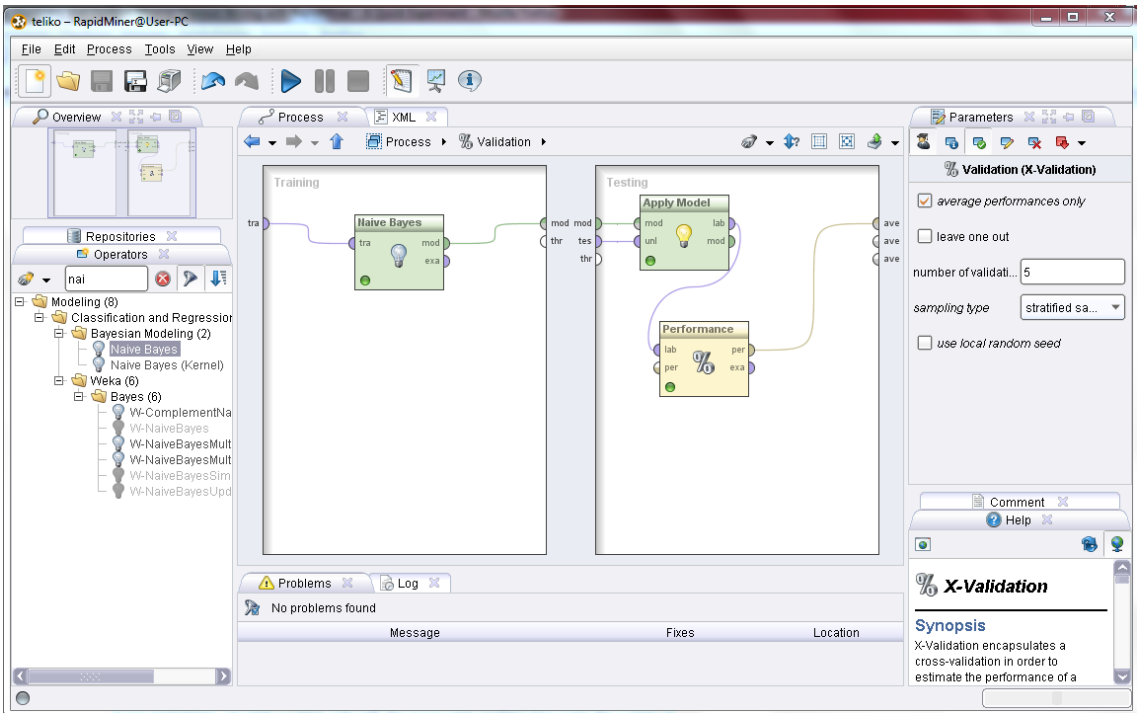
Εικόνα 44 Ο Generate Association Rules operator χρησιμοποιείται για τη δημιουργία κανόνων συσχέτισης



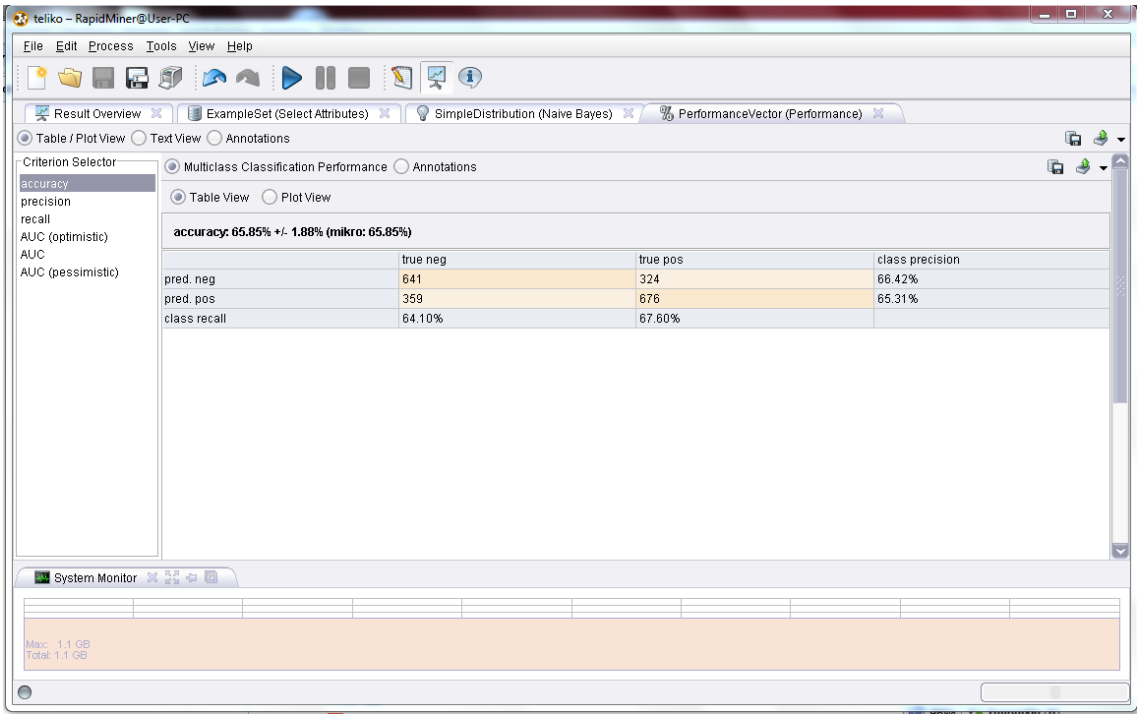
Εικόνα 45 Βρίσκοντας όμοια έγραφα



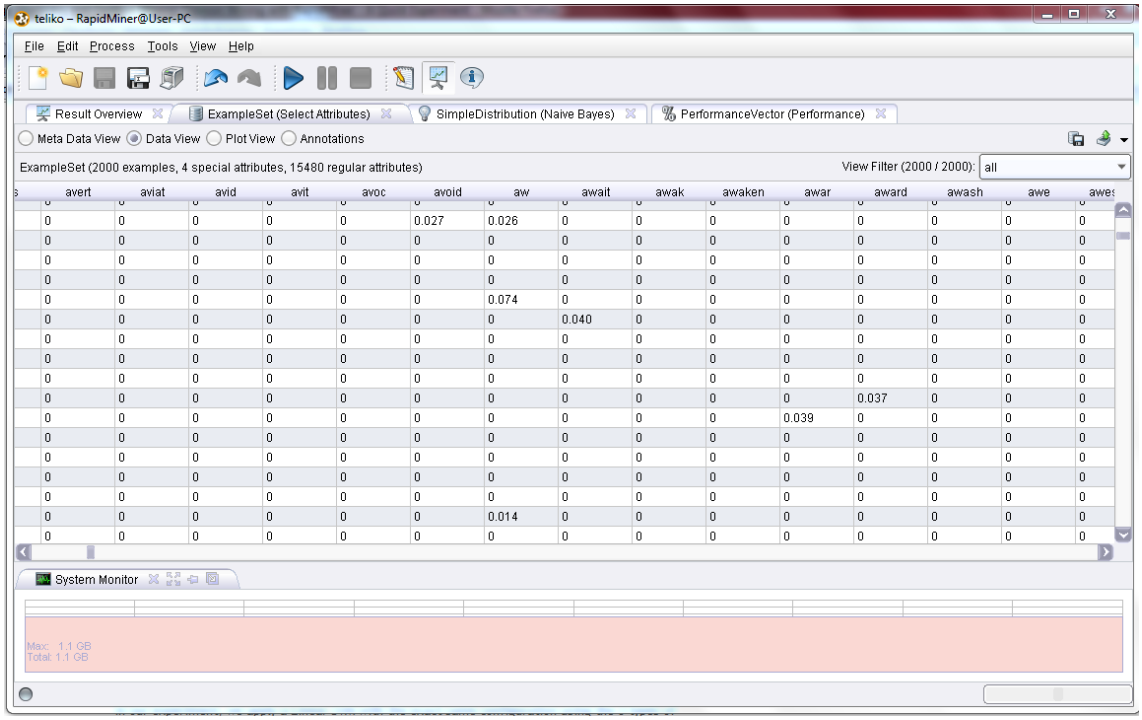
Εικόνα 46 Εισαγωγή του Validation operator



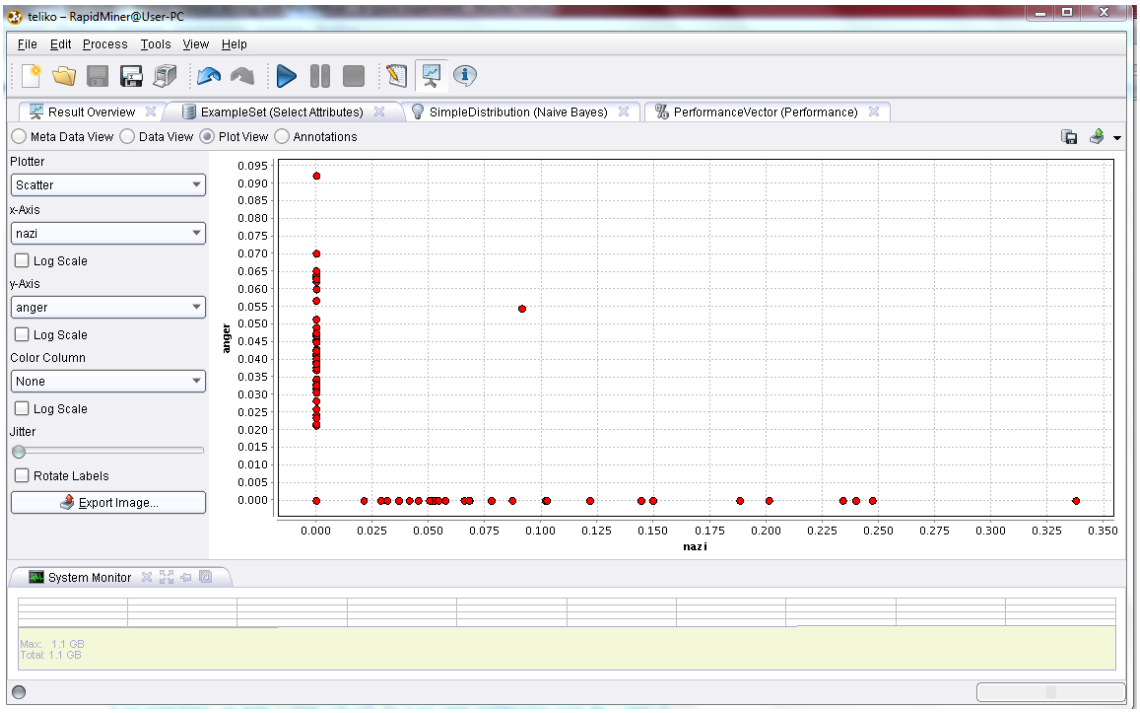
Εικόνα 47 Εισαγωγή υπόλοιπων χειριστών μαζί και του Naive Bayes



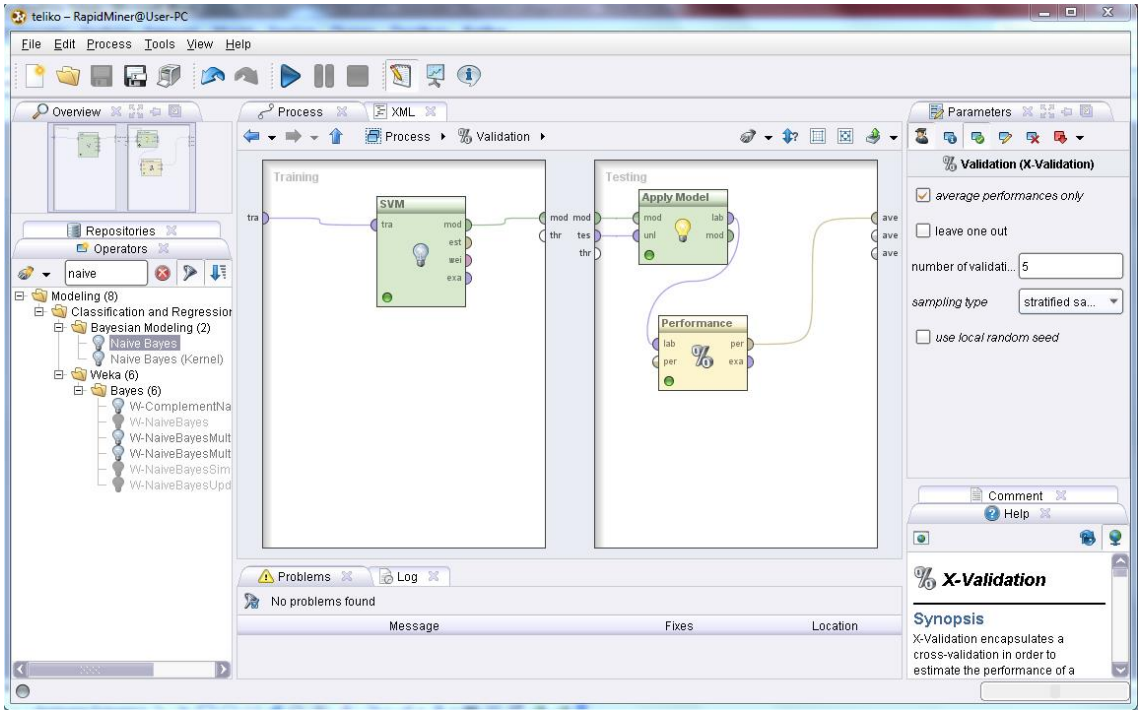
Εικόνα 48Αποτελέσματα Naive Bayes



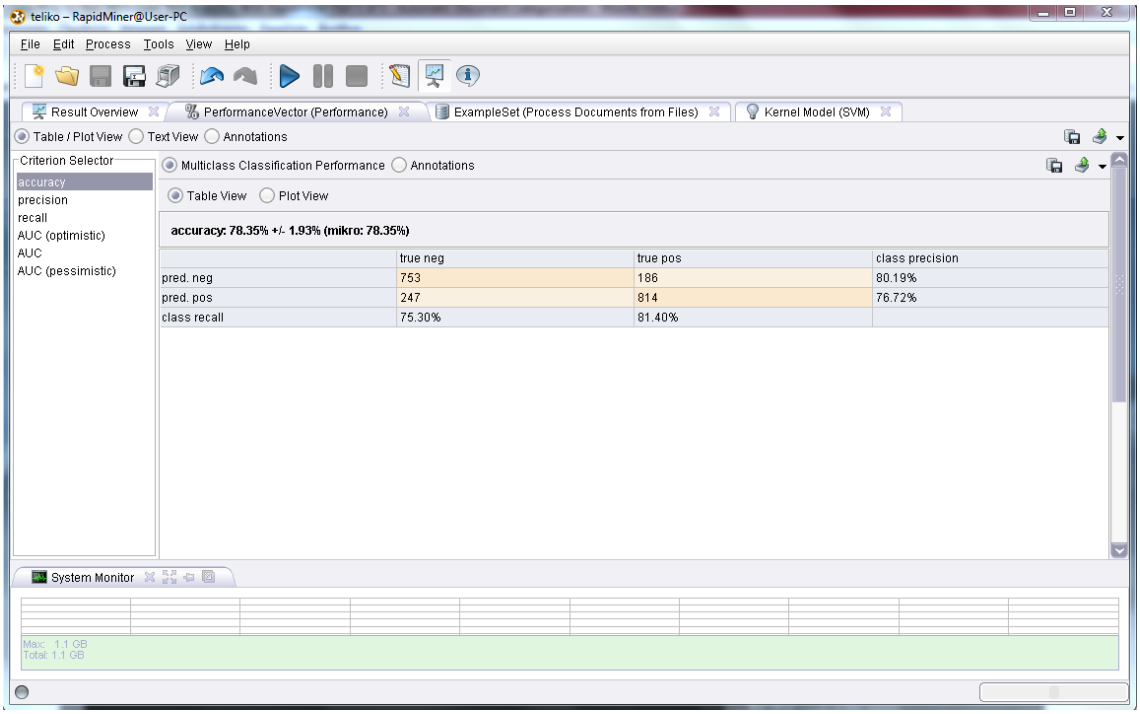
Εικόνα 49 Αποτελέσματα συχνότητας λέξεων μετά την εφαρμογή του Naive Bayes



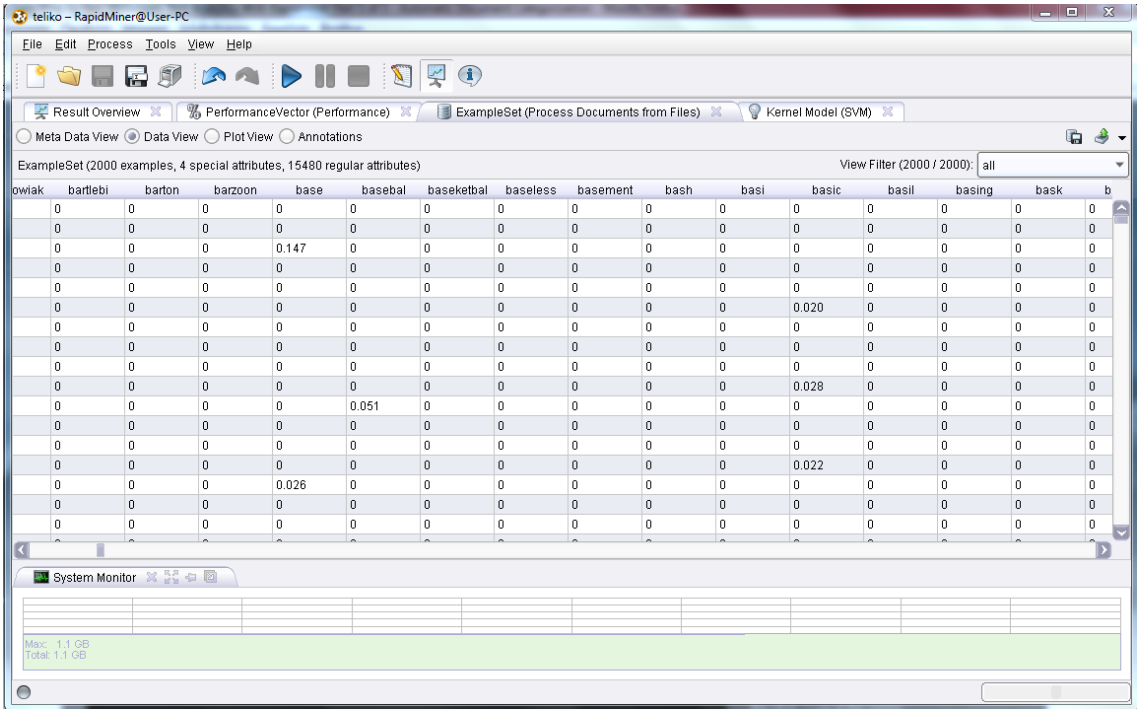
Εικόνα 50 Παράδειγμα εμφάνισης των λέξεων anger και nazi μαζί στα έγγραφα. Παρατηρούμε ότι μόνο σε ένα έγγραφο εμφανίζονται μαζί.



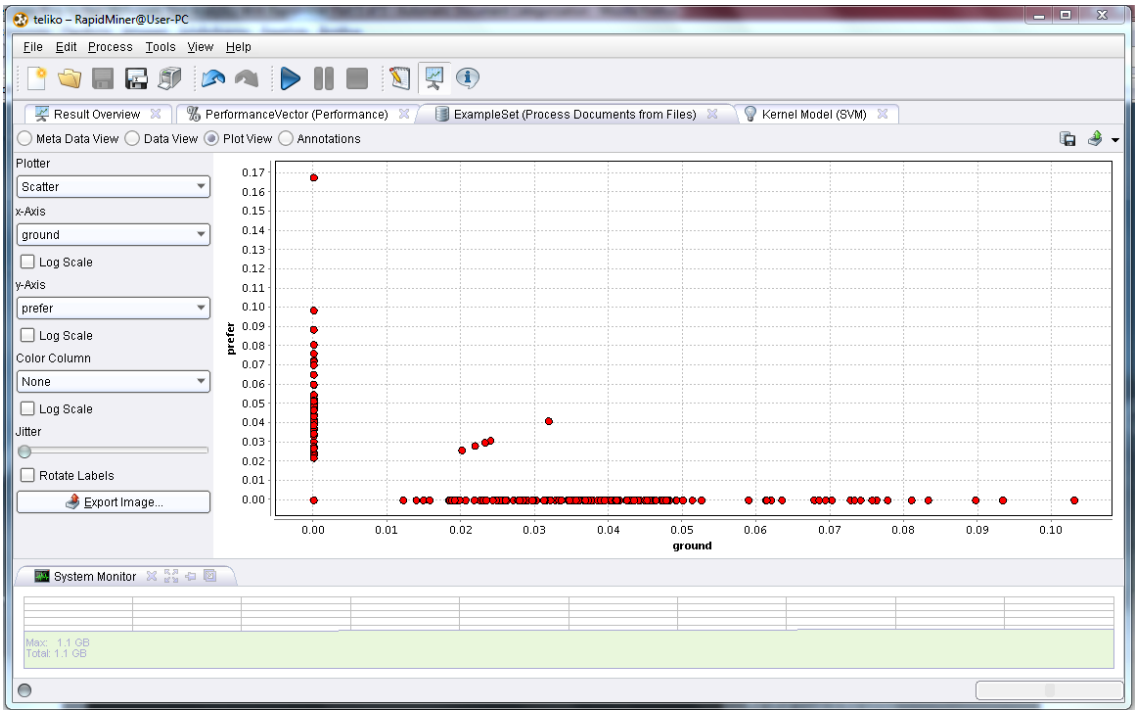
Εικόνα 51 Εισαγωγή του SVM



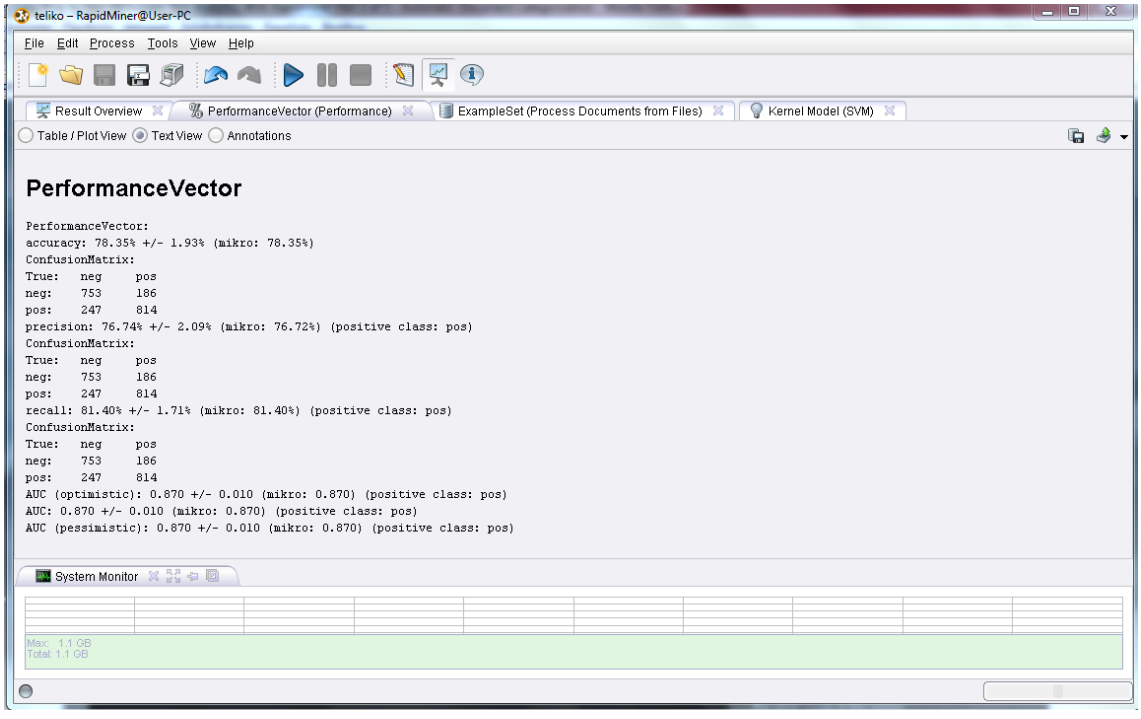
Εικόνα 52 Αποτελέσματα του SVM



Εικόνα 53 Η συχνότητα των λέξεων μετά την εφαρμογή του SVM



Εικόνα 54 Τα έγγραφα στα οποία οι λέξεις ground και prefer εμφανίζονται μαζί



Εικόνα 55 Η απόδοση του χειριστή πιο αναλυτικά