Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τμήμα Πληροφορικής
Προγράμματα Μεταπτυχιακών Σπουδών
«Πληροφορική και Επικοινωνίες»
«Διαδίκτυο και Παγκόσμιος Ιστός»

Event detection and affective analysis using Twitter texts: The case of the U.S. presidential elections of 2016

Εργασία στο μάθημα:
Εξόρυξη και ανάκτηση πληροφορίας στον Παγκόσμιο Ιστό

Φοιτητές:

Κυνηγόπουλος Γεώργιος	**Αθανασιάδης Ιωάννης**	**Κουρουπέτρογλου Πραξιτέλης-Νικόλαος**	**Τζίμας Ραφαήλ**
ΑΕΜ:631	ΑΕΜ:607	ΑΕΜ:629	ΑΕΜ:6

Διδάσκουσα καθηγήτρια:
Βακάλη Αθηνά

Ιούνιος 2016

Περιεχόμενα

1 Μοντέλα αναπαράστασης δεδομένων

1.1 Σάκοι χαρακτηριστικών

Για την αναπαράσταση του περιεχομένου μιας συλλογής εγγράφων, καθένα εκ των οποίων μπορεί να περιέχει ή όχι ένα αντιστοιχισμένο συναίσθημα, χρησιμοποιήθηκε η δομή του σάκου χαρακτηριστικών. Πρόκειται για δομή δισδιάστατου πίνακα μέσω της οποίας διατηρούμε για κάθε έγγραφο χρήσιμες πληροφορίες για τα χαρακτηριστικά που περιλαμβάνει και την κατηγορία στην οποία ενδεχομένως ανήκει. Ειδικότερα, κάθε γραμμή του πίνακα αντιστοιχεί σε ένα έγγραφο, όλες οι στήλες πλην της τελευταίας αντιστοιχούν σε χαρακτηριστικά που ενυπάρχουν στη συλλογή των εγγράφων, ενώ η τελευταία στήλη υπάρχει για να αναπαριστά την κατηγορία κάθε εγγράφου. Η γενική μορφή του σάκου παρουσιάζεται στον Πίνακα 1.

Πίνακας 1. Σάκος χαρακτηριστικών

	Χαρακτηριστικό 1	Χαρακτηριστικό …	Χαρακτηριστικό N	Κατηγορία εγγράφου
Έγγραφο 1				
Έγγραφο …				
Έγγραφο K				

Πρακτικά, η κλάση του project που υλοποιεί σε βασικό επίπεδο την ανωτέρω δομή είναι η AbstractBagOfFeatures. Η κλάση αυτή περιλαμβάνει έναν δισδιάστατο πίνακα που αναπαριστά το σάκο, ο οποίος εκτός των δεικτών σε γραμμές και στήλες που είναι ακέραιοι αριθμοί δέχεται ακέραιες τιμές και στα κελιά του. Δεδομένου ότι στη συλλογή των εγγράφων υπάρχουν χαρακτηριστικά και κατηγορίες που είναι συνήθως συμβολοσειρές, κάθε αντικείμενο της κλάσης αυτής περιλαμβάνει δύο maps, έναν για τα χαρακτηριστικά και έναν για τις κατηγορίες. Οι maps αυτοί είναι αντικείμενα της κλάσης LinkedHashMap και δέχονται ως κλειδιά διαφορετικές συμβολοσειρές και ως τιμές μοναδικούς κωδικούς για καθεμία από αυτές τις συμβολοσειρές. Με τον τρόπο αυτό διατηρούνται σε μορφή συμβολοσειράς τα χαρακτηριστικά και οι κατηγορίες των εγγράφων της συλλογής στους maps και μέσω των τιμών στους maps που αντιστοιχούν σε αυτές τις συμβολοσειρές καθίσταται εφικτή η δεικτοδότηση στον σάκο.

Σε ένα τέτοιο σάκο χαρακτηριστικό μπορεί να είναι μία λέξη, ένας συνδυασμός δύο διαδοχικών λέξεων (bi-gram), ένα emoticon, ένα σημείο στίξης όπως τα «!» και «?», ενώ κατηγορία είναι πάντα μία λέξη. Ειδικότερα, ως λέξεις ο σάκος διατηρεί ουσιαστικά, ρήματα, επίθετα και επιρρήματα, ενώ ως δυάδες μόνο τους εξής συνδυασμούς: <επίρρημα><επίθετο> και <επίρρημα><επίρρημα>. Αυτό διότι τα επιρρήματα και τα επίθετα είναι πιο περιγραφικά μέρη του λόγου ως προς τα συναισθήματα σε σχέση με τα υπόλοιπα και συναντώνται γραμματικά στην αγγλική γλώσσα σε αυτή τη διάταξη. Συμπληρωματικά, η κλάση αυτή περιλαμβάνει τη

μέθοδο getHeader() για την ανάκτηση της επικεφαλίδας του πίνακα με τις ονομασίες των χαρακτηριστικών και τις κατηγορίες καθώς επίσης και τη μέθοδο getDataRow() για την ανάκτηση του περιεχομένου μίας γραμμής του σάκου, δίνοντας ως παράμετρο τον αριθμό της γραμμής. Τέλος, η abstract κλάση σάκου περιλαμβάνει και δύο abstract μεθόδους, την shouldFeatureBeIncludedInbag() που δέχεται ως παράμετρο συμβολοσειρά και την getClassValueForDataRow() που δέχεται ως παράμετρο αριθμό γραμμής.

Δεδομένου ότι οι σάκοι που περιλαμβάνουν και κατηγορικές ετικέτες για τα έγγραφα τα οποία αναπαριστούν διαφέρουν από σάκους που δεν περιλαμβάνουν τέτοια πληροφορία την ανωτέρω κλάση επεκτείνουν με διαφορετικό τρόπο οι κλάσεις LabeledBagOfFeatures και UnlabeledBagOfFeatures. Για παράδειγμα, η πρώτη διατηρεί και δύο επιπλέον maps σχετικούς με τη διαχείριση των κατηγοριών των εγγράφων, ενώ στη δεύτερη περίπτωση κάτι τέτοιο δεν υφίσταται. Εξάλλου, καθεμία από αυτές υλοποιεί διαφορετικά τις προαναφερθείσες abstract μεθόδους. Στην πρώτη κλάση από τις ανωτέρω η εξαγωγή ενός χαρακτηριστικού από το σάκο καθίσταται εφικτή μόνο όταν το χαρακτηριστικό εμφανίζεται πάνω από ένα κατώτατο όριο συχνότητας το οποίο είναι προκαθορισμένο, ενώ στη δεύτερη δεν ισχύει κάτι ανάλογο. Ομοίως, η τιμή της κατηγορίας που εξάγεται για ένα οποιοδήποτε έγγραφο από το σάκο είναι μία από ένα πεπερασμένο γνωστό πλήθος στην πρώτη περίπτωση, ενώ στη δεύτερη είναι πάντοτε ο χαρακτήρας «?».

1.2 Μοντέλο θεμάτων

Για τον εντοπισμό θεμάτων αξιοποιήθηκε το Mallet (http://mallet.cs.umass.edu), ένα σχετικό εργαλείο που έχει αναπτυχθεί σε γλώσσα Java, το οποίο είναι ενσωματωμένο στο project ως βιβλιοθήκη (.jar). Λαμβάνοντας υπόψη αυτό το γεγονός, για την υλοποίηση του συγκεκριμένου εγχειρήματος δημιουργήθηκε ένα πακέτο Java το οποίο αξιοποιεί αυτή τη βιβλιοθήκη και περιλαμβάνει συνολικά 10 κλάσεις. Δύο εξ αυτών είναι από τις σημαντικότερες κλάσεις του πακέτου καθώς μέσω αυτών διεξάγεται η ανάλυση των αναδυόμενων θεμάτων που εντοπίζονται σε δοθέντα κείμενα από το Mallet. Πρόκειται για τις κλάσεις: Topic και MalletTopics.

Η κλάση Topic αναπαριστά ένα θέμα, όπως αυτό προκύπτει από το εργαλείο Mallet. Χρησιμοποιείται δηλαδή για να μπορούμε να διαχειριζόμαστε με τον ίδιο τρόπο καθένα από τα αποτελέσματα του Mallet, έκαστο εκ των οποίων αντιστοιχεί σε ένα θέμα. Ειδικότερα, αναπαριστούμε κάθε εντοπισμένο θέμα ως ξεχωριστό αντικείμενο Java αυτής της κλάσης και καλώντας τη μέθοδο setWord(), δίδοντας της ως παραμέτρους λέξη και συχνότητα, αποθηκεύονται τα πληροφοριακά στοιχεία ενός θέματος σε ένα αντικείμενο LinkedHashMap (το οποίο αποτελεί μεταβλητή της κλάσης Topic). Πέραν αυτής, υπάρχουν και άλλες μέθοδοι με τις οποίες μπορούμε να αντλήσουμε πληροφορίες σχετικά με το θέμα. Παραδείγματος χάριν, μπορούμε να ανακτήσουμε το σύνολο των λέξεων του θέματος ή τη συχνότητα που αντιστοιχεί σε μία λέξη.

Η κλάση MalletTopics, η οποία χρησιμοποιείται για τον εντοπισμό θεμάτων μέσω του εργαλείου Mallet, δημιουργεί και διαχειρίζεται αντικείμενα της κλάσης Topic. Για τη δημιουργία ενός στιγμιότυπου της καλείται ο constructor της με τις εξής τέσσερις παραμέτρους: α) αριθμός θεμάτων, β) αριθμός λέξεων για κάθε θέμα, γ) αριθμός επαναλήψεων του αλγορίθμου, δ) επιλογή για εμφάνιση ή μη συχνοτήτων. Στη συνέχεια καλείται η μέθοδος run() δίνοντας σε αυτή ως παραμέτρους το αρχείο εισόδου (το οποίο περιλαμβάνει τα tweets για χρονικό διάστημα διάρκειας δύο ωρών) και το αρχείο εξόδου στο οποίο καταγράφονται τα αποτελέσματα. Σε κάθε γραμμή του αρχείου των αποτελεσμάτων καταγράφονται ο αριθμός του θέματος, η κατανομή του σε σχέση με τα υπόλοιπα θέματα που αντλήθηκαν από το ίδιο αρχείο, οι λέξεις που το περιγράφουν μαζί με τη συχνότητα εμφάνισης τους στα tweets.

2 Μέθοδοι επεξεργασίας και υπολογισμού

2.1 Ανάκτηση ανεπεξέργαστων tweets και αποθήκευση στη MongoDB

Στη συγκεκριμένη εργασία επικεντρωθήκαμε στην ανάλυση των tweets από τους τρεις υποψηφίους των προεδρικών αμερικανικών εκλογών και αναζητήσαμε tweets που αναφέρουν τα εξής κανάλια τους στο Twitter: realDonaldTrump, HillaryClinton, και BernieSanders.

Για την εύρεση των κατάλληλων tweets έχει γίνει χρήση του Streaming API του Twitter (https://dev.twitter.com/streaming/overview), από το development section του Twitter (https://dev.twitter.com/). Το συγκεκριμένο API παρέχει στους developers το 1% της ροής δεδομένων – tweets από το σύνολο των σχετικών statuses που αναρτώνται στο Twitter. Για να χρησιμοποιήσει το συγκεκριμένο API κάθε μηχανικός λογισμικού πρέπει να δημιουργήσει τα κατάλληλα διαπιστευτήρια ως Twitter developer (στο https://apps.twitter.com/) ώστε να του δοθούν τα κατάλληλα κλειδιά για να χρησιμοποιεί το συγκεκριμένο API. Απαραίτητη προϋπόθεση για αυτό είναι ο μηχανικός λογισμικού να διαθέτει ήδη ένα λογαριασμό στο Twitter.

Προκειμένου να χρησιμοποιηθεί το παραπάνω API στο project, το οποίο βασίστηκε στη γλώσσα προγραμματισμού Java, αξιοποιήθηκε το Twitter4j (http://twitter4j.org). Το Twitter4j είναι μία ανεπίσημη Java βιβλιοθήκη για το Twitter API που δημιουργήθηκε με στόχο την ενσωμάτωση του Twitter API σε εφαρμογές βασισμένες στη Java. Ειδικότερα, δημιουργήθηκαν στο project οι κατάλληλες κλάσεις για ανάκτηση tweets από τη ροή του Twitter, οι οποίες αξιοποιούν δυνατότητες της εν λόγω βιβλιοθήκης. Για παράδειγμα, τα tweets από τη ροή του Twitter ανακτήθηκαν μέσω της κλάσης TwitterStreaming, όπου χρησιμοποιούνται δυνατότητες του Twitter4J αφού ελεγχθεί πρώτα η εγκυρότητα των στοιχείων διαπίστευσης.

Για την καλύτερη και πιο εστιασμένη διαλογή tweets από τη ροή του Twitter πραγματοποιούνταν ένα αρχικό φιλτράρισμα κατά την ανάκτηση δεδομένων από τη ροή, βασιζόμενο στην ποιότητά τους. Τα κριτήρια φιλτραρίσματος που επιλέχθηκαν ήταν τα εξής:

- Tweets μόνο στα αγγλικά
- Tweets με στόχο μόνο τους υποψηφίους (realDonaldTrump, HillaryClinton, BernieSanders)
- Tweets με πάνω από 4 λέξεις
- Αποφυγή tweets από twitter bots.

Το αποτέλεσμα αυτής της ανάκτησης ήταν κατά το διάστημα μεταξύ 24/05/2016 4:00 τοπική ώρα και 24/05/2016 4:00 τοπική ώρα να συλλεχθούν 1,107,246 tweets.

Μόλις ολοκληρώθηκε η συγκεκριμένη διαλογή των αρχείων αποθηκεύτηκαν σε μία NoSQL βάση δεδομένων, τη MongoDB (http://www.mongodb.com), τοποθετώντας τα tweets σε κατάλληλο collection "rawTweetsForAnalysis", εντός μίας βάσης δεδομένων με ονομασία "dbTweetsForAnalysis". Για τη συγκεκριμένη βάση, χρησιμοποιήθηκε η έκδοση 3.0 της MongoDB και έγινε χρήση συμπίεσης WiredTiger (www.wiredtiger.com) με τον αλγόριθμο snappy για καλύτερη διαχείριση των tweets.

2.2 Προεπεξεργασία ανεπεξέργαστων tweets

Η προεπεξεργασία ανεπεξέργαστων κειμένων εγγράφων, στην περίπτωση μας tweets, πραγματοποιήθηκε μέσω των κλάσεων Java που υπάρχουν στο πακέτο textprocessing και δημιουργήθηκαν για το σκοπό αυτό. Ειδικότερα, μία βασική προσέγγιση για το στάδιο της προεπεξεργασίας ήταν η εκτέλεση της μεθόδου preprocess της κλάσης Preprocessor, δίδοντας της ως παραμέτρους ένα αρχείο εισόδου, ένα αρχείο εξόδου και μία δυαδική τιμή η οποία σηματοδοτούσε την ύπαρξη ή μη κατηγορικής ετικέτας συναισθήματος σε κάθε κείμενο του αρχείου εισόδου. Λαμβάνοντας υπόψη τη δυαδική αυτή τιμή, η preprocess καλούσε με τη σειρά της μία εκ δύο μεθόδων της κλάσης StringUtils, δηλαδή την getTextPreprocessedWithTags() εφόσον το αρχείο εισόδου περιελάμβανε τέτοιες συναισθηματικές ετικέτες ή διαφορετικά αντί αυτής την getTextPreprocessedWithoutTags() όταν ίσχυε το αντίθετο.

Και οι δύο προαναφερθείσες μέθοδοι της StringUtils περιλαμβάνουν στην αρχή της εκτέλεσης τους μία κλήση στη μέθοδο doDefaultPreprocessing() της ίδιας κλάσης. Μέσω της τελευταίας διεξάγεται η βασική προεπεξεργασία κειμένου, η οποία συνίσταται: α) στον καθαρισμό του από θόρυβο μέσω κανονικών εκφράσεων για τη διαχείριση ειδικών συμβόλων, β) το μετασχηματισμό των negations αλλά και των emoticons, και γ) τη διατήρηση ή απαλοιφή πιθανών retweets και mentions, ανάλογα με τις παραμέτρους που δίδονται κάθε φορά στη μέθοδο. Οι μετασχηματισμοί των emoticons από την αρχική τους μορφή σε αριθμητική επιλέχθηκε καθώς σύμβολα όπως οι παρενθέσεις που περιέχονται συνήθως σε αυτά είναι ασύμβατα με εργαλεία επεξεργασίας φυσικής γλώσσας, ιδίως κατά τη φάση προσδιορισμού των λημμάτων που εμπεριέχουν τα υπό επεξεργασία φυσικής γλώσσας κείμενα.

Μετά την ολοκλήρωση της βασικής προεπεξεργασίας σε επόμενο στάδιο διεξάγονταν επεξεργασία φυσικής γλώσσας αξιοποιώντας τόσο την ειδική για το σκοπό αυτό βιβλιοθήκη Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/) όσο και τα

streams σε συνδυασμό με τα lambda expressions που προστέθηκαν στην έκδοση 8 της Java. Αναλυτικότερα, διαχωρίζονταν το δοθέν κείμενο σε προτάσεις, κάθε πρόταση σε λήμματα και φιλτράρονταν όσα λήμματα είτε αποτελούσαν διευθύνσεις ιστοσελίδων στον παγκόσμιο ιστό είτε ταυτίζονταν με κάποια από τις λεγόμενες stop words. Στα λήμματα που απέμεναν ελέγχονταν αν περιείχαν πολλαπλές διαδοχικά φορές τον ίδιο χαρακτήρα, οπότε αν αυτό ίσχυε και συγχρόνως δεν αποτελούσαν λέξεις ενός έγκυρου corpus τότε πραγματοποιούνταν προσπάθεια μείωσης των πολλαπλών διαδοχικών εμφανίσεων του, έως ότου βρεθεί αντίστοιχη, έγκυρη λέξη στο corpus μικρότερου μήκους. Σε περίπτωση που δεν εντοπίζονταν τελικά τέτοια λέξη, το λήμμα διατηρούνταν στη αρχική του μορφή, καθώς για παράδειγμα θα μπορούσε να αφορά μια συντομογραφία (abbreviation) που χρησιμοποιείται κατά κόρον σε γραπτά μηνύματα (π.χ. lol).

Τέλος, οι αλγόριθμοι που αναπτύχθηκαν για την ανάγνωση κειμένων από αρχεία και την επεξεργασία φυσικής γλώσσας μπορούν να εκτελούνται παράλληλα για κάθε γραμμή εισόδου, πρόταση ή λήμμα σε επεξεργαστές με πολλαπλούς πυρήνες, δημιουργώντας ένα διαφορετικό νήμα (thread) επεξεργασίας για κάθε γραμμή, πρόταση ή λήμμα και αναθέτοντας καθένα εξ αυτών σε διαφορετικό πυρήνα. Αυτό επιτυγχάνεται καθώς οι αλγόριθμοι έχουν σχεδιαστεί με βάση το νέο πακέτο stream της πρόσφατης έκδοσης 8 της Java, αξιοποιώντας ορισμένες από τις εγγενείς δυνατότητες παράλληλης επεξεργασίας συλλογών δεδομένων που προσφέρει. Τέτοιες είναι οι μέθοδοι filter() και map(), τις οποίες αξιοποιήσαμε κατάλληλα.

2.3 Ανακάλυψη οντοτήτων

Στο συγκεκριμένο στάδιο διεξήχθη επεξεργασία των tweets για την ανακάλυψη οντοτήτων και μερών του λόγου από τα κείμενα που συλλέχθηκαν από το Twitter. Η επεξεργασία διεξήχθη επαναληπτικά για καθένα από τα μη επεξεργασμένα tweets της συλλογής, μέσω της κλάσης MongoDBConverterFromRawToFullJsonStructure. Η προεπεξεργασία των αντλούμενων tweets από τη σχετική συλλογή της βάσης δεδομένων, πραγματοποιήθηκε μέσω κλήσης σε μεθόδους της κλάσης StringUtils. Ως προς τον εντοπισμό οντοτήτων, όσες ανακαλύφθηκαν είναι εκ των παρακάτω τύπων:

1. Person
2. Location
3. Organization

Τέλος, το τελικό JSON αρχείο που αποθηκεύονταν κατά το πέρας της επεξεργασίας έχει τα εξής πεδία:

- "_id", το κλειδί του JSON document που αποδίδεται από τη βάση
- "persons", η οντότητα persons,
- "urls", τα expanded urls
- "followers", οι followers του χρήστη για το συγκεκριμένο tweet
- "hashtags", τα hashtags του tweet
- "mentions", τα mentions του tweet
- "partOfSpeech", τα μέρη του λόγου του tweet

- "organizations", η οντότητα organizations,
- "created_at", ημερομηνία δημιουργίας του tweet
- "locations", η οντότητα locations
- "text", το επεξεργασμένο κείμενο του tweet
- "refId", το Id του προτοτύπου tweet που βρίσκεται στη συλλογή των μη επεξεργασμένων tweet, για μελλοντική αναφορά
- "timestampMs", το timestamp του tweet σε χιλιοστά του δευτερολέπτου.

2.4 Εξαγωγή από τη MongoDB συλλογών επεξεργασμένων tweets ανά δίωρο

Για τη ανάδειξη αναδυόμενων θεμάτων και στο μέρος του έργου που αφορούσε στην κατηγοριοποίηση των κειμένων ανάλογα με το εμπεριέχον συναίσθημα χρειάστηκε να γίνει κατάτμηση του συνολικού χρόνου λήψης δεδομένων σε μικρότερα χρονικά διαστήματα και ο διαχωρισμός των επεξεργασμένων tweets ανά χρονικό διάστημα. Έπειτα από πειραματισμούς ως κατάλληλο χρονικό εύρος για το διαχωρισμό αυτό επιλέχθηκε το χρονικό διάστημα των δύο ωρών. Αναπτύχθηκε λοιπόν το πρόγραμμα MongoBatchExportFilteredPerTimeWindow το οποίο εκτελεί στο system console την εντολή mongoexport, παρέχοντας στο χρήστη ένα τμήμα της συλλογής json αρχείων επεξεργασμένων κειμένων που διατηρείται στη βάση MongoDB με ονομασία "filteredTweetsTextOnly", με βάση πάντοτε το χρονικό διάστημα που ορίζει ο χρήστης.

2.5 Ανεύρεση αναδυόμενων θεμάτων με το Mallet

Τα CSV αρχεία που δημιουργήθηκαν στο προηγούμενο βήμα (2.4), όπως έχει ήδη αναφερθεί, έχουν υποστεί αρχική προεπεξεργασία προκειμένου να αφαιρεθεί ο θόρυβος από τα ανεπεξέργαστα δεδομένα, όπως παραδείγματος χάριν οι λεγόμενες stop words. Αυτά τα αρχεία CSV αξιοποιήθηκαν για την ανεύρεση αναδυόμενων θεμάτων με το εργαλείο Mallet. Ωστόσο, για να καταστεί εφικτό κάτι τέτοιο χρειάστηκε να προηγηθεί μια επιπρόσθετη επεξεργασία στα προαναφερθέντα αρχεία προκειμένου να διαμορφωθούν άλλα, αντίστοιχα σε πλήθος, τα οποία να είναι σε συμβατά με τη μορφή εισόδου που απαιτείται από το εργαλείο. Τελικά, για το Mallet χρησιμοποιήθηκαν 36 ειδικά αρχεία όπου το καθένα περιελάμβανε κείμενα χρηστών του Twitter για διαφορετικό, διαδοχικό δίωρο, με κάποιου είδους αναφορά στις επερχόμενες προεδρικές εκλογές των ΗΠΑ. Εντός κάθε τέτοιου αρχείου η μορφή κάθε γραμμής ήταν: <id><tab><X><tab><text>. Το id είναι μία μοναδική τιμή που αντιστοιχεί σε κάθε tweet, ενώ το δεύτερο πεδίο χρησιμοποιείται ως ετικέτα εγγράφου και για την εφαρμογή αυτή δεν θα χρησιμοποιηθεί. Έτσι, θέσαμε την τιμή του σε X, χωρίς να έχει κάποια σημασία στη διαδικασία. Τέλος, το "text" περιλαμβάνει το επεξεργασμένο tweet (κείμενο).

Στη συνέχεια, αφού τα αρχεία εισόδου απέκτησαν την επιθυμητή μορφή, χρησιμοποιήθηκε η βιβλιοθήκη Mallet προκειμένου να εντοπίσουμε πιθανά αναδυόμενα θέματα. Πρέπει να σημειωθεί ότι με την χρήση αυτής της βιβλιοθήκης έγινε επιπλέον επεξεργασία προκειμένου να απομακρυνθούν λέξεις, οι οποίες

δυσχεραίνουν αυτή τη διαδικασία. Αναλυτικότερα, απομακρύνθηκαν σύμβολα (π.χ. '!') και ορισμένες επιπλέον κοινότυπες λέξεις (π.χ. «people»), σύμφωνα με τις οδηγίες χρήσης του εργαλείου. Επιπρόσθετα, επειδή παρατηρήθηκε ότι τα ονόματα των υποψηφίων (π.χ. «realdonaldtrump») εμφανίζονταν σχεδόν σε όλα τα θέματα ως οι δημοφιλέστερες λέξεις, αποφασίστηκε να μην χρησιμοποιηθούν διότι ουσιαστικά δεν βοηθούσαν στην αναγνώριση θεμάτων. Αυτό το μοντέλο χρησιμοποιήθηκε για κάθε αρχείο ξεχωριστά.

Αξίζει ακόμη να σημειωθεί ότι έγινε παραμετροποίηση στο εργαλείο όσον αφορά το εξαγόμενο πλήθος των θεμάτων ανά χρονική περίοδο, το πλήθος των λέξεων ανά θέμα, και η χρονική διάρκεια κάθε περιόδου. Πραγματοποιήθηκαν πλήθος πειραμάτων προκειμένου να αποφασιστεί η βέλτιστη τιμή για καθεμία από τις τρεις κύριες μεταβλητές Ειδικότερα, δοκιμάστηκαν τα εξής: για το πλήθος των θεμάτων ανά χρονική περίοδο οι τιμές «10,5,4,3», για το πλήθος των λέξεων ανά θέμα οι τιμές «5-15» και για το χρονικό διάστημα κάθε περιόδου οι τιμές 1-4 ώρες. Οι τελικές τιμές που αποφασίστηκαν ήταν: 3 θέματα κάθε 2 ώρες με 10 λέξεις να περιγράφουν το κάθε από αυτά.

Γενικά, όπως διαπιστώθηκε από τον πειραματισμό με το εργαλείο, σε κάθε αναδυόμενο θέμα μιας χρονικής περιόδου αποδίδεται ένας συντελεστής με πεδίο τιμών από 0 έως 1, ο οποίος αντανακλά τη βαρύτητα του θέματος στη συγκεκριμένη χρονική περίοδο. Επιπρόσθετα, το άθροισμα των συντελεστών βαρύτητας όλων των αναδυόμενων θεμάτων για δεδομένη χρονική περίοδο ισούται με 100%. Κατά τη φάση της εκτέλεσης πειραμάτων με τις ανωτέρω παραμέτρους προέκυψε ότι στις περισσότερες περιπτώσεις για κάθε χρονικό διάστημα δύο ωρών δεν εντοπίζονταν παραπάνω από τρία σημαντικά θέματα. Στα περισσότερα δε δίωρα εμφανίζονταν ένα μόνο θέμα να λαμβάνει συντελεστή βάρους μεγαλύτερο από 95%. Δεδομένου ότι όλοι οι συντελεστές μιας χρονικής περιόδου έχουν άθροισμα 100%, αυτό σημαίνει ότι συχνά ένα αναδυόμενο θέμα είναι κατεξοχήν κυρίαρχο συγκριτικά με τα υπόλοιπα. Για το λόγο αυτό αποφασίστηκε το εξαγόμενο πλήθος αναδυόμενων θεμάτων να είναι μόνο τρία για κάθε χρονικό διάστημα (δίωρο). Όσον αφορά το πλήθος των λέξεων ανά θέμα, επιλέχθηκε η τιμή δέκα, προκειμένου να έχουμε στη διάθεση μας την αναγκαία πληροφορία για σαφή αναγνώριση του αναδυόμενου θέματος. Το μοντέλο θεμάτων σχηματίστηκε μετά 1000 επαναλήψεις σύμφωνα με τις προαναφερθείσες τελικές τιμές για τις τρεις παραμέτρους, ακολουθώντας τις οδηγίες χρήσης του εργαλείου.

Στη συνέχεια χρησιμοποιήθηκε ένα κατώφλι με τιμή το 0.3 για το συντελεστή βαρύτητας, προκειμένου να αντλούμε μόνο τα πιο σημαντικά αναδυόμενα θέματα ανά δίωρο. Τέλος, εστιάσαμε στις λέξεις που είχαν τη μεγαλύτερη εμφάνιση σε όλα τα θέματα του τριημέρου, προκειμένου να εξετάσουμε τις διακυμάνσεις της συχνότητας με την οποία εμφανίζονται στην πάροδο του χρόνου. Επιλέχθηκαν οι έξι κορυφαίες λέξεις στη βάση σχετικής μετρικής προκειμένου να τις αναλύσουμε περαιτέρω και να καταλήξουμε σε συμπεράσματα, μέσω της απεικόνισης τους σε γράφημα.

2.6 Κατασκευή σάκων χαρακτηριστικών

Ως προς την κατασκευή των δύο ειδών σάκων χαρακτηριστικών, δηλαδή σάκων με ήδη γνωστές ή μη τις κατηγορίες για τα έγγραφα της συλλογής που αναπαριστούν εφαρμόστηκε η ίδια γενική στρατηγική με επιμέρους διαφοροποιήσεις ανά είδος. Αναλυτικότερα, υπάρχουν στο ίδιο πακέτο με τις κλάσεις που αναπαριστούν τους σάκους οι αντίστοιχες κλάσεις κατασκευαστή σάκου, μία abstract και δύο που την επεκτείνουν.

Η abstract κλάση περιέχει μόνο μία μέθοδο, την exportBagToFile η κλήση της οποίας έχει ως αποτέλεσμα την εξαγωγή ενός σάκου σε αρχείο arff, κατάλληλο για το λογισμικό WEKA. Εντός της μεθόδου αυτής καλείται μία φορά η getHeader() του σάκου για την εξαγωγή της επικεφαλίδας στο αρχείο και επαναληπτικά η getDataRow() για την εξαγωγή του περιεχομένου όλων των γραμμών του σάκου.

Η LabeledBagOfFeaturesBuilder που την επεκτείνει λειτουργεί σύμφωνα με τον παρακάτω αλγόριθμο:

Βήμα 0. Αν το αρχείο εισόδου csv δεν περιέχει επεξεργασμένα έγγραφα αλλά ανεπεξέργαστα, προώθησε το στον προεξεπεργαστή και περίμενε μέχρι να ολοκληρώσει την εκτέλεση του για να μεταβείς στο Βήμα 1.

Βήμα 1. Σάρωσε το αρχείο εισόδου με μορφή csv για να κατασκευάσεις τους maps που χρησιμεύουν αργότερα ως δείκτες στα κελιά του σάκου, αναγνωρίζοντας τα χαρακτηριστικά που περιλαμβάνει εσωτερικά η συλλογή εγγράφων του αρχείου αυτού καθώς και τις κατηγορίες εγγράφων που περιέχει.

Βήμα 2. Δημιούργησε ένα σάκο N γραμμών (αριθμός εγγράφων στη συλλογή) και K στηλών (συνολικό πλήθος χαρακτηριστικών στη συλλογή συν ένα για την καταχώρηση της κατηγορίας κάθε εγγράφου) και αρχικοποίησε τον, θέτοντας τιμή μηδέν σε όλα τα κελιά του.

Βήμα 3. Σάρωσε εκ νέου από την αρχή για δεύτερη φορά το ίδιο αρχείο εισόδου και συμπλήρωσε τα κελιά του σάκου με τιμές είτε τις συχνότητες εμφάνισης των χαρακτηριστικών ανά έγγραφο είτε αντί αυτών την τιμή ένα όταν ένα χαρακτηριστικό υπάρχει τουλάχιστον μία φορά σε ένα έγγραφο, συμπληρώνοντας και την κατηγορία που αντιστοιχεί σε κάθε έγγραφο.

Βήμα 4. Εξήγαγε από το σάκο μόνο όσα χαρακτηριστικά εμφανίζουν συχνότητα εμφάνισης πάνω από ένα κατώτατο όριο για μια οποιαδήποτε κατηγορία εγγράφων, δημιουργώντας ένα αρχείο σε μορφοποίηση arff κατάλληλο για χρήση ως σύνολο εκπαίδευσης αλγορίθμων μηχανικής μάθησης από το λογισμικό WEKA.

Αντίστοιχα, η UnlabeledBagOfFeaturesBuilder που επεκτείνει τον ίδιο abstract κατασκευαστή σάκου λειτουργεί σύμφωνα με τον επόμενο αλγόριθμο:

Βήμα 1. Σάρωσε το αρχείο εισόδου με μορφή arff που αποτελεί σύνολο εκπαίδευσης για αλγορίθμους μηχανικής μάθησης και εμπεριέχει σάκο χαρακτηριστικών και ανάκτησε τα χαρακτηριστικά αυτά καθώς και τα ονόματα των κατηγοριών των εγγράφων που αναπαριστά ο σάκος.

Βήμα 2. Σάρωσε το αρχείο εισόδου με μορφή csv για να μετρήσεις το πλήθος των γραμμών του.

Βήμα 3. Δημιούργησε ένα σάκο N γραμμών (αριθμός εγγράφων στη συλλογή) και K στηλών (συνολικό πλήθος χαρακτηριστικών στη συλλογή συν ένα για την καταχώρηση της κατηγορίας κάθε εγγράφου) και αρχικοποίησε τον, θέτοντας τιμή μηδέν σε όλα τα κελιά του.

Βήμα 4. Σάρωσε εκ νέου από την αρχή για δεύτερη φορά το ίδιο αρχείο εισόδου με μορφή csv και συμπλήρωσε τα κελιά του σάκου με τιμές είτε τις συχνότητες εμφάνισης των χαρακτηριστικών ανά έγγραφο είτε αντί αυτών την τιμή ένα όταν ένα χαρακτηριστικό υπάρχει τουλάχιστον μία φορά σε ένα έγγραφο, συμπληρώνοντας ένα «?» στη στήλη για την κατηγορία κάθε εγγράφου.

Βήμα 5. Εξήγαγε από το σάκο όλα τα χαρακτηριστικά, δημιουργώντας ένα αρχείο σε μορφοποίηση arff κατάλληλο για χρήση ως σύνολο δοκιμής αλγορίθμων μηχανικής μάθησης από το λογισμικό WEKA.

2.7 Κατασκευή μοντέλων ανάλυσης συναισθημάτων με το WEKA

Σε αυτό το κομμάτι της εργασίας έγινε η ανάλυση των δεδομένων που είχαμε στο εργαλείο μηχανικής μάθησης WEKA. Σε πρώτο στάδιο έγινε πειραματισμός με αφετηρία το πρωτεύον αρχείο που λάβαμε από τη βοηθό του μαθήματος, το οποίο περιείχε 2681 μη επεξεργασμένα κείμενα tweets καθένα από τα οποία συνοδεύονταν από μία ετικέτα συναισθήματος, από τις 14 που αναφέρθηκαν προηγουμένως. Το αρχείο αυτό υπέστη εκτενή προεπεξεργασία με τη διαδικασία που περιγράφηκε παραπάνω και συγχρόνως μέσω κατάλληλου προγραμματιστικού κώδικα σε γλώσσα Java ήρθε σε συμβατή μορφή arff ώστε να γίνεται αποδεκτό από το λογισμικό μηχανικής μάθησης και εξόρυξης δεδομένων WEKA.

Έπειτα, σε πρώτο στάδιο, δοκιμάστηκαν οι αλγόριθμοι Naïve Bayes, SVM, Random Forest και Multilayer Perceptron ώστε να διαπιστωθεί σε γενικές γραμμές ποιος είναι καλύτερος για το υπό εξέταση ζήτημα της ορθής απόδοσης συναισθημάτων σε κείμενο προερχόμενο από το Twitter. Στο τέλος αυτού του βήματος είχαμε μία σαφή εικόνα της αποτελεσματικότητας και αποδοτικότητας των αλγορίθμων, με τα πρώτα συμπεράσματα να υποδηλώνουν ότι τα τεχνητά νευρωνικά δίκτυα και ο απλός Naive Bayes δεν αποδίδουν τόσο ικανοποιητικά όσο τα δέντρα απόφασης και ο SVM.

Σε δεύτερο στάδιο, και έχοντας γνώση των προηγουμένων, παράχθηκε μία συλλογή από 24 αρχεία arff για την κατασκευή μοντέλων κατηγοριοποίησης συναισθημάτων με τεχνικές μηχανικής μάθησης, τα οποία διέφεραν μεταξύ τους ανά δύο σε μία από τις παραμέτρους που χρησιμοποιήθηκαν για την παραγωγή τους, οπότε αντίστοιχα διέθεταν εσωτερικά και διαφορετικά στοιχεία εκπαίδευσης για την κατασκευή μοντέλων μηχανικής μάθησης. Οι διαφορές αυτές στις παραμέτρους παραγωγής των αρχείων arff για μηχανική μάθηση διακρίνονται εξωτερικά από το ίδιο το όνομα τους, ήτοι: α) το πρώτο σημείο στο όνομα κάθε αρχείου σημαίνει αν για κάθε γραμμή του εξαγόμενου πίνακα bag of features, δηλαδή για κάθε tweet, καταγράφουμε τη συχνότητα εμφάνισης N του feature στο tweet η απλώς θέτουμε μόνο μία binary τιμή 0/1 ανάλογα με το αν υπάρχει η όχι, β) το δεύτερο σημείο υποδηλώνει αν για την

κατασκευή του αρχείου arff λαμβάνονται υπόψη τα bi-grams, και γ) το τρίτο σημείο στο όνομα κάθε αρχείου σημαίνει πόσες φορές έπρεπε να εμφανίζεται το συγκεκριμένο χαρακτηριστικό στο σύνολο του dataset ώστε να το δεχθούμε στον πινάκα bag of features. Δηλαδή, το 10 σημαίνει ότι αν η Χ λέξη δεν υπήρχε τουλάχιστον 10 φορές στο dataset δεν την ενσωματώναμε στον πινάκα bag of features ή σε μία στρατηγική που δοκιμάσαμε μεταγενέστερα το 10 σημαίνει ότι το χαρακτηριστικό αυτό πρέπει να υπάρχει ως ζεύγος με μία οποιαδήποτε ίδια κλάση τουλάχιστον δέκα φορές στη συλλογή για να το συμπεριλάβουμε στο σάκο. Η γενική μορφή του ονόματος κάθε αρχείου arff για κατασκευή μοντέλου ανάλυσης συναισθημάτων ήταν η εξής:

<BagOfFeatures>_<TRUE/FALSE>_<TRUE/FALSE>_<Number>.arff

Εν συνεχεία παρήχθησαν και αρχεία arff για κατασκευή μοντέλου μηχανικής μάθησης λαμβάνοντας υπόψη αφενός όλα τα συλλεγμένα tweets από το stream είτε παραλείποντας εξολοκλήρου tweets που περιείχαν είτε URLs είτε URLs και mentions και retweets. Καθώς η προσέγγιση της παράλειψης τέτοιων tweets δεν απέδωσε ικανοποιητικά στις δοκιμές, δε χρησιμοποιήθηκε περαιτέρω τέτοια μεθοδολογία. Σε όλες τις προαναφερθείσες προσεγγίσεις διατηρούνται στον εξαγόμενο πίνακα bag of features τα emoticons όχι με τη μορφή που περιλαμβάνονται σε ένα κείμενο, αλλά μέσω μίας ειδικής αριθμητικής τιμής για καθένα από αυτά όπως προαναφέρθηκε.

Τέλος, το τρίτο στάδιο περιλάμβανε πειράματα στη βάση του συνδυασμού όλων αυτών των αρχείων arff με τους καλύτερους αλγορίθμους μηχανικής μάθησης για το εξεταζόμενο ζήτημα, όπως αυτοί αναδείχθηκαν από τη μελέτη της βιβλιογραφίας και τον πειραματισμό μας στο πρώτο στάδιο.

3 Παράθεση και σχολιασμός αποτελεσμάτων

3.1 Αποτελέσματα ανεύρεσης αναδυόμενων θεμάτων με το Mallet

Σε αυτό το μέρος παρουσιάζονται τα αποτελέσματα της μεθοδολογίας ανεύρεσης αναδυόμενων θεμάτων. Με τη χρήση του εργαλείου Mallet βρέθηκαν τρία θέματα για κάθε χρονικό διάστημα δύο ωρών στη διάρκεια των τριών ημερών. Έτσι τελικά εντοπίστηκαν συνολικά 108 θέματα. Ωστόσο, όπως προκύπτει από το συντελεστή βαρύτητας που περιλαμβάνεται στα αποτελέσματα που εξάγονται από το Mallet, δεν κρίθηκαν όλα τα θέματα το ίδιο σημαντικά, καθώς τις περισσότερες φορές υπήρχε μόνο ένα κυρίαρχο θέμα ανά δίωρο. Για το λόγο αυτό χρησιμοποιήθηκε ως κατώφλι η τιμή 0.3 για τον συντελεστή βαρύτητας, προκειμένου να φιλτράρουμε τα πιο σημαντικά θέματα. Με αυτήν την μέθοδο αναδείχθηκαν 37 θέματα, το οποίο σημαίνει ότι μόνο σε ένα δίωρο εντοπίστηκαν παραπάνω από ένα σημαντικά θέματα. Τα αποτελέσματα είναι διαθέσιμα στη web εφαρμογή που δημιουργήθηκε για τις ανάγκες της εργασίας, αλλά δεν παρατίθενται εδώ, διότι ο όγκος τους είναι μεγάλος. Ωστόσο, θα αναφέρουμε στη συνέχεια κάποια ενδεικτικά παραδείγματα σε σχέση με τον εντοπισμό των θεμάτων αυτών.

Επίσης, ένας επιπλέον στόχος είναι η εξέταση της πορείας των θεμάτων στην πάροδο του χρόνου, προκειμένου να αναδειχθούν θέματα τα οποία συζητήθηκαν έντονα μόνο για περιορισμένο χρονικό διάστημα. Για την επίτευξη του παραπάνω εφαρμόσαμε δύο προσεγγίσεις. Στην πρώτη προσέγγιση εντοπίσαμε τις λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης στο σύνολο των tweets και στη συνέχεια εξετάσαμε τη συχνότητα που είχαν αυτές σε κάθε χρονική περίοδο. Στη δεύτερη προσέγγιση εντοπίσαμε τις λέξεις που εμφανίζονται σε περισσότερα αναδυόμενα θέματα επί του συνόλου αυτών στο τριήμερο και στη συνέχεια ακολουθήσαμε παρόμοια μεθοδολογία με την προηγούμενη προσέγγιση. Στους Πίνακες 2 και 3 μπορούμε να παρατηρήσουμε την εφαρμογή των δύο μεθοδολογιών σε σχέση με το πρώτο χρονικά θέμα που εντοπίστηκε.

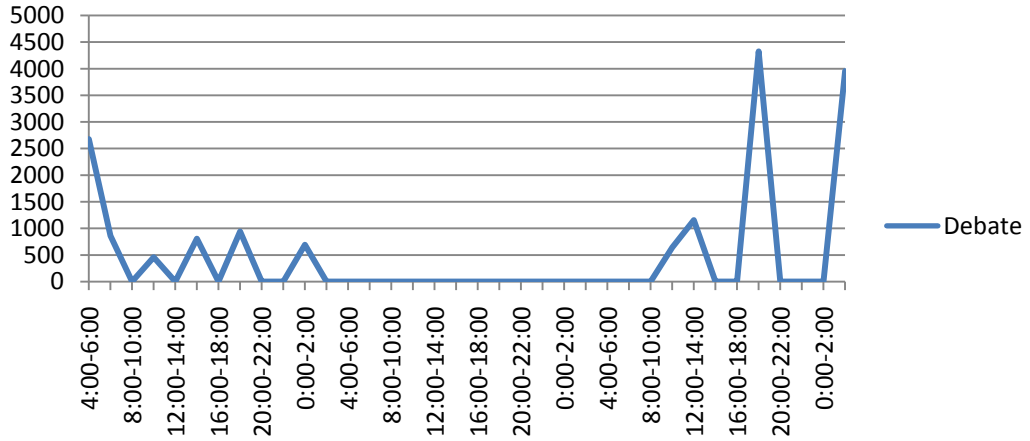
Πίνακας 2: Top-6 λέξεις - συχνότητες εμφάνισης σε όλα τα θέματα

1η προσέγγιση	debate (63700)	vote (38991)	feelthebern (24381)	cnn (22405)	california (21249)	president (20160)
θέμα 1 (συχνότητα)	2676	0	584	0	847	0

Πίνακας 3: Top-6 λέξεις - πλήθος συμπεριλήψεων στα θέματα

2η προσέγγιση	vote (53)	feelthebern (32)	debate (31)	america (28)	lie (26)	president (26)
θέμα 1 (συμπερίληψη)	0	1	1	0	0	0

Από τους Πίνακες 2 και 3 μπορούμε να παρατηρήσουμε ότι βάσει των κριτηρίων που διαλέξαμε οι περισσότερες λέξεις είναι κοινές και για τις δύο προσεγγίσεις, αλλά με διαφορετική σειρά. Στη συνέχεια, διαμορφώσαμε και εξετάσαμε το Διάγραμμα 1 για να εντοπίσουμε τις μεταβολές της συχνότητας της λέξης debate στη διάρκεια των τριών ημερών.



Διάγραμμα 1: Συχνότητα εμφάνισης της λέξης Debate στη διάρκεια των τριών ημερών

Όπως παρατηρούμε στο Διάγραμμα 1, η λέξη debate εμφανίζεται κυρίως την πρώτη και την τελευταία ημέρα. Με μια γρήγορη αναζήτηση παρατηρήθηκε ότι στις 24 Μαΐου η Hillary Clinton αρνήθηκε την πραγματοποίηση του προγραμματισμένου debate με τον Bernie Sanders στην Καλιφόρνια και για το λόγο αυτό συζητήθηκε ιδιαίτερα. Ένα σχόλιο του Sanders ήταν: «I am disappointed but not surprised by Secretary Clinton's unwillingness to debate before the largest and most important primary in the presidential nominating process» το οποίο εντοπίστηκε ως topic με τις παρακάτω λέξεις:

<debate poll primary california surprise disappointed train feelthebern foxnews secretary>

Στη συνέχεια ο Donald Trump ζήτησε από τον Sanders να πραγματοποιηθεί debate μεταξύ τους και αυτό φαίνεται από την αύξηση της συχνότητας της λέξης στις 26 Μαΐου. Εντοπίστηκαν επιπλέον θέματα, η λεπτομερής ανάλυση των οποίων είναι εκτός στόχων της παρούσας τεχνικής αναφοράς και για το λόγο αυτό δεν θα γίνει περαιτέρω παράθεση πληροφοριών για αυτά.

3.2 Αποτελέσματα ανάλυσης συναισθημάτων με το WEKA

Σε αυτό το σημείο παρουσιάζονται τα αποτελέσματα της κατασκευής μοντέλου για τη συναισθηματική κατηγοριοποίηση κειμένων και της αξιοποίησης του στα tweets που συλλέξαμε από τη ροή του Twitter για τις προεδρικές εκλογές των ΗΠΑ με το λογισμικό WEKA. Επισημαίνεται στο σημείο αυτό ότι σε όλα τα πειράματα χρησιμοποιήθηκε η προσέγγιση cross validation με 10 folds, για την εγκυρότερη καταγραφή της συμπεριφοράς των μοντέλων σε μη ταξινομημένα κείμενα. Επίσης, τονίζεται ότι τα μετασχηματισμένα σε αριθμητικές τιμές emoticons διατηρούνταν πάντοτε ως χαρακτηριστικά στους σάκους καθώς εμφάνιζαν υψηλές συχνότητες στη συλλογή κειμένων που συλλέξαμε, ενώ το ίδιο δε συνέβαινε για παράδειγμα με ορισμένες ακολουθίες λέξεων (bi-grams), για τις οποίες εξετάστηκε κατά πόσο συμβάλλουν στην ακρίβεια της συναισθηματικής κατηγοριοποίησης κειμένων.

Όπως προκύπτει από τον Πίνακα 4, ήδη από τα πρώτα πειράματα ξεχώρισαν οι καλύτεροι αλγόριθμοι για την αναπαράσταση δεδομένων που κατασκευάσαμε, από τους τέσσερις τους οποίους εξετάσαμε. Παρατηρείται ότι ο απλός Naive Bayes δεν απέδωσε ικανοποιητικά ανεξάρτητα από τις παραμέτρους κατασκευής των σάκων χαρακτηριστικών και η ακρίβεια του κυμάνθηκε στο 32%. Αντίστοιχα και τα τεχνητά νευρωνικά δίκτυα δεν κατάφεραν να επιτύχουν ικανοποιητική ακρίβεια (μέγιστη 29.5%), ενώ σε αρκετά πειράματα δεν κατέστη εφικτό να ολοκληρώσουν την απαιτούμενη επεξεργασία εντός ενενήντα λεπτών που ήταν και το ανώτατο όριο που θέσαμε για αυτή. Είναι ξεκάθαρο επίσης ότι ως καλύτεροι αλγόριθμοι αναδεικνύονται ο SVM, γεγονός που συνάδει με τη βιβλιογραφία, και τα δέντρα απόφασης, με τον αντιπροσωπευτικό αλγόριθμο Random Forest, καθώς εμφανίζουν τις καλύτερες τιμές ακρίβειας κατηγοριοποίησης. Λαμβάνοντας υπόψη τα αποτελέσματα αυτά διεξήχθησαν στη συνέχεια επιπρόσθετα πειράματα, εστιάζοντας στους δύο τελευταίους αλγορίθμους και στις διαφορετικές παραμέτρους κατασκευής των σάκων χαρακτηριστικών.

Πίνακας 4. Σύγκριση όλων των αλγορίθμων

Frequencies instead of binary values in bag	Include bigrams	Lowest Total Frequency Per Feature	Naïve Bayes	Multilayer Percepton	SVM	Random Forest
No	Yes	10	32.8236	**1.5 ώρα** για 20%	37.9709	36.7027
Yes	No	10	32.0776	**1.5 ώρα** για 20%	38.0082	35.8448
Yes	Yes	1	32.5998	27.1913	36.1432	**39**.3510
Yes	Yes	10	32.0776	29.5039	**38**.0828	36.3297

Προχωρώντας στη σύγκριση αποτελεσμάτων για τους δύο καλύτερους αλγορίθμους, μέσω του Πίνακα 5 παρατηρείται ότι ο SVM επιτυγχάνει υψηλότερη ακρίβεια όσο λιγότερα χαρακτηριστικά περιλαμβάνει ο σάκος τον οποίο αξιοποιεί για την κατασκευή μοντέλου κατηγοριοποίησης (38%), ενώ το αντίθετο συμβαίνει για τον αλγόριθμο Random Forest του οποίου τα ποσοστά ακρίβειας είναι υψηλότερα όταν εξετάζει σάκους με όλα τα διαθέσιμα χαρακτηριστικά (39%). Αξίζει εδώ να σημειωθεί ότι η μείωση διαστάσεων στον σάκο χαρακτηριστικών πραγματοποιείται λαμβάνοντας υπόψη τη συχνότητα εμφάνισης κάθε χαρακτηριστικού συνολικά στη συλλογή κειμένων και διατηρώντας για τους αλγορίθμους μηχανικής μάθησης στους σάκους μόνο όσα ξεπερνούν ένα κατώτατο όριο (Lowest Total Frequency Per Feature). Αναλυτικότερα, παρατηρείται ότι η αύξηση του ορίου αυτού, η οποία συνεπάγεται μείωση των διαστάσεων του σάκου συμβάλει θετικά στην ακρίβεια που επιτυγχάνει ο αλγόριθμος SVM.

Πίνακας 5. Σύγκριση SVM & Random Forest

Frequencies instead of binary values in bag	Include bigrams	Lowest total frequency per feature in document collection	SVM	Random Forest
Yes	No	1	35.8821	**39.2391**
Yes	No	3	35.0988	37.2622
Yes	No	5	36.7400	36.7400
Yes	No	10	**38.0082**	35.8448
Yes	Yes	1	36.1432	**39.3510**
Yes	Yes	3	35.0988	35.8821
Yes	Yes	5	36.7400	37.0011
Yes	Yes	10	**38.0828**	36.3297

Συνεχίζοντας, ο Πίνακας 6 αποτυπώνει τις επιδόσεις που επιτεύχθηκαν κατά τη διάρκεια των πειραμάτων χρησιμοποιώντας για τους αλγορίθμους SVM και Random βέλτιστη παραμετροποίηση με βάση τα ανωτέρω, δηλαδή ελάχιστη συχνότητα εμφάνισης χαρακτηριστικού 10 στη συλλογή κειμένων για τον SVM και 1 για τον Random Forest (SVM-10 και RandomForest-1, αντίστοιχα). Στο σημείο αυτό αναδεικνύεται ότι η συμπερίληψη στο σάκο χαρακτηριστικών αγγλικών bigrams της μορφής <επίρρημα><επίθετο> και <επίρρημα><επίρρημα> βελτιώνει την ακρίβεια των αλγορίθμων, αλλά σε πολύ μικρό βαθμό. Ομοίως, η καταμέτρηση της συχνότητας εμφάνισης κάθε χαρακτηριστικού για κάθε κείμενο της συλλογής βελτιώνει μεν την ακρίβεια των αλγορίθμων αλλά όχι κατά πολύ. Το γεγονός αυτό ερμηνεύεται λαμβάνοντας υπόψη τον περιορισμό των 140 χαρακτήρων που ισχύει για τα εξεταζόμενα κείμενα, δηλαδή τα tweets, πράγμα το οποίο συνεπάγεται ότι σπάνια ένα χαρακτηριστικό εμφανίζεται πολλές φορές στο ίδιο κείμενο.

Πίνακας 6. Σύγκριση SVM-10 (συχνότητα>10) & Random Forest-1 (συχνότητα >1)

Frequencies instead of binary values in bag	Include bigrams	SVM-10	RamdomForest-1
No	No	37.9336	39.0899
No	Yes	37.9709	37.4860
Yes	No	38.0082	39.2391
Yes	Yes	**38.0828**	**39.3510**

Γεγονός είναι ότι επίσης ότι ο αλγόριθμος SVM 10 είναι ταχύτερος του Random Forest 1, καθώς εξετάζει σάκους χαρακτηριστικών πολύ μικρότερων διαστάσεων για την κατασκευή μοντέλου ταξινόμησης κειμένων σε κατηγορίες συναισθημάτων. Λαμβάνοντας υπόψη το παραπάνω, στον Πίνακα 7 παρουσιάζονται τα αποτελέσματα εκτέλεσης του SVM με διαφορετικό κατώτατο όριο ελάχιστης συχνότητας χαρακτηριστικών στη συλλογή κειμένων για τη συμπερίληψη τους στο σάκο. Παρατηρείται ότι η περαιτέρω μείωση διαστάσεων του σάκου χαρακτηριστικών έχει θετικό αντίκτυπο στην ακρίβεια του αλγορίθμου, μέχρι ένα ορισμένο σημείο.

Πίνακας 7. SVM και κατώτατο όριο συχνότητας χαρακτηριστικού στη συλλογή

Frequencies instead of binary values in bag	Include bigrams	Lowest total frequency per feature in document collection	SVM
Yes	Yes	15	38.6796
Yes	Yes	25	**39.6121**
Yes	Yes	35	39.4256
Yes	Yes	45	38.9407

Ειδικότερα, η συμπερίληψη χαρακτηριστικών στο σάκο με κατώτατο όριο συχνότητας εμφάνισης στη συλλογή κειμένων ίσο με 25 επιφέρει τα καλύτερα αποτελέσματα συγκριτικά με όλα τα πειράματα που αναφέρθηκαν έως τώρα. Επιπρόσθετα, ο χρόνος εκτέλεσης του αλγορίθμου SVM-25 είναι κατά πολύ μικρότερος σε σύγκριση με τον RandomForest-1.

Τέλος, στον Πίνακα 8 παρουσιάζονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν με διαφορετική στρατηγική ως προς το κατώτατο όριο συχνότητας F χαρακτηριστικών για συμπερίληψη στο σχετικό σάκο που προωθείται σε αλγόριθμο μηχανικής μάθησης. Εν προκειμένω, αντί να εξετάζεται η συχνότητα κάθε χαρακτηριστικού N αυτοτελώς στο σύνολο της συλλογής κειμένων, σε αυτή τη στρατηγική εξετάζονταν αν κάθε χαρακτηριστικό N υπήρχε στη συλλογή τουλάχιστον F φορές με την ίδια κατηγορία (κλάση) συναισθήματος, ως ζεύγος. Όπως προέκυψε, η στρατηγική αυτή είχε ως αποτέλεσμα την αύξηση της ακρίβειας κατηγοριοποίησης του αλγορίθμου SVM καθώς ξεπέρασε το 40%. Ως καλύτερη παραμετροποίηση αναδείχθηκε η χρήση συχνοτήτων, bigrams σε συνδυασμό με τη στρατηγική αυτή.

Πίνακας 8. SVM και κατώτατο όριο χαρακτηριστικού-κλάσης στη συλλογή

Frequencies instead of binary values in bag	Include bigrams	Lowest total frequency per feature-class pair in document collection	Accuracy %	Time to build classification model (sec)
Yes	Yes	3	38.0455	124.36
Yes	Yes	5	40.0970	45.32
Yes	Yes	10	**40.3954**	**16.54**
Yes	Yes	15	39.9851	7.72
Yes	Yes	25	39.0153	4.63

Ολοκληρώνοντας τα πειράματα για την κατασκευή μοντέλου κατηγοριοποίησης κειμένων ως προς το συναίσθημα που εκφράζουν και έχοντας υπόψη τα ανωτέρω αποτελέσματα προχωρήσαμε στην κατηγοριοποίηση των tweets που συλλέξαμε για τις προεδρικές εκλογές στις ΗΠΑ το 2016, αξιοποιώντας τη βέλτιστη προαναφερθείσα προσέγγιση. Η κατηγοριοποίηση tweets διεξήχθη και στα 36 σύνολα δεδομένων ροής ανά δίωρο, ενδεικτικό συγκεντρωτικό αποτέλεσμα των οποίων για ένα από αυτά παρουσιάζεται στον Πίνακα 9. Η κατηγοριοποίηση των 49997 tweets που περιλαμβάνονταν στο συγκεκριμένο δίωρο διήρκεσε 8,44 δευτερόλεπτα με το WEKA. Όπως προέκυψε τα συναισθήματα για τα tweets που σχετίζονται με τις προεδρικές εκλογές των ΗΠΑ το 2016 είναι κατά φθίνουσα σειρά εμφάνισης: enthusiasm, joy και anger. Ωστόσο, αυτά διακρίνονται στο 22,1% των κειμένων, καθώς το 77,2% των tweets είναι συναισθηματικά ουδέτερα.

Πίνακας 9. Ανάλυση συναισθημάτων tweets σχετικών με τις εκλογές στις ΗΠΑ

Συναίσθημα	Πλήθος tweets	Ποσοστό % επί του συνόλου
neutral	38637	**77.2**
anger	671	**1.3**
anxiety	102	0.2
calm	2	0
disgust	0	0
enthusiasm	8395	**16.9**
fear	0	0
interested	170	0.3
joy	1945	**3.9**
nervous	0	0
rejection	0	0
sadness	70	0.1
shame	0	0
surprise	0	0

4 Περιγραφή web εφαρμογής

Η web εφαρμογή οπτικοποίησης των αποτελεσμάτων φιλοξενείται στη διεύθυνση http://us16electionsanalysis.netai.net/ μέσω του παρόχου φιλοξενίας ιστοσελίδων 000webhost (www.000webhost.com). Ο συγκεκριμένος πάροχος μπορεί να φιλοξενεί σελίδες υπό περιορισμούς δυνατοτήτων αλλά εντελώς δωρεάν, συνεπώς για τη συγκεκριμένη ανάλυση και για την οπτικοποίηση που αναφέρεται παρακάτω είναι ιδανικός για χρήση. Αναλυτικά, οι τεχνολογίες που χρησιμοποιήθηκαν είναι:

- HTML
- Bootstrap CSS
- Javascript
- Jquery
- Google Charts Library

Επιπρόσθετα, η διαδικτυακή εφαρμογή αποτελείται από 4 τμήματα:

- Home
- Descriptive Statistics
- Topics Detection
- Affective Analysis

4.1 Γενικές στατιστικές πληροφορίες

Το συγκεκριμένο τμήμα αξιοποιεί μεθόδους περιγραφικής στατιστικής και αποτελείται από διάφορα υπο-τμήματα:

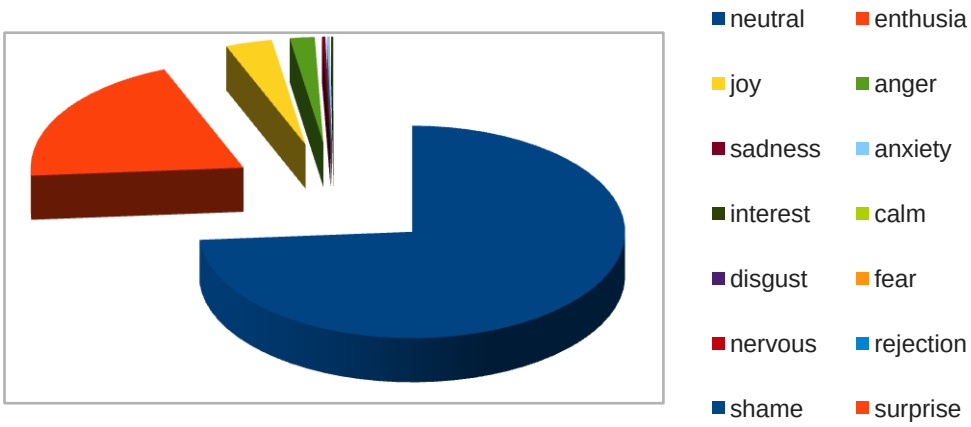
1. Obtained Tweets General Statistics
    - Τμήμα που περιλαμβάνει γενικά στατιστικά στοιχεία όπως:
        1. Στατιστικά στοιχεία για τον ρυθμό μετάδοσης των tweets κατά τη διάρκεια της ανάκτησης tweets σε διαστήματα ανά ώρα.
        2. Το σύνολο των αναφορών στους 3 βασικούς υποψηφίους (Trump, Hillary, Sanders) @mentions & #hashtags και η οποία τους κατά τη διάρκεια της ανάλυσης.
2. RealDonaldTrump General Statistics
    - Τμήμα που περιλαμβάνει γενικά στατιστικά στοιχεία όπως:
        1. Το σύνολο των αναφορών για το λογαριασμό του realDonaldTrump, (@mentions & #hashtags) κατά τη διάρκεια της ανάλυσης.
        2. Το σύνολο των αναφορών για το λογαριασμό του realDonaldTrump, (@mentions) κατά τη διάρκεια της ανάλυσης.
        3. Το σύνολο των αναφορών για το λογαριασμό του realDonaldTrump, (#hashtags) κατά τη διάρκεια της ανάλυσης.
3. HillaryClinton General Statistics
    - Τμήμα που περιλαμβάνει γενικά στατιστικά στοιχεία όπως:
    - Το σύνολο των αναφορών για το λογαριασμό του HillaryClinton, (@mentions & #hashtags) κατά τη διάρκεια της ανάλυσης.
    - Το σύνολο των αναφορών για το λογαριασμό του HillaryClinton, (@mentions) κατά τη διάρκεια της ανάλυσης.
    - Το σύνολο των αναφορών για το λογαριασμό του HillaryClinton, (#hashtags) κατά τη διάρκεια της ανάλυσης.
4. BernieSanders General Statistics
    - Τμήμα που περιλαμβάνει γενικά στατιστικά στοιχεία όπως:
        1. Το σύνολο των αναφορών για το λογαριασμό του BernieSanders, (@mentions & #hashtags) κατά τη διάρκεια της ανάλυσης.
        2. Το σύνολο των αναφορών για το λογαριασμό του BernieSanders, (@mentions) κατά τη διάρκεια της ανάλυσης.
        3. Το σύνολο των αναφορών για το λογαριασμό του BernieSanders, (#hashtags) κατά τη διάρκεια της ανάλυσης.
5. Corellations between Tweets Referenced to Candidates
    - Στο συγκεκριμένο τμήμα παρουσιάζονται σε scatterplot, οι συσχετίσεις μεταξύ των αναφορών στα tweets για τους υποψηφίους. Κάθε διάγραμμα περιγράφει τη συσχέτιση ανά δύο υποψηφίων.
6. Entitites Referenced from Tweets
    - Στο συγκεκριμένο τμήμα παρουσιάζονται τα top-20 twitter accounts, hashtags, expanded urls, people, locations και organizations, από το σύνολο των tweets που συγκεντρώθηκαν κατά την διάρκεια της ανάλυσης. Πρέπει να σημειωθεί ότι τα people, locations και organizations είναι οντότητες που ανιχνεύθηκαν με την Stanford CoreNLP βιβλιοθήκη κατά τη διάρκεια της φάσης της προεπεξεργασίας των tweets.

4.2 Ανίχνευση αναδυόμενων θεμάτων

Στο συγκεκριμένο τμήμα παρουσιάζονται αρχικώς σε έναν πίνακα το θέμα και οι λέξεις-κλειδιά που το αντιπροσωπεύουν ανά χρονικό διάστημα 2 ωρών για τη διάρκεια των τριών ημερών της ανάλυσής μας. Τέλος, παρουσιάζεται ένα διάγραμμα συχνοτήτων των top-6 πιο συχνών εμφανιζόμενων όρων σε όλα τα θέματα και η αυξομείωση της συχνότητάς τους κατά τη διάρκεια της ανάλυσης.

4.3 Ανάλυση συναισθημάτων

Στο συγκεκριμένο τμήμα με χρήση ραβδογραμμάτων και διάγραμμα πίτας παρουσιάζονται οι προβλέψεις των συναισθημάτων των tweets που συλλέχθηκαν στο χρονικό διάστημα από 24/05/16 – 27/05/16. Επίσης στο τέλος παρουσιάζονται με διάγραμμα τύπου annotation graph, για τα 5 επικρατέστερα συναισθήματα (enthusiam, anger, joy, sadness, anxiety) οι διακυμάνσεις στις συχνότητές τους κατά τη διάρκεια της ανάλυσης. Ενδεικτικό παράδειγμα απεικονίζεται στο Διάγραμμα 2.



Διάγραμμα 2. Οπτικοποίηση ανάλυσης συναισθημάτων με διάγραμμα πίτας

Παράρτημα: Ροή εργασίας & αρχιτεκτονική συστήματος