Ατομική Διπλωματική Εργασία

ΕΦΑΡΜΟΓΗ ΕΥΡΕΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΟΔΟΣΦΑΙΡΙΚΩΝ ΑΓΩΝΩΝ ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ ΜΕΣΩ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER

Μιχαήλ Σαματάς

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2015

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εφαρμογή εύρεσης αποτελεσμάτων ποδοσφαιρικών αγώνων σε πραγματικό χρόνο μέσω της εξόρυξης δεδομένων από το Twitter

Μιχαήλ Σαματάς

Επιβλέπων Καθηγητής
Μάριος Δικαιάκος

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2015

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον κ. Δικαιάκο Μάριο, επιβλέπων καθηγητή σε αυτή την εργασία, για τις συμβουλές του, την καθοδήγηση και το έμπρακτο ενδιαφέρον που επέδειξε όλο αυτό το διάστημα.

Θα ήθελα επίσης να ευχαριστήσω τον κ. Αντωνιάδη Δημήτρη για τη σημαντική βοήθεια που μου έδωσε με υποδείξεις, προτάσεις και κατευθύνσεις, καθώς και όλα τα υπόλοιπα μέλη του Εργαστηρίου Δικτυακού Υπολογισμού για σχόλια και τις ιδέες τους.

Ακολούθως θα ήθελα να ευχαριστήσω το Εργαζάκειο Ίδρυμα καθώς και το Πανεπιστήμιο Κύπρου για τις υποτροφίες που μου χορήγησαν. Δεν θα βρισκόμουν σε αυτή τη θέση σήμερα χωρίς τη βοήθειά τους.

Τελειώνοντας, θα ήθελα να ευχαριστήσω τους δασκάλους μου που κάποτε με έβαλαν στο *στραβό το δρόμο*.

Κλείνοντας θα ήθελα να ευχαριστήσω τη μητέρα μου που άφησε τη ζωή της για να με μεγαλώσει και είχε τη δύναμη να με αφήσει να φύγω όταν έπρεπε.

Περίληψη

Ζούμε σε ένα περιβάλλον όπου η χρήση των κοινωνικών δικτύων αυξάνεται ραγδαία χρόνο με το χρόνο και τα API τους εμπλουτίζονται συνεχώς με στόχο να διευκολύνουν την ανάπτυξη όλο και περισσότερο δημοφιλών εφαρμογών για τα οικοσυστήματά τους. Ο ερευνητικός τομέας της εξόρυξης δεδομένων από κοινωνικά δίκτυα έχει παράξει πολλές καινοτόμες εφαρμογές τα τελευταία χρόνια. Εφαρμογές που αξιοποιούν την ανθρώπινη συμπεριφορά και επικοινωνία που υπάρχει γύρω μας για να λύσουν πολύπλοκα υπολογιστικά προβλήματα.

Στόχος αυτής της διπλωματικής εργασίας ήταν η ανάπτυξη ενός συστήματος αυτόματου εντοπισμού αποτελεσμάτων ποδοσφαιρικών αγώνων σε πραγματικό χρόνο, μέσω της εξόρυξης δεδομένων από το Twitter.

Στα πλαίσια αυτής της εργασίας μελετήθηκαν τεχνικές που χρησιμοποιούνται από ερευνητές σε παραπλήσιους τομείς έρευνας, τεχνικές που προσαρμόστηκαν και εφαρμόστηκαν ξανά σε αυτό το σύστημα με γνώμονα την ταχύτητα. Αναπτύχθηκε επίσης και μια μέθοδος εξαγωγής αξιόπιστων αποτελεσμάτων από αναξιόπιστες πηγές πληροφόρησης σε πραγματικό χρόνο, η οποία μπορεί να ανιχνεύει αποτελέσματα ποδοσφαιρικών αγώνων αλλά μπορεί επίσης να προσαρμοστεί για να ανιχνεύει οποιαδήποτε πληροφορία παίρνει έχει διακριτά ενδεχόμενα με προβλέψιμες καταστάσεις μεταπήδησης.

Για το σύστημα αυτό υλοποιήθηκαν δύο προγράμματα πελατών, για να φτάνει η πληροφορία αυτή στους χρήστες. Μια διαδικτυακή εφαρμογή, και μια εφαρμογή για κινητά τηλέφωνα android.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Σε αυτό το κεφάλαιο υπάρχουν τα εισαγωγικά θέματα αυτής της διπλωματικής εργασίας. Το περιβάλλον μέσα στο οποίο εντάσσεται, οι αρχικοί στόχοι που τέθηκαν όταν ξεκίνησε, οι ερευνητικές συνεισφορές που έγιναν μέσα από αυτή τη δουλειά και μια σύντομη περιγραφή της δομής αυτής της αναφοράς και των περιεχομένων των κεφαλαίων της.

1.1 Υποκίνηση της Εργασίας

Τα τελευταία χρόνια έχει σημειωθεί μια ραγδαία παγκόσμια αύξηση στη χρήση των κοινωνικών δικτύων. Αυτή είναι μια τάση που οι εταιρίες αυτές θέλουν να διατηρήσουν και έτσι χρόνο με το χρόνο, εξελίσσουν όλο και περισσότερο τα API τους για να ενθαρρύνουν τη δημιουργία εφαρμογών όπου εμπλουτίζουν το οικοσύστημα τους. Μέσα από αυτή τη διαδικασία έχουν αναπτυχθεί εξαιρετικά καινοτόμες ιδέες. Οι ειδοποιήσεις σήμερα ταξιδεύουν γρηγορότερα από τις σεισμικές δονήσεις [15].

Ο ερευνητικός τομέας της εξόρυξης δεδομένων [14] από κοινωνικά δίκτυα είναι ιδιαίτερα καινούργιος και πολύ ελπιδοφόρος. Χρησιμοποιώντας crowdsourcing [19], οι εφαρμογές αυτές εκμεταλλεύονται την αυθόρμητη ανθρώπινη συμπεριφορά και επικοινωνία για να λύσουν πολύπλοκα υπολογιστικά προβλήματα.

Καθημερινά εκατομμύρια κόσμου μπαίνει στο twitter και συζητάει για ποδοσφαιρικούς αγώνες που είναι σε εξέλιξη. Πολλές ερευνητικές προσπάθειες είναι σε εξέλιξη, προσπαθώντας να εξάγουν αξιοποιήσιμες πληροφορίες από αυτές της συζητήσεις. Αυτή είναι μια προσπάθεια να αυτοματοποιηθεί η διαδικασία ενημέρωσης για τις μεταβολές των αποτελεσμάτων των αγώνων, χρησιμοποιώντας τις συζητήσεις των χρηστών του Twitter.

1.2 Στόχοι της Εργασίας

Οι στόχοι που τέθηκαν για αυτή την εργασία αφορούν την ανάπτυξη ενός συστήματος που θα μπορεί να ανιχνεύει σε πραγματικό χρόνο τα αποτελέσματα ποδοσφαιρικών αγώνων που βρίσκονται σε εξέλιξη.

Το σύστημα αυτό θα πρέπει να είναι σε θέση να ενεργοποιείται, να επεξεργάζεται και να παράγει σωστά αποτελέσματα αυτόματα και χωρίς επίβλεψη.

Το σύστημα αυτό θα πρέπει να είναι γρήγορο, με ταχύτητα απόκρισης τουλάχιστον ίση με εκείνη παρόμοιων υπηρεσιών που βασίζονται στην ανθρώπινη εργασία.

Το σύστημα θα πρέπει να είναι διαθέσιμο τόσο σαν διαδικτυακή εφαρμογή, όσο και σαν εφαρμογή για κινητά.

Αυτή η διπλωματική εργασία ασχολείται με την κατασκευή αυτού του συστήματος, αλλά και με τη διερεύνηση των περιορισμών και των σχεδιαστικών αποφάσεων που πρέπει να παρθούν για να ξεπεραστούν οι τελευταίοι.

1.3 Συνεισφορές της εργασίας

Σε αυτή τη διπλωματική εργασία έχουν γίνει δύο σημαντικές συνεισφορές:

1 Προτάθηκε μια μέθοδος εξεύρεσης αποτελεσμάτων ποδοσφαιρικών αγώνων σε πραγματικό χρόνο η οποία είναι ιδιαίτερα ανθεκτική στο θόρυβο των δεδομένων.

2 Υλοποιήθηκε η εφαρμογή του συστήματος, παρέχοντας έναν έμπρακτο αυτοματισμό σε μια εργασία που απαιτούσε ανθρώπινη εργασία και επίβλεψη μέχρι σήμερα.

1.4 Περίγραμμα της Εργασίας

Σε αυτό το πρώτο κεφάλαιο παρουσιάστηκαν ο σκοπός αυτής της διπλωματικής εργασίας και οι στόχοι της.

Το δεύτερο κεφάλαιο περιέχει βασικές έννοιες που είναι απαραίτητες για την κατανόηση της υπόλοιπης εργασίας. Σε αυτό το κεφάλαιο παρουσιάζεται το Twitter Streaming API, αλλά και οι σχετικές πρόσφατες ερευνητικές ενέργειες πάνω στην χρήση του για την εξαγωγή περιλήψεων, και την αναγνώριση συμβάντων σε πραγματικό χρόνο.

Το τρίτο κεφάλαιο περιέχει μια αναλυτική περιγραφή της λειτουργίας του συστήματος, αναλύοντας όλα τα βήματα που ακολουθούνται από τη συλλογή των δεδομένων μέχρι και την εξαγωγή των αποτελεσμάτων.

Το τέταρτο κεφάλαιο περιέχει τεχνικές λεπτομέρειες και επεξηγήσεις για τις σχεδιαστικές αποφάσεις που αφορούν την υλοποίηση του πρωτοτύπου της εφαρμογής.

Το πέμπτο κεφάλαιο ασχολείται με την αξιολόγηση των μεθόδων και του συστήματος από άποψη ταχύτητας και αξιοπιστίας αποτελεσμάτων.

Στο έκτο κεφάλαιο, τέλος, παραθέτονται τα συμπεράσματα από αυτή την εργασία καθώς και προτάσεις για μελλοντικές επεκτάσεις, βελτιώσεις και άλλες κατευθύνσεις που μπορούν να έχουν μελλοντικές προσπάθειες πάνω στο αντικείμενο.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρατίθενται κάποιες βασικές θεωρητικές έννοιες που είναι απαραίτητες για την μετέπειτα κατανόηση του κειμένου.

Αρχικά υπάρχουν κάποιες πληροφορίες για το κοινωνικό δίκτυο Twitter και την υπηρεσία Twitter Streaming API, την οποία προσφέρει, πάνω στην οποία βασίζεται αυτό το σύστημα.

Έπειτα αναλύονται μια σειρά από χρήσιμες για τη συνέχεια τεχνικές οι οποίες βρίσκονται σε σχετική με το αντικείμενο βιβλιογραφία. Αυτή η βασική θεμελίωση είναι ιδιαίτερα σημαντική καθώς τα μετέπειτα κεφάλαια δείχνουν πώς προσαρμόζονται και πώς αξιοποιούνται αυτές οι τεχνικές στο σύστημα αυτής της διπλωματικής εργασίας

2.1 Ορισμός του προβλήματος

Το ζητούμενο αυτής της εργασίας είναι η περιγραφή και η κατασκευή ενός συστήματος που θα έχει ως είσοδο δεδομένα που θα έρχονται σε πραγματικό χρόνο από το Twitter Streaming API και σαν έξοδο την παρουσίαση των τρεχόντων αποτελεσμάτων ποδοσφαιρικών αγώνων σε μια ιστοσελίδα και σε μια εφαρμογή για κινητά. Η διαδικασία αυτή χωρίζεται σε αρκετά στάδια επεξεργασίας τα οποία θα παρουσιαστούν αναλυτικότερα στα επόμενα κεφάλαια αυτού του κειμένου.

Αρχικά υπάρχει το κομμάτι της συλλογής των δεδομένων. Τα εισερχόμενα δεδομένα από το Twitter Streaming API εξαρτώνται από το περιεχόμενο της επερώτησης που γίνεται προς αυτό. Έτσι γίνεται απαραίτητο να αναζητηθεί ο τύπος των δεδομένων που κρίνεται βέλτιστος για τη λειτουργία της εφαρμογής και να κατασκευαστούν επερωτήσεις που θα μεγιστοποιούν την εισροή αυτών των δεδομένων. Επιπλέον θα πρέπει να αναγνωριστούν οι τύποι των δεδομένων που δυσχεραίνουν την επεξεργασία, αποτελούν *θόρυβο* για την εφαρμογή. Η συλλογή των δεδομένων για αυτό το σύστημα αφορά την εύρεση μιας μορφής επερώτησης που να επιστρέφει όσον το δυνατό μεγαλύτερο ποσοστό ωφέλιμων δεδομένων και όσον το δυνατό μικρότερο ποσοστό θορύβου.

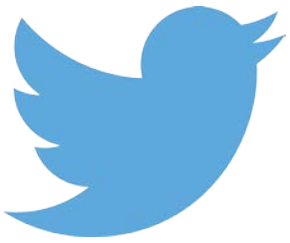
Το επόμενο στάδιο αφορά την εξαγωγή των αποτελεσμάτων σε πραγματικό χρόνο. Στόχος εδώ είναι να κατασκευαστεί μια διαδικασία όπου να παίρνει μια ροή από tweets και να μπορεί να αναγνωρίζει το τρέχον αποτέλεσμα ενός αγώνα και το πότε αυτό αλλάζει. Σε πρώτη φάση πρέπει να γίνει ένας καθαρισμός των δεδομένων, πρέπει να αποκλειστούν όσα δεδομένα δεν έχουν άμεση σχέση με ένα ζητούμενο ποδοσφαιρικό αγώνα. Έπειτα θα πρέπει να κωδικοποιηθεί η πληροφορία που περιέχουν τα tweets σε μια μορφή που να μπορεί να γίνει κατανοητή από το σύστημα. Αυτά τα βήματα θα υλοποιηθούν μέσω ενός φίλτρου γλώσσας και ενός φίλτρου εξαγωγής score term από τα tweet. Το τελευταίο βήμα αυτής της επεξεργασίας, και το πιο σημαντικό, είναι να κατασκευαστεί μια μέθοδος όπου να παίρνει αυτήν τη κωδικοποιημένη πληροφορία από τα εισερχόμενα δεδομένα, να την αξιολογεί και

να αποφασίζει πιο είναι το τρέχον αποτέλεσμα ενός αγώνα και πότε έχει υπάρξει μεταβολή του. Αυτή η λειτουργία γίνεται μέσω ενός αλγορίθμου ψηφοφορίας.

Το τελευταίο στάδιο της επεξεργασίας δεδομένων, αφορά την εξαγωγή μιας περιγραφής για κάθε μεταβολή αποτελέσματος μέσω της ανάλυσης ιστορικών δεδομένων. Στόχος είναι κάθε μεταβολή αποτελέσματος να συνοδεύεται και από μια πρόταση που να την περιγράφει. Η ταχύτητα ήταν μια βασική παράμετρος αυτού του ζητήματος, έτσι προτιμήθηκε η προσέγγιση της αναζήτησης ενός κατάλληλου tweet από τα διαθέσιμα ιστορικά δεδομένα, έναντι μιας προσέγγισης για την κατασκευή μιας αυτόματης περίληψης. Μια επιπρόσθετη παράμετρος ήταν τα αποτελέσματα αυτής της διαδικασίας να μην περιέχουν ύβρεις. Η εργασία αυτή θα εισάγει την έννοια του *χαρακτηριστικού tweet* ενός λεπτού αγώνα, του tweet που περιέχει το μεγαλύτερο αριθμό από τις πιο συχνές λέξεις αυτού του λεπτού στα δεδομένα. Τα ζητήματα σε αυτό το στάδιο της επεξεργασίας αφορούν την ανεύρεση των πιο συχνών λέξεων στη μονάδα του χρόνου, την κανονικοποίηση επιθυμητών λέξεων και τη χρήση αυτών για τον προσδιορισμό του *χαρακτηριστικού tweet.*

Μετά το πέρας της επεξεργασίας υπάρχει το στάδιο της παρουσίασης των αποτελεσμάτων. Στο στάδιο αυτό θα πρέπει να περιγραφούν, να υλοποιηθούν και να παρουσιαστούν όλα τα μέρη του συστήματος. Στο στάδιο αυτό θα πρέπει να αναλυθεί η αρχιτεκτονική του συστήματος, με έμφαση στην κατασκευή της ιστοσελίδας και της εφαρμογής android που επιτελούν το έργο της παρουσίασης των αποτελεσμάτων στον τελικό χρήστη.

2.2 Twitter



Το Twitter είναι ένα κοινωνικό δίκτυο που ιδρύθηκε το 2006. Βασικά του χαρακτηριστικά είναι ότι οι ενημερώσεις των χρηστών, *tweets,* έχουν καθαρά δημόσιο χαρακτήρα, ότι κάθε *tweet* μπορεί να έχει μέγεθος μέχρι 140 χαρακτήρες και ότι στα πλαίσια του δικτύου του δεν υπάρχουν ανταποδοτικές *φιλικές* σχέσεις μεταξύ χρηστών παρά μόνο σχέσεις ακολουθίας που δύναται να είναι μονομερείς. Το Twitter έχει 302 εκατομμύρια ενεργούς χρήστες μηνιαίως και σχεδόν 750 εκατομμύρια εγγεγραμμένους χρήστες, με τη μεγάλη πλειοψηφία των χρηστών του να είναι κάτοικοι των Η.Π.Α [13].

2.3 Twitter Streaming API

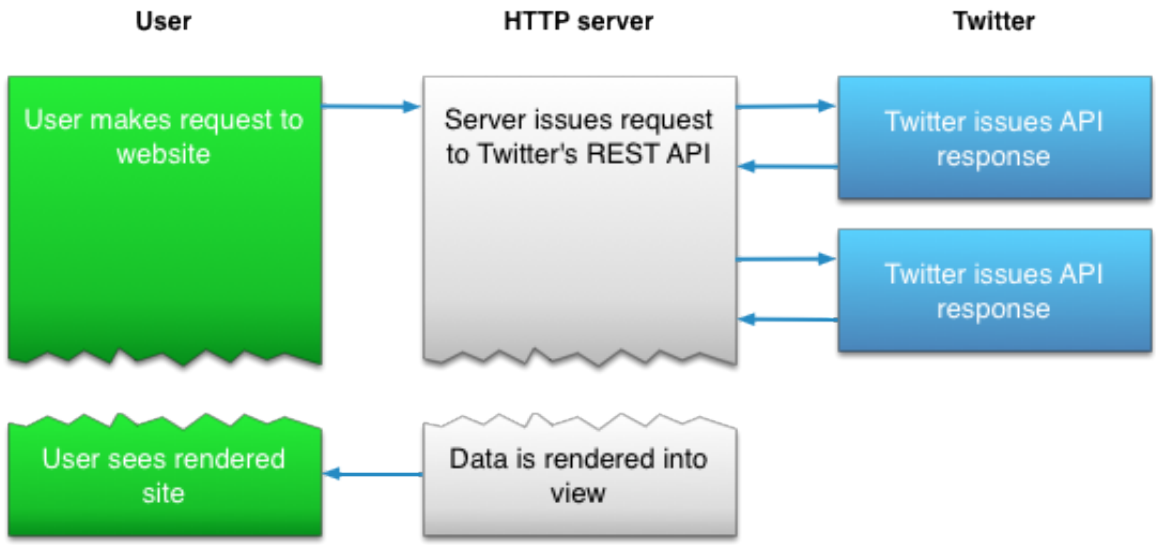
Το Twitter παρέχει εκτεταμένη πρόσβαση στα δεδομένα του σε σχέση με άλλα δημοφιλή κοινωνικά δίκτυα, κάνοντας διαθέσιμα μια πληθώρα από APIs για τους προγραμματιστές που θέλουν να αναπτύξουν εφαρμογές για το οικοσύστημά του. Οι δύο κύριες κατηγορίες API που παρέχει, είναι το REST API [17] και τα Streaming API [18].

Το REST API χρησιμοποιείται για αλληλεπίδραση με τους λογαριασμούς των χρηστών, ενέργειες όπως αποστολή και παραλαβή tweet και πραγματοποίηση follow και unfollow. Επίσης επιτρέπουν τη διεξαγωγή υψηλής αποδοτικότητας αναζητήσεων πάνω σε αποθηκευμένα δεδομένα των τελευταίων 2 εβδομάδων, επιβάλλοντας όμως όρια χρήσης.

Παραδείγματα χρήσης:

Αναζήτηση για πρόσφατα tweets που περιέχουν το hashtag #cyprus.
https://api.twitter.com/1.1/search/tweets.json?q=%23superbowl&result_type=recent

Δημοσίευση του tweet: “Hello world. #cyprus”
`https://api.twitter.com/1.1/statuses/update.json?status=Hello%20world.%20%23cyprus`



Σενάριο χρήσης του Twitter REST API.

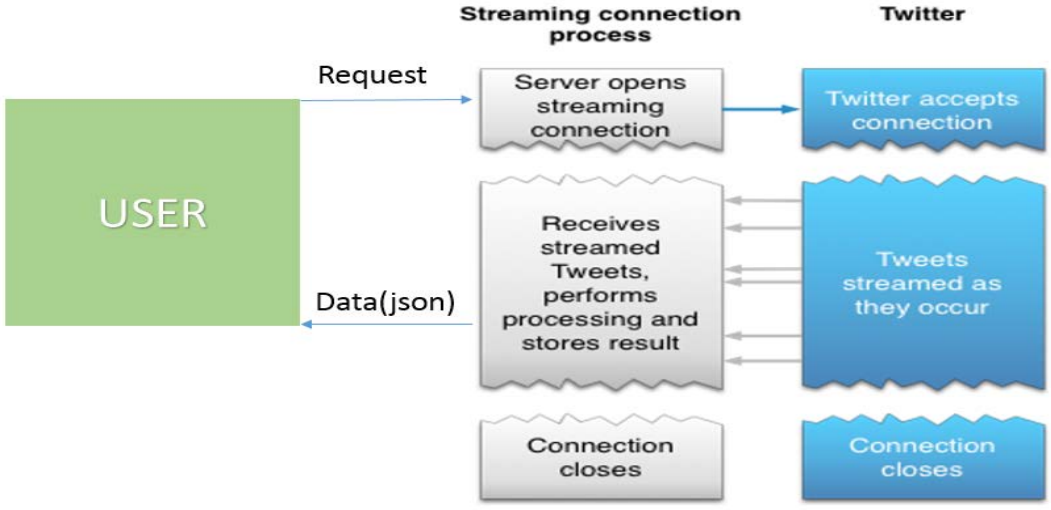
Το Streaming API είναι αυτό που χρησιμοποιείται στην παρούσα εργασία, καθώς επιτρέπει την παρακολούθηση μιας ροής tweet, που ταιριάζουν σε μια επερώτηση, σε πραγματικό χρόνο. Οι ροές του Streaming API γίνονται διαθέσιμες μέσω τριών διαφορετικών end point:

Το /firehose περιέχει το σύνολο όλων των tweet που υπάρχουν στο σύστημα κάθε δεδομένη στιγμή. Η πρόσβαση σε αυτό το end point είναι επί πληρωμή και διατίθεται αποκλειστικά σε μακροχρόνιους συνεργάτες της εταιρίας που κατασκευάζουν μεγάλης σκοπιάς συστήματα.

Το /sample κάνει διαθέσιμο ένα μικρό τυχαίο δείγμα από την κίνηση του /firehose. Το περιεχόμενο του /sample είναι το ίδιο για όλους τους χρήστες που κάνουν αίτηση σε αυτό, κάθε δεδομένη στιγμή.

Το /filter είναι το end point που αξιοποιεί η εφαρμογή αυτής της εργασίας. Το end point αυτό σου επιτρέπει να κάνεις επερωτήσεις και να παίρνεις τα δεδομένα του /firehose που ταιριάζουν στις παραμέτρους σου σε σχεδόν πραγματικό χρόνο. Ο όγκος των δεδομένων που σου κάνει διαθέσιμο το /filter δεν μπορεί να ξεπερνάει ποτέ το 1% του τρέχοντος όγκου δεδομένων στο /firehose, σε τέτοια περίπτωση υπάρχει άρνηση παράδοσης. Τα δεδομένα που εμφανίζονται στο /filter είναι αποτέλεσμα δειγματοληψίας από το /firehose. Η δειγματοληψία αυτή δεν είναι τυχαία, έτσι 2 ίδιες επερωτήσεις από διαφορετικούς χρήστες αναμένεται να επιστρέφουν παρόμοια αποτελέσματα. [20].

Το /filter δέχεται επερωτήσεις που σου επιτρέπουν να λάβεις όλα τα tweets που περικλείονται σε συγκεκριμένες γεωγραφικές περιοχές και έχουν μετα-δεδομένα γεωπληροφορίας. Εναλλακτικά σου επιτρέπει να λάβεις όλα τα tweet που περιέχουν συγκεκριμένες λέξεις, συγκεκριμένα user ids ή και συγκεκριμένες γλώσσες. Περισσότερες λεπτομέρειες υπάρχουν στην παράγραφο 3.1.1 όπου γίνεται αναλυτική αναφορά στη διαδικασία κατασκευής μιας επερώτησης για αυτό το end point.



Σενάριο χρήσης του Twitter Streaming API.

2.4 Χρήσιμες τεχνικές από τη βιβλιογραφία

2.4.1 Συλλογή Δεδομένων

Ο πιο συνηθισμένος τύπος επερώτησης στο Twitter Streaming API /filter από ερευνητές που ασχολούνται με ποδοσφαιρικούς αγώνες είναι μια επερώτηση που περιέχει σαν tracking words τα ονόματα των ποδοσφαιρικών ομάδων και όλα τα ονόματα των ποδοσφαιριστών τους. Αυτή η μέθοδος χρησιμοποιείται σε έρευνες που αφορούν το topic detection [2], σε έρευνες που αφορούν εξαγωγή περιλήψεων [1] και σε έρευνες που αφορούν την αναγνώριση του bias των χρηστών προς συγκεκριμένες ομάδες [3]. Αυτός ο τύπος επερώτησης είναι βέλτιστος σε σενάρια όπου επιδιώκεται να συλλεχθεί όσον το δυνατό μεγαλύτερος όγκος σχετικών δεδομένων, καθώς ένα tweet μπορεί να αναφέρει μόνο συγκεκριμένους ποδοσφαιριστές χωρίς κάποιο όνομα ομάδας. Ο μεγάλος αριθμός των tracking words, όμως, τείνει να αυξάνει το μέγεθος του θορύβου στα δεδομένα λόγω συνωνυμιών των ονομάτων των ποδοσφαιριστών με κοινές λέξεις, φράσεις ή τοποθεσίες.

2.4.2 Αναγνώριση Συμβάντων

Για την αναγνώριση των συμβάντων χρησιμοποιείται μια προσέγγιση που βασίζεται στη συχνότητα εμφάνισης δεδομένων στη μονάδα του χρόνου. Κάποιοι ερευνητές παρακολουθούν τη μεταβολή του συνολικού όγκου των tweet στη μονάδα του χρόνου, και χρησιμοποιούν τα spikes της για να αναγνωρίσουν τα συμβάντα σε έναν αγώνα [1]. Κάποιοι άλλοι παρακολουθούν ξεχωριστά τη μεταβολή των αναφορών σε κάθε tracking word στη μονάδα του χρόνου και αναγνωρίζουν τα συμβάντα αλλά και τους συμμετέχοντες σε αυτά από τα spikes στις μεταβολές αυτές αλλά και τη σχέση μεταξύ τους [2].

Στα πρώτα στάδια αυτής της εργασίας έγιναν προσπάθειες άμεσης εφαρμογής αυτών των τεχνικών. Τα αποτελέσματα αυτών των ενεργειών έδειξαν ότι οι μέθοδοι που βασίζονται στην παρακολούθηση της μεταβολής της συχνότητας εμφάνισης δεδομένων στη μονάδα του χρόνου, είναι πολύ αποτελεσματικές στο να αναγνωρίζουν την ύπαρξη κάποιου γεγονότος

αλλά δεν είναι αρκετές από μόνες τους για να αναγνωριστεί ο τύπος κάποιου γεγονότος (για παράδειγμα γκόλ, κίτρινη κάρτα ή φάουλ). Επιπροσθέτως είναι δύσκολη η εφαρμογή τους σε προβλήματα όπου έχει μεγάλη σημασία ο χρόνος και η επεκτασιμότητα επειδή η χρονική αποδοτικότητά τους εξαρτάται άμεσα από το διάστημα ενεργοποίησής τους (κάθε πότε ελέγχεται η μεταβολή του όγκου των δεδομένων) και κάθε διαφορετική επερώτηση έχει και διαφορετικό βέλτιστο διάστημα ενεργοποίησης.

2.4.3 Εξαγωγή Περιλήψεων

Θόρυβο ονομάζουμε το υποσύνολο των δεδομένων το οποίο δεν έχει πρακτική αξία ή νόημα για το σύστημα. Ο Θόρυβος κάνει δύσκολη την επεξεργασία των δεδομένων γιατί δυσκολεύει την εξαγωγή των χρήσιμων πληροφοριών. Για το παρόν σύστημα, θεωρούνται θόρυβος όσα tweets δεν αναφέρονται στο αποτέλεσμα του αγώνα για τον οποίο γίνεται η επερώτηση ή όσα tweets περιέχουν εσφαλμένη πληροφορία που αφορά το αποτέλεσμα του αγώνα για τον οποίο γίνεται η επερώτηση. Η αφαίρεση του θορύβου από τα δεδομένα είναι το πρώτο βήμα στη διαδικασία κατασκευής μιας περίληψης.

Το περιεχόμενο από άλλες γλώσσες αποτελεί μια πολύ συχνή πηγή θορύβου στα δεδομένα που αφορούν ποδοσφαιρικούς αγώνες. Για την αφαίρεση αυτού του τύπου θορύβου αρχικά αρχικά χρησιμοποιείται η παράμετρος tracking language στην επερώτηση στο Twitter Streaming API /filter σε συνδυασμό με ένα εξωτερικό πρόγραμμα αναγνώρισης γλώσσας επειδή το φίλτρο αυτό αποτυγχάνει συχνά.

Μια επιπλέον πηγή θορύβου είναι τα μηνύματα spam. Για την αφαίρεσή τους γίνεται αναζήτηση του περιεχομένου των tweet για λεξικογραφικούς όρους που είναι συσχετισμένοι με περιεχόμενο spam. Είναι συχνό επίσης στη βιβλιογραφία να γίνεται αυτόματη αφαίρεση όσων tweet περιέχουν urls [1].

Μετά την αφαίρεση του θορύβου τα δεδομένα κανονικοποιούνται για να μπορούν να επεξεργαστούν ευκολότερα. Η διαδικασία της κανονικοποίησης ενοποιεί τις διαφορετικές ορθογραφίες μιας λέξης και αφαιρεί επαναλαμβανόμενα γράμματα (π.χ αλλαγή “goooal” σε “goal”).

Για τη διαδικασία εξαγωγής μιας περίληψης από τα δεδομένα, κάθε tweet χωρίζεται σε προτάσεις από τα σημεία στίξης του. Από αυτές για κάθε tweet φυλάσσεται μόνο η

μεγαλύτερη πρόταση και χρησιμοποιείται για να κατασκευαστεί ένας phrase graph με την εξής διαδικασία: Κάθε λέξη είναι ένας κόμβος, με τις ακμές να βρίσκονται μεταξύ δύο γειτονικών λέξεων. Το βάρος των stop word και των hashtag τίθεται σε 0. Κάθε πρόταση κατακερματίζεται, αφαιρούνται τα διπλότυπα κομμάτια και έπειτα βαθμολογείται. Επιλέγονται οι πρώτες N σε βαθμολογία προτάσεις που δεν έχουν κοινές κανονικές λέξεις (λέξεις που δεν ανήκουν στη λίστα των stop words).

Σύμφωνα με τη βιβλιογραφία αν κανονικοποιείται το βάρος των λέξεων αναλογικά με το μέγεθος μιας πρότασης τότε ευνοούνται μικρότερες προτάσεις, ενώ αντίθετα ευνοούνται μεγαλύτερες προτάσεις με περισσότερη πληροφορία στην τελική περίληψη [1].

Αυτές οι τεχνικές εξαγωγής περιλήψεων απαιτούν μεγάλο όγκο δεδομένων για να παράξουν αποδεκτά αποτελέσματα. Σε ένα σενάριο συλλογής δεδομένων σε πραγματικό χρόνο, αυτό σημαίνει μια σημαντική χρονική καθυστέρηση στο στάδιο της συλλογής, μια καθυστέρηση λεπτών για το σύστημα αυτής της διπλωματικής εργασίας. Αυτό κρίθηκε μη αποδεκτό και έτσι ακολουθήθηκε η προσέγγιση της αναζήτησης ενός tweet όπου να περιγράφει τη μεταβολή ενός αποτελέσματος αγώνα, αντί της κατασκευής μιας κανονικής περίληψης του συμβάντος από τα δεδομένα. Η έννοια που χρησιμοποιείται για αυτό το σκοπό είναι αυτή του *χαρακτηριστικού tweet* η οποία θα οριστεί και θα αναλυθεί στο επόμενο κεφάλαιο.

Κεφάλαιο 3

Περιγραφή - Ανάλυση Συστήματος

Στο παρόν κεφάλαιο αναλύονται με λεπτομέρεια όλα τα βήματα των διαδικασιών του συστήματος, από τη συλλογή των δεδομένων μέχρι και την εξαγωγή των διαφόρων αποτελεσμάτων του.

3.1 Συλλογή Δεδομένων

3.1.1 Εύρεση της κατάλληλης επερώτησης

Το Twitter Streaming API /filter δέχεται δύο τύπους επερωτήσεων. Ο πρώτος τύπος επιστρέφει tweet που έχουν στα μετα-δεδομένα τους γεωγραφική θέση εντός ενός συγκεκριμένου bounding box. Ο δεύτερος τύπος επιστρέφει tweet που περιέχουν συγκεκριμένα user ids, συγκεκριμένες λέξεις ή και συγκεκριμένες γλώσσες [18]. Για τους σκοπούς της εφαρμογής χρησιμοποιείται ο δεύτερος τύπος επερωτήσεων καθώς πολύ μικρό ποσοστό των συνολικών tweet περιέχουν γεωπληροφορίες, οι απαιτήσεις των δεδομένων για την εφαρμογή αυτή δεν περιορίζονται σε συγκεκριμένες γεωγραφικές περιοχές και οι επερωτήσεις που ζητούν γεωπληροφορίες αυτομάτως αγνοούν όλες τις άλλες παραμέτρους των επερωτήσεων στο /filter.

Το Twitter Streaming API δεν έχει χρονικό όριο χρήσης, αλλά κάθε επερώτηση στο /filter που δεν αφορά γεωγραφικές τοποθεσίες έχει όριο 400 track keywords και 5,000 user ids.

Το κύριο ζητούμενο στη συλλογή δεδομένων για αυτή την εφαρμογή είναι η εύρεση των κατάλληλων tracking word ώστε η επερώτηση να επιστρέψει όλα τα σχετικά δεδομένα που χρειάζεται η εφαρμογή να λειτουργήσει με όσο το δυνατό λιγότερο θόρυβο. Μια δεύτερη παράμετρος σε αυτό το πρόβλημα ήταν ότι η διαδικασία εύρεσης αυτών των λέξεων θα πρέπει να αυτοματοποιείται εύκολα και να επεκτείνεται εύκολα σε άλλες γλώσσες και σε άλλα πρωταθλήματα.

Το tracking φίλτρο του /filter θεωρεί ότι όλοι οι όροι που παρακολουθεί συσχετίζονται με OR. Στα πλαίσια ενός συγκεκριμένου όρου τα κενά εκλαμβάνονται ως AND και τα κόμματα ως OR.

Για τη σύγκριση των διαφορετικών επιλογών για την κατασκευή των επερωτήσεων χρησιμοποιήθηκε μια βασική τεχνική αναγνώρισης συμβάντων σε πραγματικό χρόνο μέσα σε κοινωνικά δίκτυα: Η ανάλυση της μεταβολής της συχνότητας των εισερχόμενων δεδομένων στη μονάδα του χρόνου [1].

Για την ανάλυση των διαφορετικών επερωτήσεων, μελετήθηκαν αγώνες από τις εβδομάδες 10 και 11 (Νοέμβριος 2014) του πρωταθλήματος 2014-2015 της Premier League. Από την ανάλυση αυτή παρατηρήθηκε ότι οι πιο απλές επερωτήσεις, οι οποίες περιέχουν μόνο τα ονόματα των ποδοσφαιρικών ομάδων τείνουν να καταγράφουν παρόμοιες ακμές μετά από τη σημείωση τερμάτων (goal), σε σχέση με πιο πολύπλοκες επερωτήσεις που εμπεριέχουν και όλα τα ονόματα των ποδοσφαιριστών ή και τους επίσημους κωδικούς των ομάδων. Οι πιο απλές επερωτήσεις δεν επιστρέφουν αυξημένο όγκο δεδομένων σε μικρότερα συμβάντα του αγώνα (όπως κάρτες ή φάουλ), γεγονός που βοηθάει στην αποφυγή θορύβου για τους σκοπούς αυτής της εφαρμογής. Τέλος οι πιο απλές επερωτήσεις έχουν περίπου 2.5 φορές μικρότερο όγκο εισερχόμενων δεδομένων στη μονάδα του χρόνου, επειδή όπως αναφέρθηκε το Streaming API συσχετίζει όλα τα tracking word με το λογικό OR, κάνοντας τις μεγαλύτερες επερωτήσεις να ταιριάζουν με ένα μεγαλύτερο σύνολο από τα tweets στο /firehose.

Για την εφαρμογή τελικά επιλέχθηκε μια μικρή μορφή επερώτησης που περιέχει μόνο τα ονόματα των ομάδων σαν tracking words και την αγγλική γλώσσα σαν tracking language. Η επιλογή αυτή έγινε επειδή μια επερώτηση αυτής της μορφής αποδίδει μικρότερο ποσοστό θορύβου στα δεδομένα, τα ονόματα των ποδοσφαιριστών συχνά είναι κοινές λέξεις ή τοπωνύμια που αυξάνουν το ποσοστό του θορύβου.

Στην παραπάνω εικόνα βλέπουμε την μετρική αριθμός tweet/μονάδα του χρόνου για τον ίδιο αγώνα με δύο διαφορετικές επερωτήσεις.

Στον παρακάτω κώδικα μπορείτε να δείτε και την παράμετρο tracking language, πέρα από τα tracking words που συζητήθηκαν ήδη. Το φίλτρο της γλώσσας του Twitter Streaming API δεν είναι αλάθητο, αλλά δοκιμές έδειξαν ότι η εισαγωγή της αγγλικής σαν tracking language στις επερωτήσεις για το αγγλικό πρωτάθλημα μείωσε αρκετά το θόρυβο.

```
StatusListener listener = new StatusListener() {…
public void onStatus(Status status) {…}}
FilterQuery fq = new FilterQuery();
String keywords[] = {"Liverpool", "Leicester City"};
fq.track(keywords);
String tlang[] = {"en"};
fq.language(tlang);
twitterStream.addListener(listener);
twitterStream.filter(fq);
```

Μια τυπική επερώτηση (σε Java) για έναν αγώνα της Premier League:

3.1.2 Δομή και Αποθήκευση Δεδομένων

Τα εισερχόμενα tweet είναι σε μορφή json, γεγονός που επιτρέπει την εύκολη απομόνωση και επεξεργασία των σχετικών για την εφαρμογή δεδομένων. Τα δεδομένα αυτά είναι το περιεχόμενο του tweet και ο χρόνος δημιουργίας του.


```
"text": "On my way to Liverpool for the weekend",
"truncated": true,
"in_reply_to_user_id": null,
"in_reply_to_status_id": null,
"favorited": false,
"source": "<a href=\"http://twitter.com/\" rel=\"nofollow\">Twitter for iPhone</a>",
"in_reply_to_screen_name": null,
"in_reply_to_status_id_str": null,
"id_str": "5469180228390928",
"entities":
"contributors": null,
"retweeted": false,
"in_reply_to_user_id_str": null,
"place": null,
"retweet_count": 4,
"created_at": "Wed Apr 08 23:48:36 +0000 2015",
"retweeted_status":{…},
"user": {…},
"id": 54691802283900930,
"coordinates": null,
"geo": null
```

Το περιεχόμενο ενός εισερχόμενου tweet (μορφή json).

Η ποσότητα των δεδομένων ανά λεπτό εξαρτάται από τη δημοφιλία των όρων της αναζήτησης. Σε γενικές γραμμές υπάρχουν αποθηκευτικές ανάγκες περίπου 100KB ανά λεπτό. Για σκοπούς πειραματισμού, στα πλαίσια αυτής της εργασίας συγκεντρώθηκε ένα dataset από 16 αγώνες της Premier League και 13 αγώνες διαφορετικών πρωταθλημάτων στην περίοδο από 19/10/2014 έως 7/4/2015, με συνολικό αριθμό αποθηκευμένων tweet γύρω στα 2.160.000.

3.2 Το Πρόβλημα του Θορύβου

Θόρυβο ονομάζουμε το υποσύνολο των δεδομένων το οποίο δεν έχει πρακτική αξία ή νόημα για το σύστημα. Ο Θόρυβος κάνει δύσκολη την επεξεργασία των δεδομένων γιατί δυσκολεύει την εξαγωγή των χρήσιμων πληροφοριών. Σε αυτή την εφαρμογή, θεωρούνται θόρυβος όσα tweets δεν αναφέρονται στο αποτέλεσμα του αγώνα για τον οποίο γίνεται η επερώτηση ή όσα tweets περιέχουν εσφαλμένη πληροφορία που αφορά το αποτέλεσμα του αγώνα για τον οποίο γίνεται η επερώτηση.

Η εξόρυξη δεδομένων από κοινωνικά δίκτυα αντιμετωπίζει ιδιαίτερα το πρόβλημα του θορύβου λόγω της φύσης των δεδομένων. Οι χρήστες μπορούν πάντα να μοιράζονται λανθασμένες πληροφορίες. Οι ενημερώσεις έχουν μικρό μέγεθος, υπάρχει μεγάλη επαναληπτικότητα [1]. Η συγκεκριμένη εφαρμογή είχε να αντιμετωπίσει και κάποιες επιπρόσθετες διαστάσεις αυτού του ζητήματος. Αρχικά τα δεδομένα δεν καταφθάνουν σε απόλυτη χρονολογική σειρά από το Twitter Streaming API, ένα φαινόμενο γνωστό και ως *αναχρονισμός*. Έπειτα τα φίλτρα του Twitter Streaming API δεν είναι αλάθητα, όσον αφορά τα tracking words και ειδικά τα tracking languages. Στην ανάλυση δεδομένων σε πραγματικό χρόνο, πέρα από την ελαχιστοποίηση του θορύβου, πρέπει να ακολουθηθούν μέθοδοι όπου θα μεγιστοποιούν την ταχύτητα.

3.3 Ανάλυση Δεδομένων σε Πραγματικό Χρόνο

Παρακάτω περιγράφονται τα βήματα της διαδικασίας που ακολουθούνται, ώστε να εξαχθεί το αποτέλεσμα ενός αγώνα σε πραγματικό χρόνο από τα δεδομένα εισόδου. Η διαδικασία αυτή λαμβάνει χώρα κάθε φορά που παραλαμβάνεται ένα νέο tweet από το σύστημα.

3.3.1 Φίλτρο Γλώσσας

Στόχος αυτού του φίλτρου είναι να αποκλειστούν όσα tweets εμπεριέχουν συνομιλίες για στοιχήματα ή είναι μηνύματα spam.

Για την αφαίρεσή τους γίνεται αναζήτηση στο περιεχόμενο του tweet για λεξικογραφικούς όρους που είναι συσχετισμένοι με περιεχόμενο spam ή με περιεχόμενο στοιχημάτων. Η συνομιλία για στοιχήματα είναι σχεδόν πανομοιότυπη με πραγματικά updates που αφορούν το σκόρ, οπότε η αφαίρεσή της στην αρχή της διαδικασίας μειώνει αισθητά το θόρυβο στα επόμενα βήματα. Η επιλογή των λέξεων που ενεργοποιούν αυτό το φίλτρο έγινε αυθαίρετα μετά από πειραματισμό και είναι ένα θέμα το οποίο θα μπορούσε να εξεταστεί με τη χρήση μηχανικής μάθησης στο μέλλον για καλύτερα αποτελέσματα.

3.3.2 Φίλτρο εξαγωγής score term

Σκοπός αυτού του φίλτρου είναι να κωδικοποιήσει την πληροφορία που περιέχουν τα tweets σε μια μορφή που είναι κατανοητή και χρήσιμη για την μετέπειτα επεξεργασία. Η διαδικασία αυτή έχει σαν είσοδο το περιεχόμενο εισερχόμενου tweet και μέσω κάποιων φίλτρων κανονικών εκφράσεων απορρίπτει όσα από αυτά δεν περιέχουν πληροφορία για σκόρ και δεν έχουν μια συγκεκριμένη δομή. Η διαδικασία έχει σαν έξοδο μια συμβολοσειρά που αναφέρεται στο αναγνωρισμένο αποτέλεσμα (π.χ «3-2») αν αναγνωριστεί, αλλιώς έχει σαν έξοδο μια κενή συμβολοσειρά.

Το φίλτρο αυτό υλοποιείται από 2 κανόνες που χρησιμοποιούν κανονικές εκφράσεις, οι οποίοι παραθέτονται παρακάτω με σχετικά πραγματικά παραδείγματα:

Να περιέχει μέσα του τη δομή «[0-9]-[0-9]». \\D*[0-9]-[0-9]\\D*

Περνάει: RT @OptaJoe: 0-0 - The last five Premier League games involving Arsenal have all been 0-0 at half-time. Sleepy.

Δεν Περνάει: @BeingFaridKhan @_OfficialAgent_ where are these then. I only have the arsenal game

Να μην περιέχει πάνω από 2 αριθμούς. Άρνηση στο \\D*[0-9]\\D*[0-9]\\D*[0-9]\\D*

Περνάει: RT @gunnersfr: Alexisssssssss goallllllllllllllll 1-0 pour Arsenal !

Δεν Περνάει: RT @premierleague: HALF-TIME #BPL SCORESArsenal 0-0 SaintsChelsea 2-0 SpursEverton 1-0 HullSunderland 1-2 Man City

Είναι συχνό φαινόμενο ποδοσφαιρικές ομάδες να έχουν ονόματα πόλεων. Επειδή η συλλογή των δεδομένων γίνεται με μια επερώτηση που περιέχει μόνο τα ονόματα των ομάδων και μέχρι τώρα έχει καθαριστεί μόνο ο θόρυβος από μηνύματα spam και συνομιλία για στοιχήματα, σε αυτό το στάδιο υπάρχει στα δεδομένα σημαντικό ποσοστό θορύβου, συνομιλία που αναφέρεται απλά στις πόλεις και δεν έχει σχέση με τον τρέχον ποδοσφαιρικό αγώνα.

Ο συνδυασμός αυτών των δύο κανονικών εκφράσεων αποκλείει όλα τα tweet που δεν περιέχουν ακριβώς δύο αριθμούς. (Το πρώτο φίλτρο ικανοποιείται από δύο αριθμούς μιας συγκεκριμένης δομής, και το δεύτερο από το πολύ δύο αριθμούς). Ο δεύτερος κανόνας υπάρχει για να αποκλείει όσα tweets αναφέρονται σε πολλούς αγώνες, ή σε αγώνες και άλλα αριθμητικά δεδομένα όπως προσωπικά ρεκόρ ποδοσφαιριστών, δεδομένα δηλαδή που θα δυσκόλευαν την άμεση και αξιόπιστη κωδικοποίηση του περιεχομένου σε ένα μόνο αποτέλεσμα ενός μόνο αγώνα.

Ο πρώτος κανόνας διασφαλίζει ότι υπάρχει πληροφορία για αποτέλεσμα που μπορεί να εξαχθεί άμεσα. Αξίζει να σημειωθεί ότι αρκετά tweet δεν περιγράφουν αποτελέσματα αγώνων χρησιμοποιώντας τη μορφή «[0-9]-[0-9]», αλλά χρησιμοποιώντας φυσική γλώσσα όπως «…leads one nill», «the score is two to one». Ο κανόνας αυτός θα αποκόψει αυτά τα tweets.

3.3.3 Αλγόριθμος Ψηφοφορίας

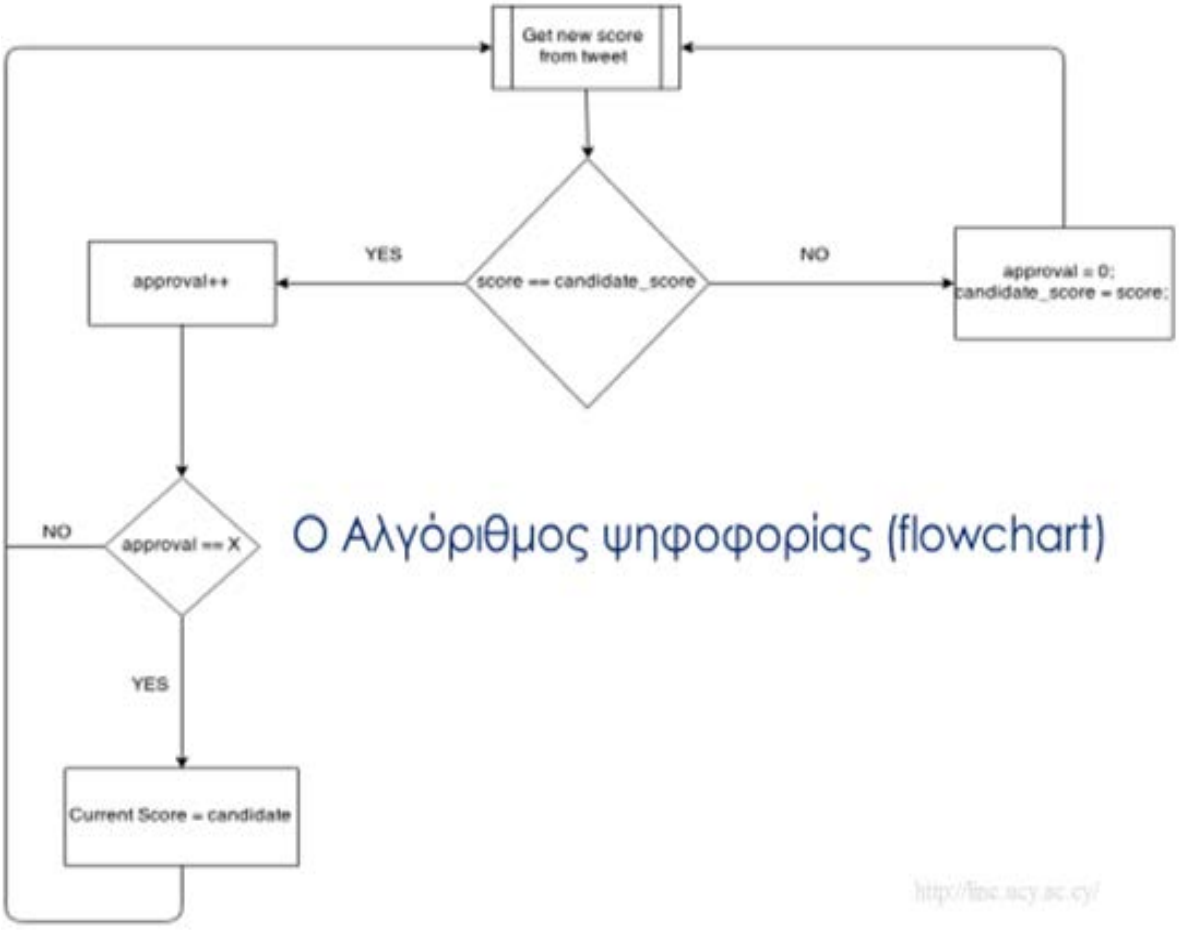
Οι αλγόριθμοι ψηφοφορίας είναι μια πολύ χρήσιμη έννοια στον υπολογισμό. Χρησιμοποιούνται όταν υπάρχει η ανάγκη να εξαχθούν αξιόπιστες πληροφορίες αξιοποιώντας μεγάλο αριθμό πηγών που μπορεί να μην είναι αξιόπιστες [4].

Σε αυτό το σύστημα, ο αλγόριθμος ψηφοφορίας που χρησιμοποιείται είναι μια διαρκής διαδικασία που έχει σαν είσοδο τα score terms που έχει εξαχθεί σαν αποτέλεσμα από το προηγούμενο φίλτρο. Υπενθυμίζεται ότι κάθε tweet που καταφθάνει στο φίλτρο εξαγωγής score term θα καταλήξει να δίνει ένα αποτέλεσμα της μορφής «[0-9]-[0-9]», το οποίο παρακάτω θα λέμε ότι το προτείνει σαν υποψήφιο, ή να δίνει μια κενή συμβολοσειρά, δηλαδή να απέχει από την διαδικασία.

Η διαδικασία αυτή έχει σκοπό να εντοπίζει τις μεταβολές στο αποτέλεσμα ενός αγώνα και σαν έξοδο το νέο αποτέλεσμα.

Ο αλγόριθμος ψηφοφορίας της εφαρμογής ενεργοποιείται για κάθε νέο tweet που περνάει το φίλτρο εξαγωγής score term με αποτέλεσμα διάφορο της κενής συμβολοσειράς και λειτουργεί ως εξής:

- Κάθε εισερχόμενο tweet που περιέχει τη δομή «[0-9]-[0-9]» στο περιεχόμενό του θεωρείται ότι προτείνει αυτή τη δομή ως υποψήφιο αποτέλεσμα.
- Κάθε νέο υποψήφιο αποτέλεσμα που προτείνεται αντικαθιστά πλήρως το τρέχον υποψήφιο αποτέλεσμα και μηδενίζει την πρόοδό του προς την εκλογή.
- Αν η τρέχουσα ψήφος συμφωνεί με το υποψήφιο αποτέλεσμα, τότε η αποδοχή του αυξάνεται κατά 1. Αν η αποδοχή του περάσει ένα όριο, τότε το υποψήφιο αποτέλεσμα εκλέγεται ως το νέο τρέχον αποτέλεσμα.



Ο αλγόριθμος επεξεργάζεται ένα μικρό υποσύνολο των εισερχομένων δεδομένων από το Twitter Streaming API. Η αποδοτικότητά του στηρίζεται στα μεγάλα spike σωστών αποτελεσμάτων που εμφανίζονται μετά από τη σημείωση κάθε τέρματος.

3.4 Ανάλυση Ιστορικών Δεδομένων

Στόχος αυτής της διαδικασίας είναι να παραχθεί μια γρήγορη μέθοδος εξαγωγής πληροφοριών για κάθε μεταβολή αποτελέσματος σε έναν αγώνα. Για να ικανοποιηθεί αυτός ο στόχος, θα ορίσουμε την έννοια του *Χαρακτηριστικού Tweet*. Η έννοια του *Χαρακτηριστικό Tweet* και οι μέθοδοι υπολογισμού του βασίζεται σε τεχνικές αυτόματης εξαγωγής περιλήψεων αθλητικών γεγονότων από το twitter [1], αλλά δεν αποτελούν τα ίδια τυπικές περιλήψεις των γεγονότων.

Η προσέγγιση αυτή, έναντι μιας προσέγγισης που να βασίζεται στην εξαγωγή περίληψης των δεδομένων, προέκυψε από την ανάγκη για ταχύτητα στην εξαγωγή αποτελεσμάτων. Η αποδοτικότητα των μεθόδων εξαγωγής περιλήψεων συναρτάται άμεσα από το μέγεθος του δείγματος, στο συγκεκριμένο σενάριο η συγκέντρωση μεγαλύτερου δείγματος κοστίζει σε χρόνο. Το αποτέλεσμα μιας κανονικής περίληψης για τους σκοπούς αυτού του συστήματος θα ήταν περιορισμένο σε μια ή δύο προτάσεις, λόγω περιορισμών χώρου στις οθόνες των κινητών συσκευών αλλά και λόγω του μεγάλου αριθμού τέτοιων περιλήψεων που θα έπρεπε

να προκύψουν για όλες τις μεταβολές στα αποτελέσματα όλων των αγώνων που διεξάγονται παράλληλα.

Αλλάζοντας το ζητούμενο από το πρόβλημα κατασκευής μιας σύντομης περίληψης κάθε μεταβολής αποτελέσματος, σε πρόβλημα επιλογής ενός tweet από τα δεδομένα που χαρακτηρίζει ικανοποιητικά μια μεταβολή αποτελέσματος δίνεται ένα ισοδύναμο αποτέλεσμα, με μικρότερη πολυπλοκότητα, με ευκολότερη χρονοδρομολόγηση και στη γενική περίπτωση μικρότερο χρόνο απόκρισης.

Η ανάλυση αυτού του σταδίου γίνεται πάνω σε αποθηκευμένα δεδομένα και απαιτεί πρώτα την εύρεση των πιο δημοφιλών λέξεων που περιέχονται στα δεδομένα στη μονάδα του χρόνου.

3.4.1 Δημοφιλείς λέξεις ανά λεπτό αγώνα

Το πρώτο βήμα σε αυτή τη διαδικασία είναι να βρεθούν οι πιο δημοφιλείς λέξεις στο επόμενο λεπτό από το οποίο σημειώθηκε μια μεταβολή στο σκορ. Από τις λέξεις αυτές αφαιρούνται όσες εμπεριέχονται στη λίστα με τα stop words, μια διαδικασία πολύ συχνή σε σχετική βιβλιογραφία [1].

Ένα παράδειγμα μιας τέτοιας ανάλυσης, 8 πιο δημοφιλείς λέξεις ανά λεπτό αγώνα (Man. United vs Stoke City 2/12/2014):

22:03	22:04	22:05	22:06	22:07	22:08	22:09	
Stoke	109 Stoke	81 Old	84 United	1041 United	2415 United	1237 United	664
Manchester	104 Man	47 Stoke	83 Stoke	750 Stoke	1599 Stoke	845 Stoke	474
United	73 Utd	40 0-0	74 Fellaini	559 Fellaini	1251 Fellaini	684 Fellaini	340
0-0	62 United	40 Trafford.	69 1	441 1	1111 Marouane	592 Marouane	283
Old	60 passes	31 gets	68 Marouane	439 Marouane	1106 1	539 1	266
vs	57 Michael	30 stuck	67 put	371 put	1043 Goal!	508 put	238
right	57 pass.	30 Fellaini	67 0.	371 Goal!	1038 put	507 0.	234
Trafford	48 completed	30 PIC:	66 Goal!	370 0.	1036 0.	505 rises	233

3.4.2 Ορισμός και Εύρεση Χαρακτηριστικού Tweet.

Χαρακτηριστικό Tweet: Το μεγαλύτερο σε μήκος tweet, από όσα περιέχουν το μεγαλύτερο αριθμό από δημοφιλείς λέξεις για ένα συγκεκριμένο λεπτό.

Η εύρεσή του γίνεται με την εξής διαδικασία:

- Εύρεση της λίστας των δημοφιλέστερων λέξεων για εκείνο το λεπτό
- Εύρεση του tweet που εμπεριέχει των μεγαλύτερο αριθμό από αυτές τις λέξεις.
- Σε περίπτωση ισοψηφίας, επιλογή του μεγαλύτερου σε μήκος tweet.

Σημείωση: Στην διαδικασία αναζήτησης λέξεων, κάποιοι συγκεκριμένοι όροι κανονικοποιούνται λεξικογραφικά (π.χ goooal σε goal, scooore σε score). Αυτό γίνεται για να αυξηθεί το συνδιασμένο βάρος τους και να εμφανίζονται συχνότερα καθώς κρίνονται πιο σχετικοί. Κάποιοι άλλοι όροι (ύβρεις) δεν κανονικοποιούνται εσκεμμένα για να εμφανίζονται σπανιότερα.

Κεφάλαιο 4

Λεπτομέρειες Υλοποίησης

Σε αυτό το κεφάλαιο περιγράφονται οι υποστηρικτικές διαδικασίες και οι σχεδιαστικές αποφάσεις που αφορούν τον αυτοματισμό του συστήματος και τη φάση του deployment της διαδικτυακής εφαρμογής και την εφαρμογής για κινητά. Στο τέλος του περιλαμβάνεται μια πλήρης εικόνα της αρχιτεκτονικής του συστήματος και μια περιγραφή της υλοποίησης και της αλληλεπίδρασης των διαφόρων κομματιών του.

4.1 Προγραμματισμός Αγώνων

Ένας από τους αρχικούς στόχους της ανάπτυξης αυτού του συστήματος ήταν να μπορεί να λειτουργεί αυτόνομα, χωρίς ανθρώπινη επίβλεψη. Λόγω περιορισμένου αριθμού κλειδιών πρόσβασης στο Twitter Streaming API, αυτό συνεπάγεται στην ανάπτυξη λειτουργικότητας για την αυτόματη κατασκευή μιας επερώτησης και την ενεργοποίησή της την κατάλληλη στιγμή πριν από έναν αγώνα.

Για να μπορεί να το πετύχει αυτό η εφαρμογή χρειαζόταν πρόσβαση σε ένα αξιόπιστο πρόγραμμα αγώνων, όπου θα περιείχε τις πληροφορίες για την ημέρα, την ώρα και τις συμμετέχουσες ομάδες σε έναν ποδοσφαιρικό αγώνα. Η ανεύρεση μιας τέτοιας πηγής, με σταθερή τοποθεσία και μορφοποίηση που να ανανεώνεται διαρκώς αποδείχτηκε δύσκολη, καθώς τα προγράμματα αγώνων μεταβάλλονται αρκετά κατά τη διάρκεια μιας χρονιάς λόγω των υποχρεώσεων των ομάδων σε διεθνείς διοργανώσεις που παίρνουν προτεραιότητα στον προγραμματισμό και των αναβολών και επαναπρογραμματισμών κάποιων αγώνων. Υπάρχουν διάφορα APIs που παρέχουν ιστορικά αθλητικά δεδομένα, αλλά τα μοναδικά που προσφέρουν πληροφορίες για μελλοντικό προγραμματισμό είναι υπηρεσίες επιχειρησιακής κλίμακας με ανάλογο κόστος και δυνατότητες.

Τελικά, για τον προγραμματισμό των αγώνων επιλέχτηκε η λύση να χρησιμοποιηθούν οι πίνακες με τα επίσημα fixtures της Premier League. Η λίστα αυτή φορτώθηκε σε ένα τοπικό αρχείο στον εξυπηρετητή. Στη λίστα αυτή υπάρχουν κάποια λάθη όσον αφορά την ώρα κάποιον αγώνων (περίπου στο 15% των περιπτώσεων) τα οποία διορθώνονται με το χέρι μια φορά το μήνα.

Η λίστα αυτή έχει την ακόλουθη μορφή (ώρα GMT+0)

11/04/2015	14:45	Swansea City **v** Everton
11/04/2015	17:00	Southampton **v** Hull City
11/04/2015	17:00	Sunderland **v** Crystal Palace
11/04/2015	17:00	Tottenham Hotspur **v** Aston Villa
11/04/2015	17:00	West Bromwich Albion **v** Leicester City
11/04/2015	17:00	West Ham United **v** Stoke City
11/04/2015	19:30	Burnley **v** Arsenal

Ένα scheduler bash script [10] που τρέχει στον εξυπηρετητή, ελέγχει διαρκώς αυτή τη λίστα και ενεργοποιεί μια επερώτηση στο Twitter Streaming API μια ώρα πριν την έναρξη κάθε αγώνα. Η σύνδεση τερματίζεται αυτόματα 3 ώρες αργότερα.

4.2 Πρόσβαση στο Twitter Streaming API

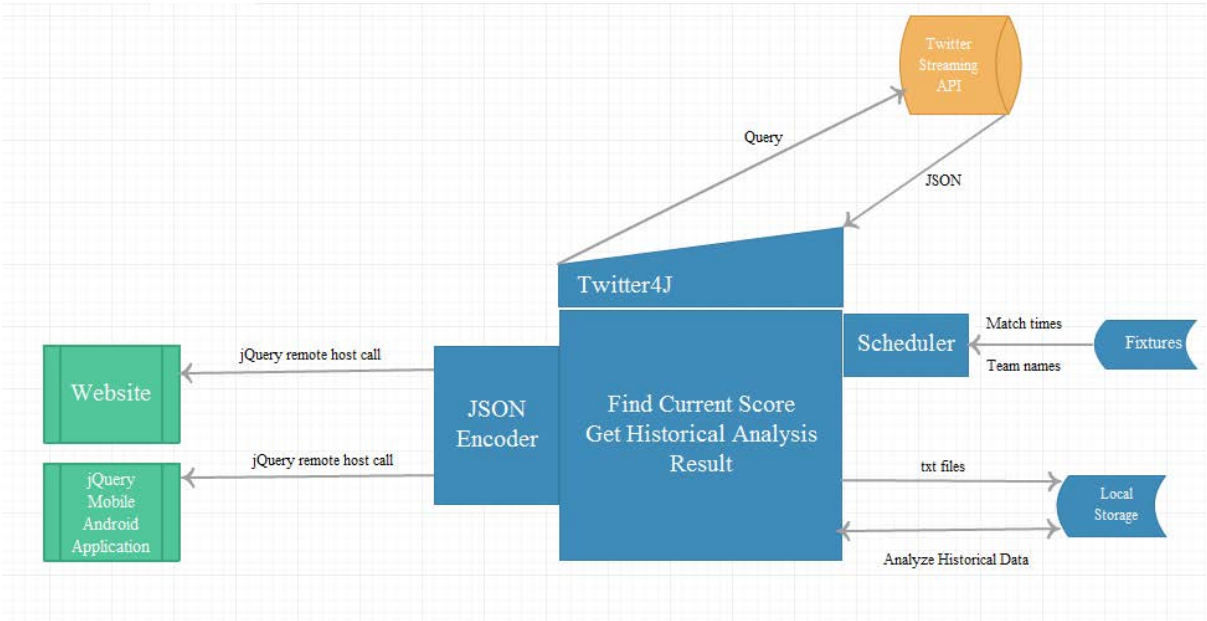
Κάθε λογαριασμός στο Twitter μπορεί να πάρει μέχρι 6 κλειδιά για πρόσβαση στο Twitter Streaming API. Κάθε κλειδί μπορεί να έχει μέχρι 1 σύνδεση ανοιχτή σε κάθε δεδομένη στιγμή. Το scheduler script ελέγχει με τη χρήση flags ποια κλειδιά είναι διαθέσιμα και αποδεσμεύει όποιο κλειδί αποσυνδεθεί από το Twitter. Για την σύνδεση, την παραλαβή και την επεξεργασία των tweets (μορφή json) χρησιμοποιείται η βιβλιοθήκη Twitter4j [5].

4.3 Αρχιτεκτονική Εφαρμογής

Μέχρι εδώ έχουν καλυφθεί τα κομμάτια της λειτουργίας του συστήματος, πώς το σύστημα παίρνει δεδομένα από το twitter, τα επεξεργάζεται και εξάγει αποτελέσματα. Αυτό που μένει να αναλυθεί είναι το κομμάτι του deployment στις διαφορετικές πλατφόρμες, το πώς αυτά τα αποτελέσματα φτάνουν στους τελικούς χρήστες.

Ένας από τους πρώτους στόχους αυτής της εργασίας ήταν η υλοποίηση του τελικού συστήματος σε περισσότερες από μια πλατφόρμες. Αυτό, σε συνδυασμό με τον περιορισμένο αριθμό από κλειδιά για το API, έσπρωξε τη σχεδίαση του συστήματος σε μια συγκεκριμένη κατεύθυνση.

Η επεξεργασία των δεδομένων δεν θα μπορούσε να γίνεται τοπικά στη μηχανή του χρήστη. Επίσης η επεξεργασία των δεδομένων μπορεί να γίνεται μόνο σε ένα σημείο σε πραγματικό



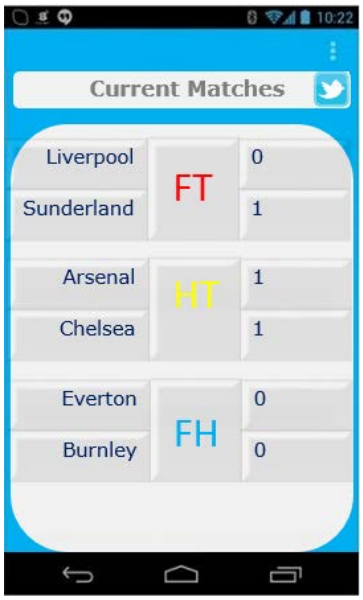
χρόνο, άρα οι διαφορετικοί πελάτες θα έπρεπε να έχουν πρόσβαση και να λαμβάνουν τα δεδομένα τους από ένα κοινό σημείο.

Εδώ φαίνεται ένα διάγραμμα του τελικού συστήματος, το οποίο ακολουθεί την αρχιτεκτονική Model-View-Controller [8].

Για την υλοποίηση της ιστοσελίδας, εξετάστηκαν αρκετές διαφορετικές επιλογές. Το <stream> tag της html5 ήταν μια καλή λύση από άποψη ταχύτητας μετάδοσης δεδομένων, αλλά η λειτουργία του διακόπτεται μετά από παρατεταμένα διαστήματα χρόνου. Η μετάδοση μέσω RSS από την άλλη κοστίζει αρκετά σε χρόνο, αν και είναι αξιόπιστη. Η χρήση της jQuery αποδείχτηκε πειραματικά ως η καλύτερη λύση για μετάδοση πληροφοριών για παρατεταμένα χρονικά διαστήματα με αποδεκτή ταχύτητα.

Ακολουθώντας την ανάπτυξη της ιστοσελίδας, η εφαρμογή για κινητά αναπτύχθηκε χρησιμοποιώντας την πιλοτική πλατφόρμας jQuery Mobile 1.4 [6].

Η κωδικοποίηση των πληροφοριών σε μορφή JSON γίνεται χρησιμοποιώντας τη βιβλιοθήκη JSON Simple [7].



Και οι δύο εφαρμογές πελατών στην ουσία αποτελούν απλά Views, πέρα από τη βασική τους διεπαφή, όλες οι σημαντικές πληροφορίες (πρόγραμμα αγώνων, αγώνες σε εξέλιξη, μεταβολές αποτελεσμάτων, χαρακτηριστικά tweet μεταβολών) έρχονται δυναμικά από απομακρυσμένο εξυπηρετητή. Λόγω αυτού, θα μπορούσε στο μέλλον να αναπτυχθεί ένα διαφορετικό σύστημα επεξεργασίας χωρίς να μεταβληθούν τα Views, ή να προστεθούν επιπλέον Views χωρίς να αλλάξει καθόλου η υπάρχουσα αρχιτεκτονική.

4.4 Εύρεση κατάστασης αγώνα

Υπάρχει ασυμφωνία μεταξύ πραγματικού χρόνου και χρόνου αγώνα, λόγω της ύπαρξης διακοπών στη ροή ενός παιχνιδιού. Οι χρήστες του twitter σπανίως αναφέρονται στο τρέχον λεπτό, εκτός αν έχει συμβεί κάποιο άλλο συμβάν, γεγονός που κάνει δύσκολη την αυτόματη ανακάλυψη του ακριβούς λεπτού αγώνα. Για να μπορεί ο χρήστης να έχει με μια ματία

εποπτεία του που βρίσκεται κάθε παιχνίδι, τα views τον ενημερώνουν για το στάδιο που βρίσκεται το τρέχον παιχνίδι. Τα στάδια είναι: **N**ot **S**tarted, **F**irst **H**alf, **H**alf **T**ime, **S**econd **H**alf, **F**inal **T**ime. Οι πληροφορίες για το τρέχον στάδιο ενός αγώνα εμπεριέχονται σε έναν ακέραιο αριθμό, που παίρνει τιμές από 1-5.

Το παιχνίδι ξεκινάει να εμφανίζεται στην κατάσταση **NS,** με το που ξεκινήσει το σύστημα να μαζεύει δεδομένα για αυτό. Μετακινείται αυτόματα στην κατάσταση **FH** στην προγραμματισμένη ώρα έναρξης. Το σύστημα χρησιμοποιεί μια μέθοδο ανίχνευσης της μεταβολής της συχνότητας εμφάνισης συγκεκριμένων λέξεων στη μονάδα του χρόνου [2] για να μετακινήσει τον αγώνα σε κάθε μια από τις επόμενες καταστάσεις. Για την κατάσταση HT, παρακολουθεί τους όρους «HT» και «Half Time», για την κατάσταση **SH**, παρακολουθεί τους όρους «under way», «started», «second half» και για την κατάσταση **FT** παρακολουθεί τους όρους «**FT**» και «**final**». Το όριο αποδοχής έχει τεθεί στο 200%, χρειάζονται 3 φορές περισσότερες εμφανίσεις των όρων ενός σταδίου σε μια μέτρηση σε σχέση με την προηγούμενη, για να την δεχτεί το σύστημα ως καινούργια κατάσταση.

Κεφάλαιο 5

Αξιολόγηση Μεθόδων

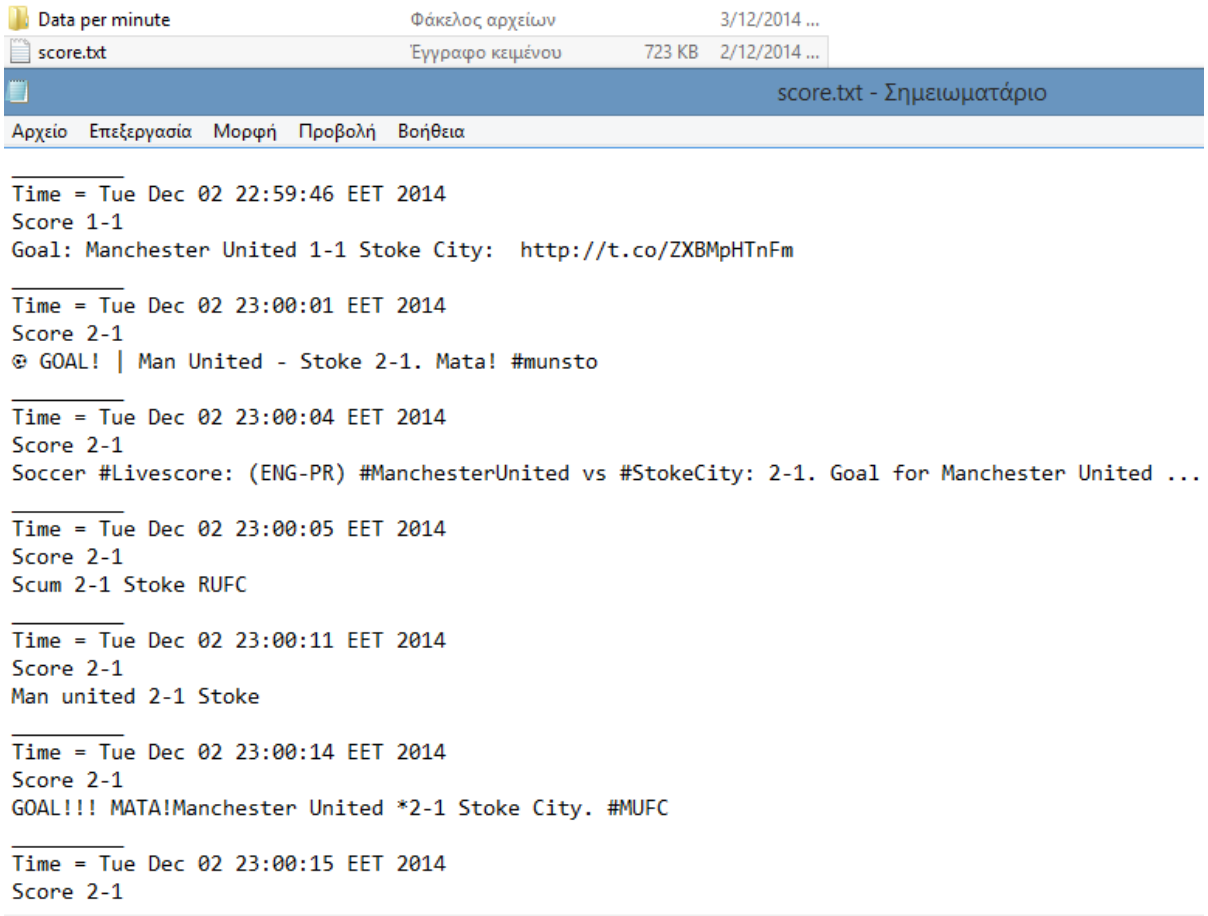
Σε αυτό το κεφάλαιο υπάρχει μια αξιολόγηση των μεθόδων που χρησιμοποιήθηκαν για τη συλλογή και την επεξεργασία των δεδομένων όσον αφορά την αξιοπιστία, την ταχύτητα και την αποδοτικότητά τους. Εδώ γίνεται επίσης μια σύγκριση της αποδοτικότητας του προτεινόμενου συστήματος έναντι παρόμοιων υπηρεσιών που λειτουργούν σήμερα.

5.1 Αξιολόγηση Φίλτρων

Τα φίλτρα γλώσσας και score term φαίνεται να επιτυγχάνουν το βασικό σκοπό τους, να προωθούν στον αλγόριθμο ψηφοφορίας tweets που αναφέρονται στο τρέχον αποτέλεσμα του ποδοσφαιρικού αγώνα.

Γίνεται φανερό, όμως, ότι το πετυχαίνουν αφαιρώντας ένα πολύ μεγάλο ποσοστό των συνολικών tweet. Σε ένα τυπικό αγώνα όπου σημειώνεται ένα τέρμα, μόνο το 5% των συνολικών tweet θα περάσει από αυτά τα φίλτρα. Σε έναν τυπικό αγώνα όπου σημειώνονται τρία τέρματα, μόνο το 8.5% των συνολικών tweet θα περάσει από αυτά τα φίλτρα.

Η λειτουργία αυτή μπορεί να είναι επιθυμητή όταν υπάρχει επάρκεια εισερχόμενων δεδομένων, αλλά είναι προφανής η ανάγκη για επέκταση του φίλτρου των Score Terms για να δέχεται και αποτελέσματα που περιγράφονται με φυσική γλώσσα.



Για κάθε αγώνα, όσα tweets περάσουν από τα φίλτρα αποθηκεύονται και σε ένα ξεχωριστό αρχείο για σκοπούς αποσφαλμάτωσης και ελέγχου. Στη δεύτερη γραμμή κάθε εγγραφής φαίνεται η ψήφος αυτού του tweet προς τον αλγόριθμο ψηφοφορίας.

5.2 Αξιολόγηση Αλγορίθμου Ψηφοφορίας

Ο αλγόριθμος ψηφοφορίας αποδίδει σε ικανοποιητικά αποτελέσματα. Από άποψη ορθότητας αποτελεσμάτων, για τους αγώνες της Premier League και με τις επερωτήσεις που περιγράφηκαν σε αυτήν την εργασία έχει βρεθεί πειραματικά ότι η τιμή 7 στο όριο εκλογής είναι η χαμηλότερη ασφαλής τιμή που μπορεί να τεθεί καθώς με αυτήν δεν σημειώνονται λανθασμένες αλλαγές στο αποτέλεσμα. Για τιμές ορίου ≤ 4 υπάρχει σημαντική πιθανότητα να αναφερθεί λανθασμένα μια μεταβολή στο αποτέλεσμα.

Από άποψη ταχύτητας, με τιμή ορίου 7, ο αλγόριθμος καταλήγει σε σωστή ενημέρωση αποτελέσματος σε ένα διάστημα 7-30 δευτερολέπτων από τη στιγμή άφιξης του πρώτου tweet που περιέχει το σωστό αποτέλεσμα. Δεν υπάρχει εύκολα εκτελέσιμος τρόπος να μετρηθούν οι επιδόσεις ανταγωνιστικών υπηρεσιών όπως το livescore.com για μεγάλο αριθμό παρατηρήσεων, παρακάτω θα παρουσιαστούν μετρήσεις από συγκεκριμένους αγώνες.

5.3 Αξιολόγηση ταχύτητας

Για την αξιολόγηση της ταχύτητας του συστήματος, ήταν αναγκαίο να γίνει σύγκριση με κάποια άλλη πλατφόρμα που παρέχει παρόμοιες υπηρεσίες. Για το σκοπό αυτό επιλέχτηκε η πολύ δημοφιλής υπηρεσία livescore.com. Η υπηρεσία αυτή δεν παρέχει χρονικές πληροφορίες για τις μεταβολές των αποτελεσμάτων στους ποδοσφαιρικούς αγώνες, έτσι κατέστη αναγκαίο να αναζητηθεί μια μέθοδος τοπικής καταγραφής των στιγμών όπου γίνονται οι ανανεώσεις.

Η τοπική αυτή καταγραφή έγινε με το open source εργαλείο CamStudio [21], με λειτουργία καταγραφής 2 frame ανά δευτερόλεπτο, για να επιτευχθεί ακρίβεια δευτερολέπτου στον προσδιορισμό κάθε μεταβολής αποτελέσματος. Η αξιολόγηση αυτή έγινε την Κυριακή 24/5/2015, την τελευταία ημέρα αγώνων του πρωταθλήματος της Premier League 2014-2015.

Πρέπει να σημειωθεί ότι η ταχύτητα απόκρισης του livescore.com εδώ συμπεριλαμβάνει και την ταχύτητα μεταφοράς των δεδομένων από το σημείο επεξεργασίας, στο cdn και τελικά στον τοπικό υπολογιστή που τρέχει την καταγραφή. Η σύνδεση του τοπικού υπολογιστή με

το cdn του livescore.com είχε σταθερό χρόνο απόκρισης 82ms καθ' όλη τη διάρκεια των δοκιμών. Λόγω του μικρού όγκου δεδομένων και του μικρού χρόνου απόκρισης μπορούμε να θεωρήσουμε την επιπλέον καθυστέρηση αμελητέα (μικρότερη του ενός δευτερολέπτου).

Έλεγχος του καταγεγραμμένου βίντεο και καταγραφή των μεταβολών με ακρίβεια δευτερολέπτου

5.3.1 Ταχύτητα του εξυπηρετητή.

Η ταχύτητα του εξυπηρετητή της εφαρμογής μετρήθηκε με βάση τη χρονική στιγμή όπου το σύστημα (ο αλγόριθμος ψηφοφορίας) κατέληξε στη μεταβολή του ποδοσφαιρικού αποτελέσματος.

Arsenal - West Bromwich Albion

Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στον εξυπηρετητή	Διαφορά
05'	1-0	17:04:41	17:04:12	0:00:29
16'	2-0	17:15:20	17:14:10	0:01:10
19'	3-0	17:18:21	17:17:06	0:01:15
39'	4-0	17:38:26	17:37:03	0:01:23
57'	4-1	18:14:41	18:15:42	-0:01:01

Stoke City - Liverpool

Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στον εξυπηρετητή	Διαφορά
24'	1-0	17:24:37	17:22:32	0:02:05
27'	2-0	17:27:39	17:26:30	0:01:09
31'	3-0	17:32:13	17:30:58	0:01:15
42'	4-0	17:43:08	17:41:46	0:01:22
45'	5-0	17:46:10	17:45:40	0:00:30
70'	5-1	18:29:53	18:28:58	0:00:55
87'	6-1	18:46:39	18:44:53	0:01:46

Chelsea - Sunderland

Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στον εξυπηρετητή	Διαφορά
28'	0-1	17:27:39	17:26:40	0:00:59
39'	1-1	17:38:26	17:36:36	0:01:50
71'	2-1	18:29:53	18:28:05	0:01:48
88'	3-1	18:46:39	18:46:04	0:00:35

Στην παραπάνω εικόνα φαίνονται μερικά παραδείγματα τέτοιων μετρήσεων. Πρέπει να σημειωθεί ότι ο χρόνος που φαίνεται για τον εξυπηρετητή μετριέται τη στιγμή που ανακαλύπτεται μια μεταβολή αποτελέσματος, πριν το στάδιο της κωδικοποίησης της πληροφορίας και της αποστολής της.

Παρατηρούμε ότι στη γενική περίπτωση ο αλγόριθμος ψηφοφορίας καταλήγει σε ένα νέο αποτέλεσμα 30-120 δευτερόλεπτα πριν να υπάρξει μια ανανέωση με αυτό το αποτέλεσμα στην υπηρεσία livescore.com.

5.3.2 Ταχύτητα της ιστοσελίδας.

Η ταχύτητα της ιστοσελίδας μετρήθηκε με βάση τη στιγμή όπου υπάρχει μια ανανέωση των αποτελεσμάτων στο view ιστοσελίδας του συστήματος.

Arsenal - West Bromwich Albion				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην ιστοσελίδα	Διαφορά ιστοσελίδας
05'	1-0	17:04:41	17:04:30	0:00:11
16'	2-0	17:15:20	17:14:33	0:00:47
19'	3-0	17:18:21	17:17:31	0:00:50
39'	4-0	17:38:26	17:37:27	0:00:59
57'	4-1	18:14:41	18:16:09	-0:01:53
Stoke City - Liverpool				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην ιστοσελίδα	Διαφορά ιστοσελίδας
24'	1-0	17:24:37	17:23:02	0:01:35
27'	2-0	17:27:39	17:26:54	0:00:45
31'	3-0	17:32:13	17:31:19	0:00:54
42'	4-0	17:43:08	17:42:08	0:01:00
45'	5-0	17:46:10	17:46:04	0:00:06
70'	5-1	18:29:53	18:29:19	0:00:34
87'	6-1	18:46:39	18:45:20	0:01:19
Chelsea - Sunderland				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην ιστοσελίδα	Διαφορά ιστοσελίδας
28'	0-1	17:27:39	17:27:06	0:00:33
39'	1-1	17:38:26	17:36:59	0:01:27
71'	2-1	18:29:53	18:28:32	0:01:21
88'	3-1	18:46:39	18:46:30	0:00:09

Πρέπει να σημειωθεί ότι ο χρόνος αυτός εμπεριέχει και το χρόνο κωδικοποίησης της πληροφορίας σε μορφή json, το χρόνο της αναμονής αιτήματος και της αποστολής στο web server και το χρόνο της αποκωδικοποίησης και ανανέωσης της σελίδας.

Στα παραπάνω παραδείγματα παρατηρούμε ότι στη γενική περίπτωση υπάρχει ανανέωση με ένα νέο αποτέλεσμα στη σελίδα του συστήματός μας 6-75 δευτερόλεπτα γρηγορότερα σε σχέση με την υπηρεσία livescore.com.

5.3.3 Ταχύτητα της εφαρμογής android.

Η ταχύτητα της ιστοσελίδας μετρήθηκε με βάση τη στιγμή όπου υπάρχει μια ανανέωση των αποτελεσμάτων στο view ιστοσελίδας του συστήματος.

Arsenal - West Bromwich Albion				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην εφαρμογή	Διαφορά εφαρμογής
05'	1-0	17:04:41	17:04:31	0:00:10
16'	2-0	17:15:20	17:14:35	0:00:45
19'	3-0	17:18:21	17:17:34	0:00:47
39'	4-0	17:38:26	17:37:32	0:00:54
57'	4-1	18:14:41	18:16:32	-0:02:16
Stoke City - Liverpool				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην εφαρμογή	Διαφορά εφαρμογής
24'	1-0	17:24:37	17:23:03	0:01:34
27'	2-0	17:27:39	17:26:55	0:00:44
31'	3-0	17:32:13	17:31:20	0:00:53
42'	4-0	17:43:08	17:42:12	0:00:56
45'	5-0	17:46:10	17:46:05	0:00:05
70'	5-1	18:29:53	18:29:25	0:00:28
87'	6-1	18:46:39	18:45:23	0:01:16
Chelsea - Sunderland				
Λεπτό Παιχνιδιού	Αποτέλεσμα	Ενημέρωση Livescore	Αποτέλεσμα στην εφαρμογή	Διαφορά εφαρμογής
28'	0-1	17:27:39	17:27:09	0:00:30
39'	1-1	17:38:26	17:37:01	0:01:25
71'	2-1	18:29:53	18:28:34	0:01:19
88'	3-1	18:46:39	18:46:33	0:00:06

Πρέπει να σημειωθεί ότι ο χρόνος αυτός εμπεριέχει και το χρόνο κωδικοποίησης της πληροφορίας σε μορφή json, το χρόνο της αναμονής αιτήματος και της αποστολής στο web server και το χρόνο της αποκωδικοποίησης της πληροφορίας από την εφαρμογή android.

Στα παραπάνω παραδείγματα παρατηρούμε ότι στη γενική περίπτωση υπάρχει ανανέωση με ένα νέο αποτέλεσμα στη σελίδα του συστήματός μας 6-74 δευτερόλεπτα γρηγορότερα σε σχέση με την υπηρεσία livescore.com.

Παρατηρούμε επίσης ότι η εφαρμογή android αποδίδει χρονικά σε παρόμοια επίπεδα με την διαδικτυακή εφαρμογή, γεγονός αναμενόμενο καθώς και οι δύο αποτελούν κυρίως λειτουργίες προβολής των ίδιων αποτελεσμάτων, τα οποία έχουν ήδη υπολογιστεί.

5.4 Αξιολόγηση περιεχομένου χαρακτηριστικών tweet.

Για την αξιολόγηση της ποιότητας του περιεχομένου των χαρακτηριστικών tweet ήταν αναγκαίο να γίνει σύγκριση με τα αποτελέσματα κάποιας άλλη πλατφόρμα που παρέχει παρόμοιες υπηρεσίες. Για το σκοπό αυτό επιλέχτηκε η δημοφιλής υπηρεσία Live Text Commentary που χρησιμοποιείται από το BBC Sports (http://www.bbc.com/sport/0/football/). Η υπηρεσία αυτή παρέχει σύντομες περιγραφές των σημαντικών στιγμών ενός ποδοσφαιρικού αγώνα, ενώ συμβαίνουν.

Οι συγκρίσεις που ακολουθούν αφορούν τον αγώνα Chelsea - Sunderland την Κυριακή 24/5, τελευταία ημέρα του πρωταθλήματος της Premier League 2014-2015.

Αρχή παράθεσης δεδομένων

Γκόλ: 0-1

BBC Sports Live Text Commentary:

25:49

Goal!

Goal! Chelsea 0, Sunderland 1. Steven Fletcher (Sunderland) header from very close range to the bottom left corner. Assisted by Adam Johnson with a cross following a corner.

Χαρακτηριστικό tweet:

> RT @premierleague: GOAL Chelsea 0-1 Sunderland (26 mins) A cross from the right comes all the way to the far post for Steven Fletcher to no…

Πιο δημοφιλείς λέξεις:

Sunderland	801	Steven	226
Chelsea	707	GOAL	213
Fletcher	401	right	187
0-1	393	way	187
post	280	far	187

Γκόλ: 1-1

BBC Sports Live Text Commentary:

36:23 **Goal!**

Goal! Chelsea 1, Sunderland 1. Diego Costa (Chelsea) converts the penalty with a right footed shot to the bottom left corner.

Χαρακτηριστικό tweet:

RT @TSBible: Didier Drogba gets injured in his final game for the club and the Chelsea players carry him off. (@MiguelDelaney) http://t.co/…

Πιο δημοφιλείς λέξεις:

Chelsea	997	Sunderland	336
Drogba	474	final	215
Costa	358	carry	215
"1-1"	349	players	203
Didier	342	GOAL	190

Γκόλ: 2-1

BBC Sports Live Text Commentary:

69:22 **Goal!**

Goal! Chelsea 2, Sunderland 1. Loïc Remy (Chelsea) right footed shot from outside the box to the bottom left corner. Assisted by Eden Hazard.

Χαρακτηριστικό tweet:

RT @premierleague: GOAL Chelsea 2-1 Sunderland (70 mins) Loic Remy puts the Blues ahead with a low shot from the edge of the box #CHESUN

Πιο δημοφιλείς λέξεις:

Chelsea	715	GOAL	228

"2-1"	453	puts	202
Sunderland	384	edge	200
Remy	335	ahead	199
Loic	247	low	199

Γκόλ: 3-1

BBC Sports Live Text Commentary:

87:20 **Goal!**

Goal! Chelsea 3, Sunderland 1. Loïc Remy (Chelsea) right footed shot from very close range to the bottom left corner. Assisted by Nemanja Matic with a cross.

Χαρακτηριστικό tweet:

> RT @premierleague: GOAL Chelsea 3-1 Sunderland (88 mins) Loic Remy slots in his second at the near post for the champions #CHESUN

Πιο δημοφιλείς λέξεις:

Chelsea	745	GOAL	193
"3-1"	459	second	190
Sunderland	392	CHELSEA	182
Remy	270	champions	175
Loic	197	slots	168

Τέλος παράθεσης δεδομένων

Παρατηρούμε ότι συχνά το χαρακτηριστικό tweet καταλήγει να είναι ένα retweet κάποιου χρήστη από τον επίσημο λογαριασμό της Premier League. Αυτό συμβαίνει γιατί ο λογαριασμός αυτός είναι αρκετά δημοφιλής κατά τη διάρκεια των αγώνων. Είναι σημειωτέο ότι τα retweets των χρηστών είναι μεγαλύτερα σε μέγεθος από τα αρχικά tweets του @premierleague, λόγω της προσθήκης της σήμανσης RT στο περιεχόμενο, άρα παρόλο που περιέχουν τον ίδιο αριθμό δημοφιλών λέξεων, θα προτιμώνται πάντα από το σύστημα έναντι των αρχικών tweet.

Στο παράδειγμα αυτό, στο «1-1» το χαρακτηριστικό tweet που επιλέχθηκε από το σύστημα δεν αναφερόταν στην αλλαγή του αποτελέσματος, παρόλο που κάποιες από τις κορυφαίες

λέξεις αναφέρονταν σε αυτό. Αυτό συνέβη διότη η επιλογή του εξαρτάται από των αριθμό των κορυφαίων λέξεων εκείνης της χρονικής περιόδου που αφορούν ένα συγκεκριμένο θέμα, και οι περισσότερες κορυφαίες λέξεις εκείνου του λεπτού αφορούσαν τον τραυματισμό ενός διάσημου ποδοσφαιριστή στο τελευταίο του παιχνίδι με την ομάδα του και όχι στη μεταβολή του αποτελέσματος.

Κεφάλαιο 6

Συμπεράσματα - Μελλοντική Εργασία

Σε αυτό το τελευταίο κεφάλαιο γίνεται μια ανασκόπηση του έργου που έγινε σε αυτή την εργασία και των αποτελεσμάτων που βγήκαν. Στο τέλος αναφέρονται οι τρόποι μελλοντικής βελτίωσης αυτού του συστήματος που έχουν αναγνωριστεί, καθώς επίσης και κάποιες σχετικές ερευνητικές εργασίες που θα μπορούσαν να ξεκινήσουν στο μέλλον, οι οποίες έγιναν εμφανείς κατά τη φάση της ανάπτυξης του παρόντος συστήματος.

6.1 Συμπεράσματα

Η εξόρυξη δεδομένων από κοινωνικά δίκτυα είναι ένας τομέας έρευνας που έχει ανοίξει πρόσφατα και κρύβει ακόμα αρκετά θέματα και κατευθύνσεις όπου μπορεί να ακολουθήσει κάποιος και να παράξει κάτι αξιόλογο. Οι εφαρμογές του εντοπισμού συμβάντων σε πραγματικό χρόνο είναι εντυπωσιακές. Η ενημέρωση για τρέχουσες φυσικές καταστροφές και για μεταβολές σε αποτελέσματα αγώνων είναι μόνο η αρχή. Ο συνδυασμός αυτών των τεχνικών με τη μηχανική μάθηση και την ανάλυση συναισθήματος και γλώσσας, είναι βέβαιο ότι θα φέρει πολλές ακόμα καινοτομίες.

Ξεκινώντας την εργασία, έγινε μελέτη των τωρινών τεχνικών εντοπισμού συμβάντων από κοινωνικά δίκτυα. Χρησιμοποιώντας την ιδιαιτερότητα των ποδοσφαιρικών αγώνων οι οποίοι έχουν διακριτά αποτελέσματα με προβλέψιμες μεταβολές σχεδιάστηκε ένας αλγόριθμος που μπορεί να αναγνωρίζει αποτελέσματα αγώνων με αξιοπιστία και ταχύτητα.

Έπειτα μελετήθηκαν τεχνικές εξαγωγής περιλήψεων από ιστορικά δεδομένα κοινωνικών δικτύων και με βάση αυτές προτάθηκε η έννοια του *χαρακτηριστικού tweet* για μια μεταβολή αποτελέσματος. Μια πληροφορία που ανταποκρίνεται στις ανάγκες τις εφαρμογής και μπορεί να υπολογιστεί σε μικρό χρονικό διάστημα.

Το σύστημα υλοποιήθηκε ταυτόχρονα σαν διαδικτυακή εφαρμογή και σαν εφαρμογή για κινητά. Η δημιουργία αυτού του συστήματος προσφέρει αξία, καθώς για πρώτη φορά αυτοματοποιεί μια διαδικασία όπου μέχρι τώρα χρειαζόταν ανθρώπινη επίβλεψη

Αυτή η διπλωματική εργασία με βοήθησε να αναπτύξω ιδιαίτερα σημαντικές δεξιότητες για την μετέπειτα επαγγελματική μου πορεία. Αρχικά με έφερε σε επαφή με τον κόσμο της ακαδημαϊκής έρευνας μέσω της συνεργασίας μου με το εργαστήριο δικτυακού υπολογισμού του Πανεπιστημίου Κύπρου, μιας συνεργασίας που με δίδαξε πώς να αξιοποιώ τους διαθέσιμους πόρους που έχω αποδοτικά. Οι μηνιαίες παρουσιάσεις προόδου στο εργαστήριο μου έμαθαν πώς να παρουσιάζω αποτελεσματικά το έργο μου και πώς να δέχομαι κριτική. Το θέμα της εργασίας με έφερε σε επαφή με αρκετούς τομείς έρευνας όπου έχουν σημαντικές μελλοντικές ευκαιρίες. Ασχολήθηκα με τον κόσμο της εξόρυξης δεδομένων, με τα εργαλεία και τις τεχνικές του είτε άμεσα είτε έμμεσα μέσω των εργασιών στο εργαστήριο. Έμαθα να χειρίζομαι το Twitter Streaming API. Έμαθα jQuery, προγραμματισμό σε bash, και έφτιαξα

μια ολοκληρωμένη εφαρμογή για κινητά τηλέφωνα. Συνολικά αυτή η εργασία με βοήθησε να αναπτυχθώ σαν προγραμματιστής αλλά και σαν άνθρωπος.

6.2 Μελλοντική Εργασία

Σε αυτή την εργασία δημιουργήθηκε ένα σύστημα το οποίο μπορεί να εντοπίζει αποτελέσματα ποδοσφαιρικών αγώνων σε πραγματικό χρόνο και να ενημερώνει με αξιοπιστία και ακρίβεια τους χρήστες μέσω μιας ιστοσελίδας και μιας εφαρμογής για κινητά. Οι βασικοί στόχοι της εργασίας έχουν επιτευχθεί στο ακέραιο. Κατά τη διάρκεια της ανάπτυξης, όμως, πάρθηκαν μια σειρά από σχεδιαστικές αποφάσεις οι οποίες αφήνουν αρκετό χώρο για μελλοντικές βελτιώσεις στο σύστημα. Επιπλέον, το θέμα της αναγνώρισης συμβάντων μέσω εξόρυξης δεδομένων από κοινωνικά δίκτυα είναι αρκετά καινούργιο ερευνητικά και έχει ακόμα αρκετές ανεξερεύνητες ευκαιρίες για μελλοντικές προσπάθειες. Κάποια από αυτά τα θέματα, τα οποία φάνηκαν κατά τη διάρκεια της ανάπτυξης θα συζητηθούν παρακάτω.

Για το θέμα των βελτιώσεων πάνω στην ίδια την εφαρμογή:

Το φίλτρο γλώσσας που χρησιμοποιείται στην αρχή της συλλογής των δεδομένων θα μπορούσε να υλοποιηθεί με ένα σύστημα μηχανικής μάθησης για καλύτερα αποτελέσματα, σε σχέση με την αφαιρετή λεξικογραφική αφαίρεση όρων που χρησιμοποιείται τώρα.

Το φίλτρο των Score Terms μπορεί να επεκταθεί για να αναγνωρίζει και αποτελέσματα στη φυσική γλώσσα, μετά από κατάλληλη σχετική έρευνα. Έτσι θα αυξηθεί το ποσοστό των tweets που συμμετέχουν στην εξαγωγή αποτελεσμάτων και πιθανώς να γίνει το σύστημα πιο ανθεκτικό σε περιπτώσεις μικρής εισροής εισερχόμενων δεδομένων.

Μπορεί να αναζητηθεί μια καλύτερη λύση για την εξασφάλιση σωστών προγραμμάτων αγώνων, από τα fixture tables που χρησιμοποιούνται τώρα.

Για το θέμα μελλοντικών εργασιών σε παραπλήσια ζητήματα:

Θα μπορούσε να γίνει μια αυτόματη περίληψη ολόκληρου του ποδοσφαιρικού αγώνα, μια κανονική περίληψη για κάθε γκόλ και μια αυτόματη αναγνώριση κάθε σκόρερ στη θέση της πληροφορίας που παρέχει η εύρεση του *χαρακτηριστικού tweet* στην παρούσα εργασία.

Θα μπορούσε να γίνει μια ανάλυση συναισθήματος στα δεδομένα από το Twitter για κάθε ποδοσφαιριστή κατά τη διάρκεια μιας ολόκληρης χρονιάς. Χαρτογραφώντας την αποτύπωση διαφορετικών συμβάντων σε αυτά τα δεδομένα πιθανών να μπορεί να γίνει κάποιου είδους μελλοντική πρόβλεψη απόδοσης.

Τέλος, ενώ αυτή η εργασία απέρριπτε τις συνομιλίες για στοιχήματα από τα δεδομένα, θα ήταν ενδιαφέρουσα μια ερευνητική προσπάθεια όπου θα συγκέντρωνε αυτές τις συνομιλίες και θα τις σύγκρινε με τα πραγματικά μελλοντικά αποτελέσματα ψάχνοντας κατηγορίες χρηστών όπου πέφτουν συχνά μέσα στις προβλέψεις τους.

Βιβλιογραφία

[1] J. Nichols, J. Mahmud and C. Drews “Summarizing Sporting Events Using Twitter” UI '12 Proceedings of the 2012 ACM international conference on Intelligent User Interfaces Pages 189-198.

[2] D. Corney, C. Martin and A. Göker “Spot the ball: Detecting Sports Events on Twitter” 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings. pp 449-454.

[3] D. Corney, C. Martin, A. Göker "Two sides to every story: Subjective event summarization of sports events using Twitter," ICMR2014 1st Workshop on Social Multimedia and Storytelling, Glasgow, UK, Apr. 2014.

[4] B. Parhami “Voting Algorithms” IEEE Transactions on reliability, vol. 43, no. 4, December 1994

[5] Twitter4j, http://twitter4j.org/en/index.html

[6] jQuery Mobile, https://jquerymobile.com/

[7] JSON Simple, https://code.google.com/p/json-simple/

[8] MVC, https://msdn.microsoft.com/en-us/library/ff649643.aspx

[9] Android Studio, https://developer.android.com/sdk/index.html

[10] Bash, http://www.gnu.org/software/bash/

[11] Eclipse, https://eclipse.org/

[12] Java, http://docs.oracle.com/javase/7/docs/api/

[13] “Twitter by the numbers” report, http://news.yahoo.com/twitter-statistics-by-the-numbers-153151584.html

[14] Data Mining, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[15] T. Sakaki, M. Okazaki and Y. Matsuo “Earthquake shakes Twitter users: real-time event detection by social sensors” WWW '10 Proceedings of the 19th international conference on World wide web Pages 851-860

[16] API, http://www.webopedia.com/TERM/A/API.html

[17] Twitter REST API, https://dev.twitter.com/rest/public

[18] Twitter Streaming API, https://dev.twitter.com/streaming/overview

[19] Crowdsourcing, http://www.merriam-webster.com/dictionary/crowdsourcing

[20] K. Joseph, P.M. Landwehr and K.M. Carley “Two 1% don’t make a whole: Comparing simultaneous samples from Twitter’s Streaming API” Carnegie Mellon University Pittsburgh, PA, USA

[21] Camstudio, http://camstudio.org/