COVID-19 ΑΝΑΛΥΣΗ

Φοίβος Τζάβελλος και Αθανάσιος Παπανικολαου

Εξόρυξη Δεδομένων 2019-20
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
{ftzavellos,apapanikolaou}@e-ce.uth.gr

Περίληψη Στην συγκεκριμένη εργασία ασχοληθήκαμε με την χρήση μηχανικής μάθησης και την αξιοποίηση τεχνικών εξόρυξης δεδομένων για την μελέτη του ιού COVID-19. Εφαρμόσαμε κάποιες μεθόδους για να προβλέψουμε την πρόοδο του ιού, να οπτικοποιήσουμε με γραφήματα την πορεία του ιού αντλώντας δεδομένα από το διαδίκτυο και να σφυγμομετρήσουμε την άποψη του κόσμου για την αντιμετώπιση της πανδημίας από τους εκάστοτε αρμόδιους.

Λέξεις Κλειδιά: coronavirus, covid19, Random Forest, Linear Regression

1 Κορωνοϊός

Οι corona viruses ή στα ελληνικά κορωνοϊοί είναι ζωονοσογόνοι ιοί (που σημαίνει ότι μεταδίδονται μεταξύ ζώων και ανθρώπων). Σε σοβαρές περιπτώσεις μπορούν να προκαλέσουν πνευμονία , SARS (severe acute respiratory syndrome), νεφρική ανεπάρκεια, ακόμα και θάνατο. Οι κορωνοϊοί είναι ασυμπτωματικοί, που σημαίνει οτι ο φορέας μπορεί να έχει τον ιο αλλά να μην εμφανίζει συμπτώματα. Ο novel coronavirus (nCov) είναι το νέο στέλεχος που δεν είχε εντοπιστεί σε ανθρώπους. Ο COVID-19 προκαλείται από έναν SARS-COV-2 κορωνοϊό. Ο πρώτος φορέας εντοπίστηκε στο Wuhan, Hubei , China , με τα συμπτώματά του να εμφανίζονται τον Νοέμβριο του ’19. Στις 30 Ιανουαρίου ο Παγκόσμιος Οργανισμός Υγείας κήρυξε το ξέσπασμα ως έκτακτη ανάγκη διεθνούς ενδιαφέροντος για τη δημόσια υγεία.

2 Related Work

Εξαιτίας της διάδοσης και της επικινδυνότητας του ιού η διεθνής επιστημονική κοινότητα έχει κινητοποιηθεί και παρατηρείται μια αύξηση των δημοσιευμένων άρθρων σχετικά με τον ιό και την χρήση μηχανικής μάθησης για την κατανόηση και την καταπολέμηση του. Οι Ensheng Dong et al [7] ανέπτυξαν ένα διαδραστικό , βασισμένο στο διαδίκτυο , dashboard για να παρακολουθούν την εξέλιξη του ιού σε πραγματικό χρόνο. Οι Adam J. Kucharski et al [8] έκαναν μια μαθηματική μελέτη βασισμένοι σε ένα στοχαστικό δυναμικό μοντέλο μετάδοσης του ιού. Οι David Baud et al [9] υπολογίζουν το ποσοστό θνησιμότητας του ιού διαιρώντας τον αριθμό των θανάτων για την δοθείσα ημέρα με τον αριθμό των επιβεβαιωμένων κρουσμάτων 14 μέρες πριν. Οι M. Barstugan et al [10] χρησιμοποιούν μηχανική μάθηση, συγκεκριμένα SVM και κάποιους αλγορίθμους εξόρυξης για να κάνουν ταξινόμηση αξονικών τομογραφιών και

να εντοπίσουν φορείς του ιού μέσα από αυτές (με ακρίβεια 99,7% στον καλύτερο συνδυασμό αλγορίθμων). Οι Zunyou Wu και Jennifer M. McGoogan [2] ανέλυσαν 75000 δεδομένα από το κινέζικο κέντρο λοιμώξεων, συγκρίνοντας τον ιό με τους SARS και MERS. Οι Gopalkrishna Barkur et al [11] κάνουν ανάλυση συναισθημάτων των μηνυμάτων Twitter για να σφυγμομετρήσουν την κοινή γνώμη σχετικά με την καραντίνα στην Ινδία λόγω του ιού.

3 Σύνολο δεδομένων

Τα dataset που χρησιμοποιήσαμε για την δημιουργία των γραφικών παραστάσεων (plots) είναι τα:

- ”covid_19_clean_complete.csv” το οποίο αντλήσαμε από τον σύνδεσμο αυτόν και περιέχει πληροφορίες για τον αριθμό των επιβεβαιωμένων κρουσμάτων, των νεκρών και των ιασμένων για κάθε χώρα και κάθε μέρα από τις 22 Ιανουαρίου 2020.
- ”population_by_country_2020.csv” το οποίο αντλήσαμε από τον σύνδεσμο αυτόν και περιέχει αρκετά στοιχεία, εμείς όμως κρατήσαμε μόνο τον πληθυσμό της κάθε χώρας ώστε να δημιουργήσουμε την στήλη κρούσματα ανά ένα εκατομμύριο κατοίκους.
- ”master_dataset.csv” το οποίο αντλήσαμε από τον σύνδεσμο αυτόν και περιέχει πληροφορίες (κοντά στις 700.000 εισαγωγές) ατομικά για ανθρώπους, όπως την ηλικία τους, το φύλο τους, τον δείκτη μάζας σώματός τους, εξαρτησιογόνες ουσίες που ίσως χρησιμοποιούν , ασθένειες, αν είναι θετικοί στον ιό, την γνώμη τους για την δράση της κυβέρνησής τους και πολλά άλλα.

Για την πρόβλεψη με μεθόδους μηχανικής μάθησης χρησιμοποιήσαμε τρία datasets τα οποία αντλήσαμε από τον ίδιο σύνδεσμο :

- το ’time_series_covid19_confirmed_global.csv’ το οποίο περιέχει πληροφορίες για τον αριθμό των επιβεβαιωμένων κρουσμάτων για κάθε χώρα και κάθε μέρα από τις 22 Ιανουαρίου 2020 (θυμίζει το πρώτο dataset αλλά είναι ξεχωριστά για επιβεβαιωμένα κρούσματα).
- το ’time_series_covid19_deaths_global.csv’ το οποίο περιέχει πληροφορίες για τον αριθμό των νεκρών για κάθε χώρα και κάθε μέρα από τις 22 Ιανουαρίου 2020.
- το ’time_series_covid19_recovered_global.csv’ το οποίο περιέχει πληροφορίες για τον αριθμό των αναρρώσεων για κάθε χώρα και κάθε μέρα από τις 22 Ιανουαρίου 2020.

Για την ανάλυση συναισθημάτων μέσω Twitter χρησιμοποιούμε ένα dataset το οποίο το δημιουργούμε εμείς (περισσότερες τεχνικές λεπτομέρειες στην ενότητα Ανάλυση Συναισθημάτων μέσω Twitter και στον σχολιασμένο κώδικα), όπως επίσης και ένα dataset για την εκπαίδευση του Random Forest με στοιχεία για το συναίσθημα των πελατών σχετικά με διάφορες αμερικάνικες αεροπορικές εταιρίες το οποίο το βρήκαμε εδώ.

4 Γραφικές Παραστάσεις

Ο στόχος μας σε αυτή την ενότητα είναι να επεξεργαστούμε κάποια σύνολα δεδομένων με τέτοιον τρόπο ώστε να εξάγουμε γραφικές παραστάσεις οι οποίες μας περιγράφουν την κατάσταση με τον κορωνοϊό. Τα δεδομένα μας έχουν τελευταία ημερομηνία ανανέωσης την 30 Απριλίου 2020. Κάνουμε κάποια επεξεργασία στο κύριο dataset και δημιουργούμε στήλες με τις ενεργές υποθέσεις για κάθε χώρα, κάποια ποσοστά όπως τους θανάτους ανά 100 υποθέσεις, τον πληθυσμό, τις υποθέσεις ανά ένα εκατομμύριο κατοίκους και την ποσοστιαία αλλαγή στον αριθμό των αρρώστων σε σχέση με την προηγούμενη βδομάδα.Με την βοήθεια της βιβλιοθήκης Plotly Express κάνουμε bar και line plots χρησιμοποιώντας αυτά τα δεδομένα , φτιάχνουμε έναν παγκόσμιο χάρτη με τα επιβεβαιωμένα κρούσματα και τους θανάτους, ένα animation με το πως εξαπλώθηκε ο ιός, heatmap, και συγκρίνουμε τον κορωνοϊό με άλλες πανδημίες όπως Ebola, SARS, MERS, H1N1 ώς προς τα επιβεβαιωμένα κρούσματα, τους θανάτους και το ποσοστό θνησιμότητας. Το συμπέρασμα από το τελευταίο είναι ότι ο COVID19 έχει το δεύτερο μικρότερο ποσοστό θνησιμότητας σε σχέση με τα προαναφερθέντα αλλά σε καθαρούς αριθμούς οι θάνατοι που προκάλεσε ο ιός είναι οι περισσότεροι. Για να οπτικοποιήσουμε δεδομένα ατομικά για υποθέσεις χρησιμοποιούμε ένα dataset που αναφέραμε στην πάνω ενότητα από το οποίο δημιουργούμε γραφικές παραστάσεις για την ηλικία των νοσούντων , το φύλο, τον δείκτη μάζας σώματος και την γνώμη τους για την δράση της κυβέρνησης, η οποία για την Ελλάδα είναι κυρίως θετική.

5 Πρόβλεψη

Για να κάνουμε την πρόβλεψη δημιουργούμε τις χρονοσειρές με τα κρούσματα, τον αριθμό των νεκρών και των ιασμένων, από τις 22 Ιανουαρίου μέχρι τις 30 Απριλίου. Η μελλοντική πρόβλεψη που κάνουμε αφορά 10 μέρες μετά την τελευταία ημερομηνία. Κάνουμε κάποια παρόμοια με την πάνω ενότητα plots χρησιμοποιώντας τις συγκεκριμένες χρονοσειρές και τους πίνακες με δεδομένα. Χρησιμοποιούμε μηχανή υποστήριξης φορέα (Support Vector Machine)για την πρόβλεψη, το οποίο εξηγείται στην υποενότητα. Η πρόβλεψή μας έχει μια υπολογίσιμη απόκλιση απο την πραγματικότητα, πράγμα που φαίνεται και από το νούμερο των μέσων απόλυτων λαθών και μέσων τετραγωνισμένων λαθών. Στην αρχή πάει αρκετά καλά , κάποια στιγμή 35 μέρες περίπου μετά την αρχή προβλέπει περισσότερους θανάτους από τα πραγματικά νούμερα και στις 70 περίπου μέρες και μετά προβλέπει λιγότερους θανάτους από την πραγματικότητα.
Χρησιμοποιούμε γραμμική παλινδρόμηση (Linear Regression) και τα αποτελέσματα είναι κάπως χειρότερα σε σχέση με SVM. Το γραμμικό μοντέλο αποτυγχάνει αρκετά να αποδώσει την πραγματική εικόνα της προόδου του ιού.

5.1 Support Vector Machines

Στη μηχανική μάθηση, οι μηχανές υποστήριξης-φορέα (SVMs, καθώς και τα δίκτυα υποστήριξης-φορέα) είναι εποπτευόμενα μοντέλα μάθησης με συναφείς αλγόριθμους μάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για την ταξινόμηση και την

ανάλυση παλινδρόμησης. Λαμβάνοντας υπόψη ένα σύνολο παραδειγμάτων εκπαίδευσης, το καθένα επισημαίνεται ότι ανήκει σε μία ή την άλλη από δύο κατηγορίες, ένας αλγόριθμος εκπαίδευσης SVM δημιουργεί ένα μοντέλο που εκχωρεί νέα παραδείγματα σε μια κατηγορία ή την άλλη, καθιστώντας τον μη πιθανό δυαδικό γραμμικό ταξινομητή. Ένα μοντέλο SVM είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο διάστημα, χαρτογραφημένα έτσι ώστε τα παραδείγματα των ξεχωριστών κατηγοριών να διαιρούνται με ένα σαφές κενό που είναι όσο το δυνατόν ευρύτερο. Στη συνέχεια, νέα παραδείγματα χαρτογραφούνται στον ίδιο χώρο και προβλέπεται να ανήκουν σε μια κατηγορία με βάση την πλευρά του κενού στην οποία πέφτουν. Εκτός από την εκτέλεση γραμμικής ταξινόμησης, τα SVMs μπορούν να εκτελέσουν αποτελεσματικά μια μη γραμμική ταξινόμηση χρησιμοποιώντας αυτό που ονομάζεται κόλπο πυρήνα, χαρτογραφώντας σιωπηρά τις εισόδους τους σε χώρους μεγάλων διαστάσεων. Πιο τυπικά, ένα μηχάνημα φορέα υποστήριξης κατασκευάζει ένα υπερπλάνο ή ένα σύνολο υπερπλάνων σε ένα χώρο υψηλού ή άπειρου διαστάσεων, το οποίο μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση ή άλλες εργασίες όπως ανίχνευση ακραίων τιμών. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερπλάνο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο εκπαίδευσης-δεδομένων οποιασδήποτε κλάσης (το λεγόμενο λειτουργικό περιθώριο), καθώς γενικά όσο μεγαλύτερο είναι το περιθώριο, τόσο χαμηλότερο είναι το σφάλμα γενίκευσης του ταξινομητή.
Ένα υπερεπίπεδο στον ν-διάστατο χώρο μπορεί να αναπαρασταθεί από την ακόλουθη εξίσωση:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \sum_{i=1}^{n} x_i w_i + b = 0$$

Διαιρώντας με $||\mathbf{w}||$, παίρνουμε

$$\frac{\mathbf{x}^T\mathbf{w}}{||\mathbf{w}||} = P_{\mathbf{w}}(\mathbf{x}) = -\frac{b}{||\mathbf{w}||}$$

που δείχνει ότι η προβολή κάθε σημείου $\mathbf{x}$ στο επίπεδο πάνω στο διάνυσμα $\mathbf{w}$ είναι πάντα $-b/||\mathbf{w}||$, δηλαδή, $\mathbf{w}$ είναι η κανονική κατεύθυνση του επιπέδου, και $|b|/||\mathbf{w}||$ είναι η απόσταση από την αρχή του επιπέδου. Να σημειωθεί ότι η εξίσωση του υπερεπιπέδου δεν είναι μοναδική. $c\,f(\mathbf{x}) = 0$ αναπαριστά το ίδιο επίπεδο για κάθε c.

Ο ν-διάστατος χώρος χωρίζεται σε δύο περιοχές από το επίπεδο. Ειδικά, ορίζουμε μια εξίσωση που αντιστοιχίζει τα σημεία $y = sign(f(\mathbf{x})) \in \{1, -1\}$,

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \begin{cases} > 0,\, y = sign(f(\mathbf{x})) = 1,\ \mathbf{x} \in P \\ < 0,\, y = sign(f(\mathbf{x})) = -1,\ \mathbf{x} \in N \end{cases}$$

Κάθε σημείο $\mathbf{x} \in P$ που είναι στην θετική πλευρά του επιπέδου αντιστοιχίζεται στο 1 , ενώ κάθε σημείο $\mathbf{x} \in N$ στην αρνητική πλευρά αντιστοιχίζεται στο -1. Ένα σημείο $\mathbf{x}$ άγνωστης κλάσης θα ανήκει στο P αν $f(\mathbf{x}) > 0$, ή στο N αν $f(\mathbf{x}) < 0$.

5.2 Linear Regression

Στα στατιστικά στοιχεία, η γραμμική παλινδρόμηση είναι μια γραμμική προσέγγιση στη μοντελοποίηση της σχέσης μεταξύ μιας κλιματικής απόκρισης (ή εξαρτημένης μεταβλητής) και μιας ή περισσότερων επεξηγηματικών μεταβλητών (ή ανεξάρτητων μεταβλητών). Η περίπτωση μιας επεξηγηματικής μεταβλητής ονομάζεται απλή γραμμική

παλινδρόμηση. Για περισσότερες από μία επεξηγηματικές μεταβλητές, η διαδικασία ονομάζεται πολλαπλή γραμμική παλινδρόμηση. Αυτός ο όρος διαφέρει από την γραμμική παλινδρόμηση πολλαπλών παραλλαγών, όπου προβλέπονται πολλαπλές συσχετιζόμενες εξαρτημένες μεταβλητές, και όχι μεμονωμένη κλιματική μεταβλητή.
Στη γραμμική παλινδρόμηση, οι σχέσεις μοντελοποιούνται χρησιμοποιώντας συναρτήσεις γραμμικής πρόβλεψης των οποίων οι άγνωστες παράμετροι μοντέλου εκτιμώνται από τα δεδομένα. Τέτοια μοντέλα ονομάζονται γραμμικά μοντέλα. Συνήθως, ο υπό όρους μέσος όρος της απόκρισης δεδομένης της τιμής των επεξηγηματικών μεταβλητών (ή προγνωστικών) θεωρείται ότι είναι μια συνάρτηση συνάρτησης αυτών των τιμών. Όχι πιο συχνά, χρησιμοποιείται η υπό όρους διάμεση τιμή ή κάποιο άλλο ποσοτικό. Όπως όλες οι μορφές ανάλυσης παλινδρόμησης, η γραμμική παλινδρόμηση επικεντρώνεται στην κατανομή πιθανότητας υπό όρους της απόκρισης, δεδομένων των τιμών των προγνωστών, και όχι στην κοινή κατανομή πιθανότητας όλων αυτών των μεταβλητών, η οποία είναι το πεδίο της πολυπαραγοντικής ανάλυσης.
Η γραμμική παλινδρόμηση ήταν ο πρώτος τύπος ανάλυσης παλινδρόμησης που μελετήθηκε αυστηρά και χρησιμοποιήθηκε εκτενώς σε πρακτικές εφαρμογές. Αυτό συμβαίνει επειδή τα μοντέλα που εξαρτώνται γραμμικά από τις άγνωστες παραμέτρους τους είναι ευκολότερα στην εφαρμογή από τα μοντέλα που δεν σχετίζονται γραμμικά με τις παραμέτρους τους και επειδή οι στατιστικές ιδιότητες των προκύπτοντων εκτιμητών είναι πιο εύκολο να προσδιοριστούν. Υπάρχουν δύο ευρύτερες και γενικές κατηγορίες linear regression:

- Εάν ο στόχος είναι πρόβλεψη, πρόβλεψη ή μείωση σφάλματος, η γραμμική παλινδρόμηση μπορεί να χρησιμοποιηθεί για την προσαρμογή ενός μοντέλου πρόβλεψης σε ένα παρατηρημένο σύνολο δεδομένων τιμών απόκρισης και επεξηγηματικών μεταβλητών. Μετά την ανάπτυξη ενός τέτοιου μοντέλου, εάν συλλέγονται πρόσθετες τιμές των επεξηγηματικών μεταβλητών χωρίς συνοδευτική τιμή απόκρισης, το προσαρμοσμένο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη της απόκρισης.
- Εάν ο στόχος είναι να εξηγήσουμε τη διακύμανση στη μεταβλητή απόκρισης που μπορεί να αποδοθεί σε παραλλαγή στις επεξηγηματικές μεταβλητές, μπορεί να εφαρμοστεί ανάλυση γραμμικής παλινδρόμησης για τον ποσοτικό προσδιορισμό της σχέσης μεταξύ της απόκρισης και των επεξηγηματικών μεταβλητών, και ιδίως για να καθοριστεί εάν ορισμένες Οι επεξηγηματικές μεταβλητές ενδέχεται να μην έχουν καθόλου γραμμική σχέση με την απόκριση ή να προσδιορίσουν ποια υποσύνολα των επεξηγηματικών μεταβλητών ενδέχεται να περιέχουν περιττές πληροφορίες σχετικά με την απόκριση.

Τα μοντέλα γραμμικής παλινδρόμησης συχνά τοποθετούνται χρησιμοποιώντας την προσέγγιση των λιγότερων τετραγώνων, αλλά μπορούν επίσης να τοποθετηθούν με άλλους τρόπους, όπως ελαχιστοποιώντας την «έλλειψη προσαρμογής» σε κάποιο άλλο πρότυπο (όπως με την παλινδρόμηση με απόλυτες αποκλίσεις) ή ελαχιστοποιώντας μια τιμωρημένη έκδοση της λειτουργίας με τα λιγότερα τετράγωνα όπως στην παλινδρόμηση κορυφής (ποινή L2-norm) και lasso (ποινή L1-norm). Αντίθετα, η προσέγγιση των λιγότερων τετραγώνων μπορεί να χρησιμοποιηθεί για να ταιριάζει σε μοντέλα που δεν είναι γραμμικά μοντέλα. Έτσι, αν και οι όροι "ελάχιστα τετράγωνα" και "γραμμικό μοντέλο" είναι στενά συνδεδεμένοι, δεν είναι συνώνυμοι.
Δοθέντος ενός συνόλου δεδομένων $\{y_i,\ x_{i1}, ..., x_{ip}\}_{i=1}^{n}$ από n στοιχεία , ένα γραμμικό

μοντέλο υποθέτει ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής y και και του p-φορέα της παλινδρόμησης x_i είναι γραμμική. Η σχέση αυτή διαμορφώνεται μέσα από έναν όρο διαταραχής ή λάθος μεταβλητή e_i - μία απαρατήρητη τυχαία μεταβλητή που προσθέτει θόρυβο με τη γραμμική σχέση ανάμεσα στην εξαρτημένη μεταβλητή και παλινδρόμηση. Έτσι, το μοντέλο παίρνει τη μορφή

$$y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i$$

$$i = 1, ..., n$$

όπου Τ συμβολίζεται ο ανάστροφος, άρα $x_i^T \beta$ είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων x_i και β.

6 Ανάλυση Συναισθημάτων μέσω Twitter

Σε αυτή την ενότητα δοκιμάζουμε την ταξινόμηση μικρών μηνυμάτων από το Twitter με γνώμονα το συναίσθημα τους, χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Τα μηνύματα στο Twitter, ή αλλιώς tweets, όπως είναι ευρέως γνωστά, περιορίζονται στους 140 χαρακτήρες. Αυτός ο περιορισμός εισάγει μια επιπρόσθετη δυσκολία για τους ανθρώπους στο να εκφράσουν τα συναισθήματα τους και συνεπώς η ταξινόμηση αυτού του συναισθήματος σε θετικό , αρνητικό ή ουδέτερο θα είναι ακόμα πιο δύσκολη. Το θέμα σύμφωνα με το οποίο κάναμε εξαγωγή των tweets είναι η αντιμετώπιση του ιού από τον Donald Trump. Αποκτήσαμε ενα Twitter Developer Account [1] και το χρησιμοποιήσαμε για να ανοίξουμε ενα stream από δεδομένα μέσω της βιβλιοθήκης tweepy. Για την προεπεξεργασία από τα χαρακτηριστικά των tweets κρατήσαμε το αναγνωριστικό τους , πότε δημιουργήθηκαν, το περιεχόμενο του κειμένου (από το οποίο αφαιρέσαμε τα emojis και τους συνδέσμους / ειδικούς χαρακτήρες), την πόλωση και την αντικειμενικότητα του κειμένου, καθώς και πληροφορίες για τον χρήστη. Δημιουργήσαμε το database στο mySQL και αποθηκεύσαμε στον πίνακα για πιο εύκολη ανάλυση / αποθήκευση. Τα συγκεκριμένα track words που χρησιμοποιήσαμε είναι τα ” Coronavirus Trump”. Η ανάλυση συναισθημάτων έγινε μέσω της βιβλιοθήκης Textblob, μιας NLP (Natural Language Processing - Επεξεργασία Φυσικής Γλώσσας) βιβλιοθήκης για επεξεργασία γραπτών δεδομένων , με έμφαση στην ανάλυση συναισθημάτων. Μας επιστρέφει το αντικείμενο sentiment που δείχνει πόσο πολωμένο είναι το συγκεκριμένο κείμενο , απο -1 για αρνητικά έως 1 για θετικά και πόσο αντικειμενικό είναι (0 για πολύ αντικειμενικό μέχρι 1 για πολυ υποκειμενικό).
Δημιουργούμε γραφικές παραστάσεις για τα δεδομένα που συλλέξαμε. Για αρχή ενα διάγραμμα γραμμής (line chart) σχετικά με την στάση απέναντι στο θέμα (κυρίως ουδέτερη). Βρίσκουμε τις πιο συχνές λέξεις που συναντάμε στα συγκεκριμένα tweets , χρησιμοποιώντας τις βιβλιοθήκες NLTK (Natural Language Toolkit) και την RE (Regular Expressions). Μέσω της RE αφαιρούμε όλους τους συνδέσμους, τα σύμβολα των retweets και ’&’, ενώνουμε όλο το κείμενο και μετατρέπουμε όλους τους χαρακτήρες σε μικρά γράμματα. Μέσω της NLTK χρησιμοποιούμε το tokenizer που διαχωρίζει την συμβολοσειρά σε υποσυμβολοσειρές, το stopwords που αφαιρεί τις λέξεις που εμφανίζονται πιο συχνά αλλά δεν ’κουβαλάνε’ συναίσθημα (π.χ. ‘had’, ‘she’, ‘all’, ‘no’, ‘when’, ‘at’,

‘any’, ‘before’, ‘them’, ‘same’) και το FreqDist που εμφανίζει την κατανομή συχνότητας των λέξεων. Δημιουργούμε ενα bar plot με τις 20 πιο συχνά εμφανιζόμενες λέξεις. Για να κατασκευάσουμε έναν διαδραστικό χάρτη των ΗΠΑ με την συχνότητα των tweets color-coded πάνω του χρησιμοποιούμε την βιβλιοθήκη Plotly. Για να βρούμε την γεωγραφική κατανομή των ανθρώπων χρησιμοποιούμε τα user profiles τους. Χρησιμοποιούμε λογαριθμική κλίμακα γιατί σε κάποιες περιπτώσεις έχουμε ακραίες τιμές (π.χ. 400+ στην Καλιφόρνια και 3 στην Βόρεια Ντακότα). Αυτές τις τρεις γραφικές παραστάσεις τις συνδυάζουμε σε μια αναλυτική μέσω του Plotly.
Σε μια κάπως διαφορετική απόπειρα μας προβλέπουμε το συναίσθημα των tweets χρησιμοποιώντας τον αλγόριθμο Random Forest της βιβλιοθήκης Sklearn. Για αρχή χρησιμοποιούμε το tweepy για να εξάγουμε τα tweets με ερώτημα αναζήτησης ’US Government’. Για σύνολο εκπαίδευσης χρησιμοποιούμε δεδομένα που βρίσκουμε στο github [3]. Προεπεξεργαζόμαστε τα δεδομένα και χρησιμοποιούμε TF - IDF για μετατροπή από κείμενο σε αριθμητικές τιμές και Random Forest για την ταξινόμηση.

6.1 TF - IDF

Term Frequency - Inverse Document Frequency ή αλλιώς συχνότητα - αντίστροφη συχνότητα εγγράφου είναι μια αριθμητική στατιστική που αντικατοπτρίζει πόσο σημαντική είναι μια λέξη σε ένα έγγραφο. Term Frequency είναι η συχνότητα με την οποία εμφανίζεται ο όρος στο έγγραφο. Inverse Document Frequency είναι ένας τρόπος να δείξουμε πόσο σημαντικός είναι ένας όρος σε ένα έγγραφο. Είναι το λογαριθμικά αντίστροφο κλάσμα των εγγράφων που περιέχουν την λέξη (το οποίο προκύπτει διαιρώντας τον συνολικό αριθμό από έγγραφα με το νούμερο των εγγράφων που περιέχουν τον όρο, και μετά παίρνοντας τον λογάριθμο αυτου του κλάσματος) [4]:

$$tfidf(t,d,D) = tf(t,d) * idf(t,D) \tag{1}$$

$$tf(t,d) = log(1 + freq(t,d)) \tag{2}$$

$$idf(t,D) = log(\frac{N}{count(d \epsilon D : t \epsilon d)}) \tag{3}$$

όπου t είναι ο όρος, d είναι το έγγραφο και D το σύνολο από έγγραφα.

6.2 Random Forest

Ο ταξινομητής Random Forest (τυχαίο δάσος) είναι ένα σύνολο ταξινομητών που χρησιμοποιούν μια σειρά από Classification and Regression Tree (CART) (δέντρα απόφασης) για να κάνει μια πρόβλεψη. Τα δέντρα που δημιουργούνται από την κατάρτιση ενός υποσυνόλου δοκιμαστικών δειγμάτων μέσω της αντικατάστασης (προσέγγιση ενσακκίσεως). Αυτό σημαίνει ότι το ίδιο δείγμα μπορεί να επιλεγεί αρκετές φορές, ενώ άλλοι δεν μπορούν. Δύο παράμετροι πρέπει να ρυθμιστούν ώστε να παραχθούν τα δέντρα στο δάσος: ο αριθμός των δέντρων απόφασης που πρόκειται να παραχθούν (Ntree) και ο αριθμός των μεταβλητών που πρέπει να επιλέγονται και να ελέγχονται για την καλύτερη διάσπαση, όταν αυξάνονται τα δέντρα (Mtry) . Θεωρητική και εμπειρική έρευνα κατέδειξε ότι η ακρίβεια ταξινόμησης επηρεάζει λιγότερο το Ntree από την παράμετρο

Mtry [5]. Η βιβλιοθήκη scikit learn μας παρέχει τον αλγόριθμο του Random Forest , στον οποίο περνάμε ως παράμετρο τον αριθμό των δέντρων (100). Χρησιμοποιούμε bootstrap, με κριτήριο το Gini impurity.

7 Συμπεράσματα

Αναλύσαμε τον ιό από διαφορετικές πλευρές και μπορέσαμε να δούμε τι αλλαγές έχει φέρει στον κόσμο και στον κάθε άνθρωπο ατομικά.
Από την ιατρική μεριά ο ιός έχει αποδειχθεί αρκετά καταστροφικός. Μέσα από τα γραφήματα βλέπουμε πως έχουν πληγεί πολλές χώρες και τα τραγικά νούμερα σε θανάτους και γενικά σε επιβεβαιωμένα κρούσματα. Ακόμα και προβλέψεις όπως με το SVM που αποτελεί από τα πιο κατάλληλα εργαλεία σε τέτοιες περιπτώσεις δεν μπόρεσε να πάρει την ίδια καμπή με αυτή του ιού. Αυτό φυσικά οφείλεται και στην ανεπάρκεια των πληροφοριών μας. Το σύστημα υγείας μιας χώρας, τα μέτρα που παίρνει η κυβέρνηση της κάθε χώρας και η υπακοή των πολιτών προς αυτά καθώς και η μέση ηλικία των πολιτών και πολλά άλλα είναι κάποιοι από τους λόγους που μία πρόβλεψη δεν μπορεί να βγει όσο σωστή θα την θέλαμε.
Από την συναισθηματική μεριά είδαμε πως πολλοί άνθρωποι είναι εξαγριωμένοι με την κατάσταση, με τις αποφάσεις της χώρας τους ή γενικά με το τι συμβαίνει τριγύρω τους. Παρόλα αυτά είδαμε και αρκετές θετικές απαντήσεις και μία αισιοδοξία προς το μέλλον.
Τέλος βλέπουμε και μέσα από την εργασία μας και γενικά στις ειδήσεις το πως ανεβαίνει ο αριθμός των αναρρώσεων, πως αρχίζει και μειώνεται η εκθετική αύξηση των κρουσμάτων και πως αρχίζουν οι πληγές του κόσμου να επουλώνονται, κάτι που όλοι περιμένουμε και ελπίζουμε να συνεχιστεί μέχρι να τελειώσει αυτός ο τρόμος.

Αναφορές

1. Chulong Li Real-time Twitter Sentiment Analysis for Brand Improvement and Topic Tracking (Chapter 1/3) https://towardsdatascience.com/real-time-twitter-sentiment-analysis-for-brand-improvement-and-topic-tracking-chapter-1-3-e02f7652d8ff Aug 26, 2019
2. Zunyou Wu, Jennifer M. McGoogan Vital Surveillances: The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020 Συνδεσμος στο πολύ ενδιαφέρον άρθρο
3. https://raw.githubusercontent.com/kolaveridi/kaggle-Twitter-US-Airline-Sentiment-/master/Tweets.csv
4. Bartosz Góralewicz: The TF*IDF Algorithm Explained https://www.onely.com/blog/what-is-tf-idf/
5. Mariana Belgiua, Lucian Drăguţb: Random forest in remote sensing: A review of applications and future directions Department of Geoinformatics Z GIS, Salzburg University
6. Rabinowitz, P.: On subharmonic solutions of a Hamiltonian system. Comm. Pure Appl. Math. 33, 609–633 (1980)
7. Ensheng Dong,Hongru Du,Lauren Gardner An interactive web-based dashboard to track COVID-19 in real time The Lancet Infectious Diseases,Elsevier,Available online 19 February 2020

8. Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, Nicholas Davies, Amy Gimma, Kevin van Zandvoort, Hamish Gibbs, Joel Hellewell, Christopher I Jarvis, Sam Clifford, Billy J Quilty, Nikos I Bosse, Sam Abbott, Petra Klepac, Stefan Flasche, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, The Lancet Infectious Diseases, 2020
9. David Baud, Xiaolong Qi ,Karin Nielsen-Saines, Didier Musso, Léo Pomar ,Guillaume Favre Real estimates of mortality following COVID-19 infection The Lancet Infectious Diseases, 2020
10. M Barstugan, U Ozkaya, S Ozturk Coronavirus (covid-19) classification using ct images by machine learning methods 2020 , https://arxiv.org/abs/2003.09424
11. Barkur, Gopalkrishna, and Giridhar B. Kamath Vibha. ”Sentiment Analysis of Nationwide Lockdown due to COVID 19 Outbreak: Evidence from India.” Asian Journal of Psychiatry (2020).