ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΙΔΙΚΕΥΣΗ: "ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ"

Ανάλυση δεδομένων COVID-19 βασισμένη σε τεχνικές ανάλυσης γράφων

ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Επιβλέπουσα Καθηγήτρια: Μαρία Χαλκίδη

Ιωάννα Κανδή, Κωνσταντίνος Μαυρογιώργος

ΑΜ: ΜΕ2136, ΜΕ2144

Email: {ioannakandi,kostismvg}@gmail.com

Περιεχόμενα

Κατάλογος Εικόνων

Περίληψη

Η παρούσα εργασία αποτελεί την απαλλακτική εργασία για το μάθημα «Εξόρυξη και Ανάλυση Δεδομένων» του 2ου εξαμήνου της μεταπτυχιακής ειδίκευσης «Προηγμένα Πληροφοριακά Συστήματα» του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς. Ειδικότερα, σκοπός της εργασίας αυτής είναι η ανάλυση δεδομένων COVID-19 βασισμένη σε τεχνικές ανάλυσης γράφων. Πιο συγκεκριμένα, στην παρούσα εργασία περιγράφεται το σύνολο δεδομένων που επιλέχθηκε και η προεπεξεργασία που εφαρμόστηκε σε αυτό. Τέλος, αναλύεται ο αντίστοιχος γράφος που δημιουργήθηκε και τα αποτελέσματα της ανάλυσης που εφαρμόστηκε πάνω σε αυτόν.

1.Θεματολογία Συστήματος

1.1 Ορισμός του προβλήματος

Η ποικιλομορφία που συναντάται πλέον στα δεδομένα έχει οδηγήσει στην ύπαρξη διαφορετικών προσεγγίσεων που επιχειρούν να τα αναλύσουν, έτσι ώστε να εξάγουν γνώση από αυτά. Ιδιαίτερης σημασίας αποτελούν τα δεδομένα που δύνανται να αναπαρασταθούν σε γράφους, καθώς αυτά συνήθως περιλαμβάνουν σύνθετες σχέσεις μεταξύ τους και, συνεπώς, η ανάλυσή τους απαιτεί μία διαφορετική προσέγγιση. Αναλύοντας γράφους και εφαρμόζοντας τεχνικές που ανήκουν στο πεδίο του «Community Detection», καθίσταται δυνατή η εύρεση ομάδων στα προαναφερθέντα δεδομένα, βάσει συγκεκριμένων χαρακτηριστικών που ενδεχομένως να έχουν ιδιαίτερη αξία για περαιτέρω μελέτη **Error! Reference source not found.**.

1.2 Στόχοι της εργασίας

Στη παρούσα εργασία, αρχικά, περιγράφεται το σύνολο δεδομένων που επιλέχθηκε. Στη συνέχεια, παρουσιάζεται η προεπεξεργασία που εφαρμόστηκε στα συγκεκριμένα δεδομένα. Τέλος, αναλύεται ο αντίστοιχος γράφος που δημιουργήθηκε και τα αποτελέσματα της ανάλυσης που εφαρμόστηκε πάνω σε αυτόν.

1.3 Δομή της εργασίας

Στο Κεφάλαιο 2 – Δεδομένα περιγράφονται τα δεδομένα που συλλέχθηκαν στα πλαίσια της εργασίας και η προεπεξεργασία που εφαρμόστηκε σε αυτά. Στο Κεφάλαιο 3 – Ανάλυση Γράφου η ανάλυση που εφαρμόστηκε στα δεδομένα και τα αποτελέσματα αυτής. Τέλος, στο Κεφάλαιο 4 – Συμπεράσματα πραγματοποιείται μία ανασκόπηση των πεπραγμένων της εργασίας και καταγράφονται μελλοντικές βελτιώσεις που θα μπορούσαν να γίνουν στη συγκεκριμένη προσέγγιση.

2. Δεδομένα

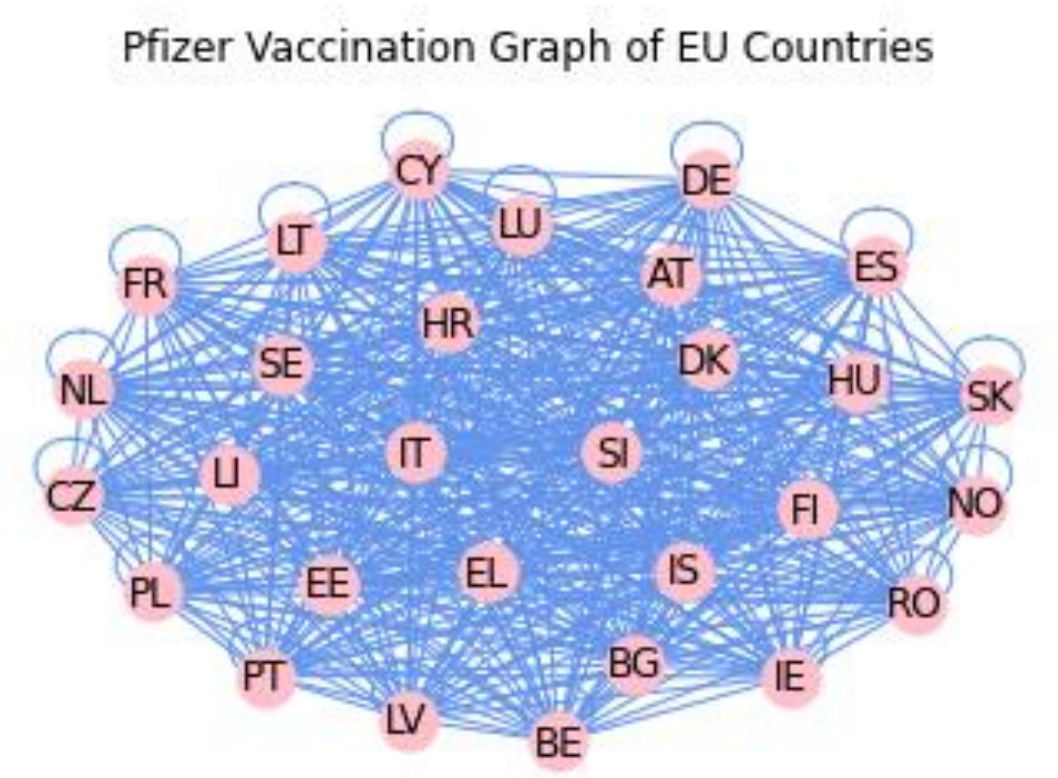
Τα δεδομένα που ανακτήθηκαν είναι διαθέσιμα εδώ και αφορούν τους εμβολιασμούς κατά της ασθένειας COVID-19 στις χώρες της Ευρωπαϊκής Ένωσης, από την έναρξη της πανδημίας το 2019 μέχρι την αρχή του 2022. Ειδικότερα, στο σύνολο των δεδομένων περιλαμβάνονται για κάθε εβδομάδα και για κάθε χώρα, οι εμβολιασμοί ανά ηλικιακή ομάδα και ανά τύπο εμβολίου (εταιρεία η οποία έχει δημιουργήσει το εμβόλιο).

Στα πλαίσια της συγκεκριμένης εργασίας, επιλέχθηκε να δημιουργηθούν ομάδες χωρών, βάσει του πλήθους των εμβολιασμών με χρήση του εμβολίου της εταιρείας «Pfizer» που διενεργήθηκαν στο γενικό πληθυσμό κάθε χώρας, από την έναρξη της πανδημίας το 2019 μέχρι και την αρχή του 2022.

Για να πραγματοποιηθεί η αντίστοιχη ανάλυση στα δεδομένα, χρειάστηκε να εφαρμοστεί συγκεκριμένη προεπεξεργασία σε αυτά με χρήση της γλώσσας προγραμματισμού Python. Αρχικά, επιλέχθηκαν οι στήλες «ReportingCountry» (αναφέρεται στην εκάστοτε χώρα της ΕΕ), «NumberDosesReceived» (αναφέρεται στο πλήθος των εμβολιασμών για κάθε χώρα), «TargetGroup» (αναφέρεται στην ηλικιακή ομάδα στην οποία πραγματοποιήθηκαν οι αντίστοιχοι εμβολιασμοί σε κάθε χώρα) και «Vaccine» (αναφέρεται στο εμβόλιο το οποίο επιλέχθηκε για τους αντίστοιχους εμβολιασμούς στην εκάστοτε χώρα). Εν συνεχεία, επιλέχθηκαν οι εγγραφές στις οποίες η στήλη «TargetGroup» είχε την τιμή «ALL», καθώς στη συγκεκριμένη ανάλυση χρειάζεται ο συνολικός αριθμός εμβολιασμών για κάθε χώρα και όχι ο διαχωρισμός του πληθυσμού ανά ηλικιακή ομάδα. Επίσης, επιλέχθηκαν οι εγγραφές στις οποίες η στήλη «Vaccine» είχε την τιμή «COM», η οποία αντιστοιχεί σε εμβόλια της εταιρείας «Pfizer». Έπειτα, τα δεδομένα ελέγχθηκαν για τυχών ακραίες τιμές. Κατά τον συγκεκριμένο έλεγχο δεν εντοπίστηκαν ακραίες τιμές. Εν συνεχεία, δεδομένου πως ζητήθηκε η δημιουργία ενός γράφου στον οποίο οι κόμβοι είναι οι χώρες και ακμές είναι η «ομοιότητα» μεταξύ τους, υπολογίστηκε η Ευκλείδεια απόσταση μεταξύ των χωρών, βάσει του αριθμού των εμβολιασμών που διενεργήθηκαν σε αυτές. Οι συγκεκριμένες αποστάσεις αποτελούν τις ακμές στον αντίστοιχο γράφο. Συνεπώς, χώρες που είχαν παρόμοιο αριθμό εμβολιασμών μεταξύ τους, φαίνεται να απέχουν μικρή απόσταση στον γράφο, ενώ αν υπάρχει μεγάλη διαφορά στον αριθμό εμβολιασμών, τότε και οι αντίστοιχοι κόμβοι απέχουν μεγαλύτερη απόσταση. Ύστερα από την προεπεξεργασία των δεδομένων προέκυψε το σύνολο δεδομένων που παρουσιάζεται στην Εικόνα 1. Στη συγκεκριμένη εικόνα φαίνεται ένα υποσύνολο από τα δεδομένα, όπου η στήλη «Source» αναφέρεται στη χώρα από την οποία «ξεκινάει» μία ακμή , η στήλη «Target» αναφέρεται στη χώρα στην οποία κατευθύνεται η συγκεκριμένη ακμή και η στήλη «Weight» αναφέρεται στο «βάρος» της ακμής (δηλαδή στην Ευκλείδεια απόσταση μεταξύ των χωρών) Για παράδειγμα, η Ευκλείδεια απόσταση μεταξύ της χώρας 1 (Αυστρία) και της χώρας 9 (Ελλάδα) είναι 3888513, ενώ η Ευκλείδεια απόσταση της Αυστρίας με τον εαυτό της είναι προφανώς 0. Αναφορικά με τον αντίστοιχο γράφο που δημιουργήθηκε, αυτός παρουσιάζεται στην Εικόνα 2.

```
Source,Target,Weight
1,1,0
1,2,6273378
1,3,17943273
1,4,21889250
1,5,2996083
1,6,122094080
1,7,16108413
1,8,21716013
1,9,3888513
1,10,63599416
1,11,11704638
1,12,112617127
1,13,17849973
1,14,10587733
1,15,15191103
1,16,22956103
1,17,71598227
1,18,23700123
1,19,17042825
1,20,22453893
1,21,21260658
1,22,8219439
1,23,12637668
1,24,36220707
1,25,2127043
1,26,5446538
1,27,3552648
1,28,20090583
1,29,15999213
```

Εικόνα 1: Ενδεικτικό παράδειγμα συνόλου δεδομένων, έπειτα από προεπεξεργασία



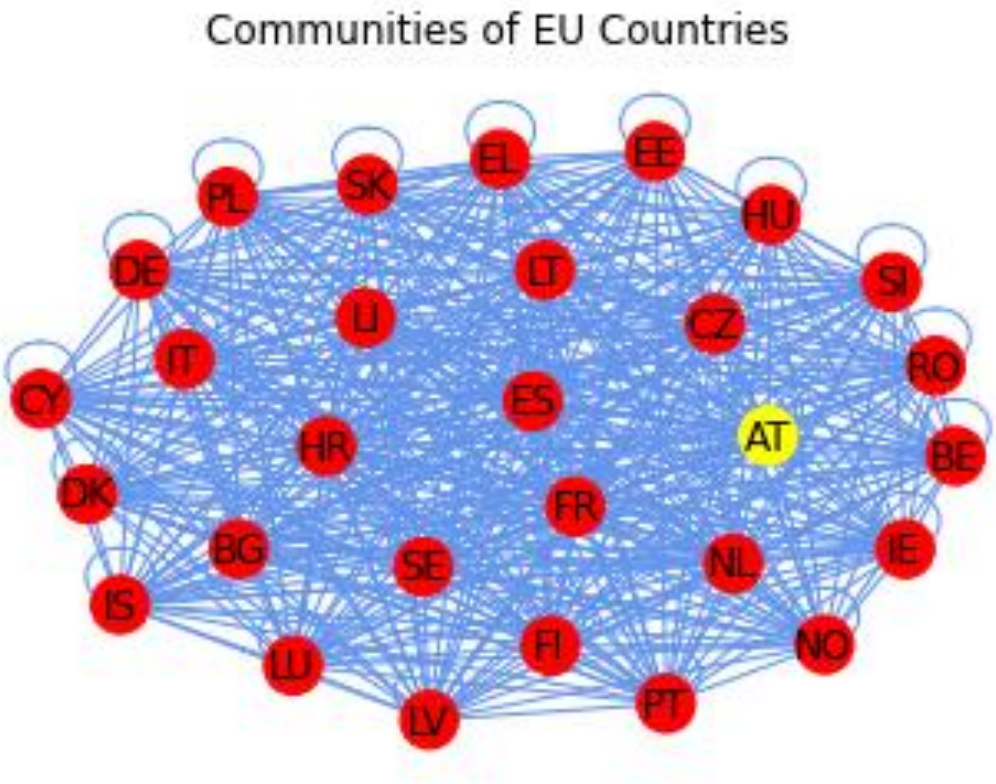
Εικόνα 2: Γράφος χωρών της ΕΕ που αξιοποίησαν το εμβόλιο της Pfizer

Οι κόμβοι του παραπάνω γράφου αντιστοιχούν στις χώρες που αναγράφονται παρακάτω:

Belgium (BE), Greece (EL), Lithuania (LT), Portugal, (PT), Bulgaria (BG), Spain (ES), Luxembourg (LU), Romania (RO), Czechia (CZ), France (FR), Hungary (HU), Slovenia (SI), Denmark (DK), Croatia (HR), Malta (MT), Slovakia (SK), Germany (DE), Italy (IT), Netherlands (NL), Finland (FI), Estonia (EE), Cyprus (CY), Austria (AT), Sweden (SE), Ireland (IE), Latvia, (LV), Poland (PL), Iceland (IS), Norway (NO)

3. Ανάλυση Γράφου

Αξιοποιώντας τον γράφο που φαίνεται στο Κεφάλαιο 2 και χρησιμοποιώντας τον αλγόριθμο Girvan-Newman **Error! Reference source not found.**, δημιουργήθηκε νέος γράφος στον οποίο οι χώρες είχαν διαχωριστεί σε communities, όπως φαίνεται στην Εικόνα 3.



Εικόνα 3: Communities χωρών βάσει του αλγόριθμου Girvan-Newman

Ειδικότερα, ο συγκεκριμένος αλγόριθμος δημιούργησε δύο communities. Στο πρώτο community ανήκει η Αυστρία και στο δεύτερο οι υπόλοιπες χώρες της ΕΕ. Εφαρμόστηκαν και άλλοι αλγόριθμοι όπως ο αλγόριθμος του Louvain **Error! Reference source not found.**. Ωστόσο, στη συγκεκριμένη περίπτωση δημιουργήθηκε μόνο ένα community το οποίο περιελάμβανε όλες τις χώρες της ΕΕ. Γι' αυτό το λόγο προτιμήθηκε ο αλγόριθμος Girvan-Newman.

Αναφορικά με τις αποστάσεις που δοκιμάστηκαν, εκτός από την Ευκλείδεια χρησιμοποιήθηκαν και οι αποστάσεις Manhattan, Minkowski και κανονικοποιημένη Ευκλείδεια. Σε όλες τις περιπτώσεις, τα communities που δημιουργήθηκαν ήταν τα ίδια που σχηματίστηκαν και με τη χρήση της Ευκλείδειας απόστασης.

Επίσης, αξίζει να σημειωθεί ότι ο κώδικας της εφαρμογής είναι ανεβασμένος και σε ιδιωτικό repository στο GitHub, στη διεύθυνση https://github.com/ioannakandi/graphAnalysis. Για να έχει κανείς πρόσβαση θα πρέπει να επικοινωνήσει σε ένα από τα emails που αναγράφονται στο εξώφυλλο της εργασίας, έτσι ώστε να προστεθεί ως contributor.

4. Συμπεράσματα

Συνοψίζοντας, η παρούσα εργασία ήταν απαραίτητη για να επιτευχθεί μια πρώτη επαφή με την ανάλυση δεδομένων βασισμένη σε τεχνικές ανάλυσης γράφων. Ωστόσο, υπάρχουν περιθώρια βελτίωσης. Ειδικότερα, ως μέτρο ομοιότητας μεταξύ των χωρών θα μπορούσαν να αξιοποιηθούν άλλες αποστάσεις, όπως η απόσταση Μανχάταν , ή ακόμα και μία εντελώς διαφορετική μετρική. Με αυτό τον τρόπο θα μπορούσαν να εφαρμοστούν εκ νέου άλλοι αλγόριθμοι, οι οποίοι στην ανάλυση που πραγματοποιήθηκε δεν ήταν αποτελεσματικοί. Επίσης, θα μπορούσαν να επιλεχθούν διαφορετικά σύνολα δεδομένων από το ίδιο πεδίο, έτσι ώστε να αξιολογηθεί η αποτελεσματικότητα της παρούσας προσέγγισης.

Βιβλιογραφία

[1] Mohamed, E. M., Agouti, T., Tikniouine, A., & El Adnani, M. (2019). A comprehensive literature review on community detection: Approaches and applications. Procedia Computer Science, 151, 295-302.

[2] Despalatović, L., Vojković, T., & Vukičević, D. (2014, May). Community structure in networks: Girvan-Newman algorithm improvement. In 2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 997-1002). IEEE.

[3] Que, X., Checconi, F., Petrini, F., & Gunnels, J. A. (2015, May). Scalable community detection with the louvain algorithm. In 2015 IEEE International Parallel and Distributed Processing Symposium (pp. 28-37). IEEE.