.

Ακρωνύμιο πρότασης:

DataPopEU

Τίτλος πρότασης:

Καινοτόμες μέθοδοι και δεδομένα υψηλής ποιότητας για τη μελέτη του λαϊκισμού και του Ευρωσκεπτικισμού

Έκθεση σχετικά με τη διαδικασία συλλογής των δεδομένων των μέσων κοινωνικής δικτύωσης (R code report)

Αύγουστος 2021

ΕΛΙΔΕΚ

ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΡΕΥΝΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΕΙΣΑΓΩΓΗ

Το έργο DataPopEU έχει ως στόχο τη συγκέντρωση και τον συνδυασμό διαφορετικών δεδομένων που προκύπτουν από το διαδίκτυο για τη μελέτη του λαικισμού και του Ευρωσκεπτικισμού στην Ελλάδα. Σκοπός της παρούσας αναφοράς είναι να παρουσιάσει τη διαδικασία με την οποία συγκεντρώνουμε τα δεδομένα από τα μέσα κοινωνικής δικτύωσης και τον κώδικα που χρησιμοποιούμε. Συγκεκριμένα, συλλέξαμε δεδομένα από δύο μέσα κοινωνικής δικτύωσης (Twitter και Facebook) και αφορούσαν είτε τις δημοσιεύσεις συγκεκριμένων λογαριασμών είτε δημοσιεύσεις που περιείχαν συγκεκριμένες λέξεις (λέξεις κλειδιά). Καθώς οι τελευταίες βουλευτικές εκλογές έγιναν το καλοκαίρι του 2019, η συλλογή των δεδομένων ξεκίνησε νωρίτερα απ' ό,τι το έργο DataPopEU (το οποίο ξεκίνησε τον Δεκέμβριο 2019). Αυτή η περίπτωση είχε προβλέφθει ήδη από τον αρχικό σχεδιασμό του έργου και έτσι ξεκινήσαμε να συλλέγουμε δημοσιεύσεις από τα μέσα κοινωνικής δικτύωσης από τις 27 Μαίου 2019. Ωστόσο, έπειτα από τους περιορισμούς που επέβαλε το Facebook το 2019 σχετικά με την εξόρυξη των δεδομένων του δεν κατέστη δυνατή η συνέχεια της συλλογής δεδομένων από τη συγκεκριμένη πλατφόρμα.

FACEBOOK

ΠΕΡΙΓΡΑΦΗ ΣΥΓΚΕΝΤΡΩΣΗΣ ΔΕΔΟΜΕΝΩΝ

Αρχικά δημιουργήσαμε μια λίστα με τα ονόματα και άλλες πληροφορίες για τους/τις υποψήφιους/ες 5 κομμάτων (ΝΔ, ΣΥΡΙΖΑ, ΚΙΝΑΛ, ΕΛΛΗΝΙΚΗ ΛΥΣΗ, ΜεΡΑ 25). Οι υποψήφιοι του ΚΚΕ δεν συμμετείχαν στη λίστα αφού η πολιτική του κόμματος αποτρέπει την ύπαρξη ατομικών λογαριασμών ή σελίδων γενικότερα στο διαδίκτυο. Στη συνέχεια για καθέναν/καθεμία από τους/τις υποψήφιους/ες έγινε αναζήτηση για να διαπιστωθεί το αν είχε δική του/της δημόσια σελίδα στο Facebook και καταχωρήθηκαν τα handles[1] ή/και τα ids για

1 Facebook handle είναι το όνομα μετά το facebook.com στην μπάρα διευθύνσεων όταν κάποιος επισκέπτεται το προφίλ

όσους/ες είχαν τελικά δημόσια ανοιχτή σελίδα. Από το σύνολο των υποψηφίων βουλευτών (1973), οι 920 είχαν μια δημόσια ανοιχτή σελίδα στο Facebook.

Χρησιμοποιώντας είτε το id, είτε το handle και την βιβλιοθήκη **Rfacebook** στην R μπορέσαμε να αντλήσουμε με αυτόματο τρόπο τα δεδομένα. Η συλλογή των δεδομένων από το Facebook έλαβε χώρα τις πρώτες δύο εβδομάδες του Ιουλίου και αφορούσε τις δημοσιεύσεις των υποψηφίων βουλευτών. Προγραμματίσαμε τον κώδικα που συγκέντρωνε τα δεδομένα να τρέχει αυτόματα κάθε 15 λεπτά μέχρι να τελειώσει η συγκέντρωσή τους. Εξαιτίας των προαναφερόμενων περιορισμών, η συλλογή των δεδομένων ήταν εφικτή μόνο στις περιπτώσεις που ο υποψήφιος/α διατηρούσε δημόσια ανοιχτή σελίδα στο Facebook. Οι δημοσιεύσεις των υποψηφίων που συγκεντρώσαμε αφορούσαν την περίοδο 27 Μαΐου εώς 7 Ιουλίου 2019. Επιλέχθηκε η 27η Μαΐου γιατί μετά το αποτέλεσμα των Ευρωεκλογών της 26ης Μαΐου ήταν πλέον ξεκάθαρο ότι πολύ σύντομα θα γίνονταν εθνικές εκλογές.

Όπως προαναφέρθηκε οι δημόσια ανοιχτές σελίδες στο Facebook ήταν 920. Από τους 920 συγκεντρώσαμε δημοσιεύσεις από 903 υποψηφίους. Οι υπόλοιποι 17 δεν είχαν καμία δημοσίευση στη σελίδα τους. Ο μέγιστος αριθμός δημοσιεύσεων που καταφέραμε να συλλέξουμε για κάποιους υποψήφιους είναι 25 καθώς αυτός ήταν ο μέγιστος αριθμός που επέτρεπε το Facebook API. Ωστόσο, κάποιοι υποψήφιοι/ιες είχαν λιγότερες από 25 δημοσιεύσεις στο διάστημα για το οποίο ενδιαφερόμαστε. Για αυτούς/ες ήταν δυνατόν να συγκεντρωθούν όλες οι δημοσιεύσεις τους. Στις δημοσιεύσεις των υποψηφίων θα αναζητηθούν τα tweets που μπορεί να σχετίζονται με τον λαϊκισμό και τον Ευρωσκεπτικισμό.

Ωστόσο, μετά τη συλλογή των προεκλογικών δημοσιεύσεων δεν ήταν πλέον δυνατή η περαιτέρω συλλογή δεδομένων από το Facebook. Συγκεκριμένα, μετά το σκάνδαλο με τη διάθεση των δεδομένων του Facebook στην Cambridge Analytica και τον ρόλο τους στην προεκλογική εκστρατεία των υποψηφίων για την προεδρία στην Αμερική το 2016, η πολιτική του Facebook σχετικά με την εξόρυξη των δεδομένων του άλλαξε. Ήδη από το 2018 άρχισαν να επιβάλλονται κάποιοι περιορισμοί στη διαθεσιμότητα των δεδομένων του. Μέχρι το καλοκαίρι του 2019 ήταν δυνατή μόνο η συγκέντρωση δεδομένων από δημόσια ανοιχτές σελίδες και δεν ήταν εφικτό πλέον να συγκεντρωθούν δεδομένα με λέξεις-κλειδιά που αφορούν τον λαϊκισμό και τον Ευρωσκεπτικισμό, ούτε δεδομένα από προσωπικούς λογαριασμούς στο Facebook (όπως

αρχικά είχαμε σχεδιάσει).

Το Facebook, από τον Σεπτέμβριο του 2019 κι έπειτα, επέβαλε ακόμα πιο αυστηρούς περιορισμούς στην εξόρυξη των δεδομένων του και αυτό κατέστησε απαγορευτική τη συλλογή των δεδομένων του (τουλάχιστον από ερευνητές) με αυτοματοποιημένο τρόπο. Έτσι τον Σεπτέμβριο του 2019, το Facebook API που είχαμε δημιουργήσει για τη συγκέντρωση των δεδομένων κατέστη προς το παρόν ανενεργό προκειμένου να δοθεί νέα άδεια για τη λειτουργία του από το Facebook. Το API δεν ξαναπήρε έγκριση, παρά τις προσπάθειές μας, καθώς η πολιτική του Facebook εκείνο τον καιρό ευνοούσε κυρίως τη συλλογή δεδομένων από επιχειρήσεις παρά από ερευνητές. Τον τελευταίο καιρό, το Facebook έχει αρχίσει και κάνει κάποια βήματα σχετικά με τη διαθεσιμότητα των δεδομένων του στους ερευνητές. Για παράδειγμα έχει δώσει πρόσβαση σε συγκεκριμένες βάσεις δεδομένων και έχει ανακοινώσει τη δημιουργία Researcher API μέσα στο 2021. Προς το παρόν αυτές οι δράσεις αφορούν μόνο δεδομένα που παράγονται γεωγραφικά από τις ΗΠΑ.

ΚΩΔΙΚΑΣ ΓΙΑ ΤΗ ΣΥΓΚΕΝΤΡΩΣΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ FACEBOOK

```
# This cron was running every 15 minutes
# load libraries
library("Rfacebook")
library("readr")
library("dplyr")
# load facebook handles
load("face_handle")
# load the row number of previous handle
load("last_a")
# increasing by 1 to get the next facebook page
start_a<-last_a+1
# load credentials for facebook api
load("my_oauth")
# get facebook page data
```

```
temp_fb_posts_parl19 <-getPage(face_handle[start_a], token = my_oauth,  since='2019/05/27', until='2019/07/07')
# next few lines run only the first time
# to create the object fb_posts_parl19
# fb_posts_parl19<-temp_fb_posts_parl19
# save("fb_posts_parl19", file ="fb_posts_parl19")
# load facebook posts already mined
load("fb_posts_parl19")
# bind new data to previous data
fb_posts_parl19 <- fb_posts_parl19 %>%
 bind_rows(temp_fb_postsocial_media_data_report.odts_parl19)
# save data
save(fb_posts_parl19, file = "fb_posts_parl19")
# save row number of facebook handle from this session
last_a <-start_a
save("last_a", file= "last_a")
```

TWITTER

ΠΕΡΙΓΡΑΦΗ ΣΥΓΚΕΝΤΡΩΣΗΣ ΔΕΔΟΜΕΝΩΝ

Η συλλογή των δεδομένων από το Twitter έχει ξεκινήσει επίσης από την 27η Μαίου 2019 και πραγματοποιείται με δύο βασικούς τρόπους.

Ο πρώτος αφορά τους λογαριασμούς των υποψηφίων βουλευτών στις εκλογές του 2019. Ήδη από όταν έγινε φανερό ότι τον Ιούλιο του 2019 η χώρα πήγαινε σε βουλευτικές εκλογές δημιουργήθηκε μια λίστα με τα ονόματα των υποψηφίων βουλευτών των 5 κοινοβουλευτικών κομμάτων (ΝΔ, ΣΥΡΙΖΑ, ΚΙΝΑΛ, ΕΛΛΗΝΙΚΗ ΛΥΣΗ, ΜεΡΑ 25). Οι υποψήφιοι του ΚΚΕ, όπως και στη λίστα για τα Facebook handles των υποψηφίων, δεν συμμετείχαν αφού η πολιτική του κόμματος αποτρέπει την ύπαρξη ατομικών λογαριασμών ή σελίδων γενικότερα στο διαδίκτυο. Στη συνέχεια ξεκίνησε η αναζήτηση των λογαριασμών των υποψηφίων στο Twitter και συγκεντρώθηκαν τα Twitter handles για όσους/ες υποψήφιους/ες είχαν λογαριασμό στο

Twitter. Με τα Twitter handles των υποψηφίων και χρησιμοποιώντας τη βιβλιοθήκη **rtweet** ξεκίνησε η συλλογή των tweets των υποψηφίων βουλευτών. Στις 27 Μαίου 2019, ξεκινήσαμε να συγκεντρώνουμε tweets από όσα Twitter handles βρήκαμε αρχικά και στη συνέχεια προσθέταμε και καινούργια όσο η αναζήτησή μας οδηγούσε στην εύρευση λογαριασμών Twitter και άλλων υποψηφίων βουλευτών. Η συλλογή των tweet των υποψηφίων βουλευτών συνεχίζεται μέχρι και σήμερα και θα συνεχιστεί μέχρι το τέλος του έργου DataPopEU.

Συνολικά βρήκαμε 540 λογαριασμούς υποψηφίων στο Twitter αλλά δεν ήταν όλοι δημόσια ανοιχτοί. Μπορέσαμε να μαζέψουμε δεδομένα κάποια στιγμή από όταν ξεκινήσαμε την αναζήτηση για περίπου 500 από αυτούς. Συνολικά έχουν μαζευτεί περίπου 300000 tweets. Ωστόσο, υπάρχει περίπτωση οι αριθμοί αυτοί να αλλάξουν όταν θα γίνει ο τελικός καθαρισμός των δεδομένων για παράδειγμα μπορεί να έχουν περάσει tweets που να έχουν γίνει σε ημερομηνία εκτός των χρονικών ορίων της έρευνας ή μπορεί κάποιος υποψήφιος να έχει διαγράψει τον λογαριασμό του κάποια στιγμη στη διάρκεια της συλλογής των δεδομένων.

Ο δεύτερος τρόπος συλλογής δεδομένων από το Twitter αφορά την αναζήτηση συγκεκριμένων tweets με λέξεις κλειδιά ή φράσεις ή ακόμα και συνδυασμούς λέξεων που σχετίζονται με τον λαϊκισμό και τον Ευρωσκεπτικισμό. Η συλλογή αυτών των δεδομένων ξεκίνησε το φθινόπωρο του 2020, όταν καταρτίστηκε το Λεξικό για την αναγνώριση του λαϊκισμού και του Ευρωσκεπτικισμού (Παραδοτέο 3.1), το οποίο αποτέλεσε και τη βάση για τις λέξεις κλειδιά που επιλέχθηκαν για τη συγκέντρωση των tweets. Το λεξικό είναι στην ουσία το αποτέλεσμα της λεξικομετρικής ανάλυσης άρθρων του έντυπου τύπου που αφορούσαν το λαϊκισμό και τον Ευρωσκεπτικισμό. Ωστόσο, η εκφορά του λόγου είναι πολύ διαφορετική στον τύπο από ότι στο Twitter εξαιτίας συγκεκριμένων χαρακτηριστικών του όπως ο περιορισμός των χαρακτήρων, το πολύ μεγάλο εύρος της θεματολογίας κτλ. Γενικά λοιπόν, ο λόγος στο Twitter είναι πολύ πιο σύντομος, άρα και πιο περιεκτικός, πιο καθημερινός και υπάρχει η δυνατότητα διαλόγου. Επίσης, μπορεί να πραγματεύεται πολλά και διαφορετικά θέματα με αποτέλεσμα η σημασία των λέξεων να είναι αρκετά ευμετάβλητη σε αντίθεση με τον έντυπο τύπο στον οποίο συνήθως υπάρχει μια συγκεκριμένη θεματολογία.

Σύμφωνα με τα παραπάνω λοιπόν, χρειάστηκε οι λέξεις του λεξικού να αναπροσαρμοστούν για να χρησιμοποιηθούν για την αναζήτηση των tweets. Για να φέρει η

αναζήτησή μας το επιθυμητό αποτέλεσμα, δηλαδή tweets που να σχετίζονται με τον λαϊκισμό και τον Ευρωσκεπτικισμό και να προέρχονται από πολίτες ο οποίοι έχουν περισσότερες πιθανότητες να είναι και ψηφοφόροι, αναζητήσαμε χειροκίνητα την καθεμία από τις λέξεις αυτές στο Twitter για να διαπιστώσουμε εάν όντως τα tweets που την χρησιμοποιούσαν ήταν σχετικά με την έρευνά μας. Επιπλέον, αναζητήσαμε και τις πιο συχνές λέξεις όπως αυτές είχαν προκύψει από το Populismus και χρησιμοποιήθηκαν και για τον εντοπισμό των άρθρων που τελικά αποδελτιώθηκαν στο έργο DataPopEU. Με αυτόν τον τρόπο έγιναν αμέσως φανερές οι πρώτες προσαρμογές που έπρεπε να γίνουν για να φέρει η αναζήτηση το επιθυμητό αποτέλεσμα.

Έτσι λοιπόν οδηγηθήκαμε στον αποκλεισμό των αγγλικών λέξεων καθώς οδηγούσαν σε δημοσιεύσεις που μπορεί να μην σχετίζονταν με τα υπό μελέτη φαινόμενα ή να αφορούσαν άλλες χώρες π.χ. money, sovereignty κτλ. Επιπλέον, αποκλείσαμε λέξεις που μπορεί το νόημα τους να αλλάξει πάρα πολύ ανάλογα με το περιεχόμενο της φράσης π.χ. αριστερά, κόσμος κ.α. γιατί ο τρόπος με τον οποίο χρησιμοποιούνταν στα περισσότερα tweets δεν αφορούσε καθόλου τα υπό μελέτη φαινόμενα. Προσθέσαμε κάποιες λέξεις που παραδοσιακά μπορεί να σχετίζονται με τον λαϊκιστικό και τον Ευρωσκεπτικιστικό λόγο και δεν υπήρχαν στο λεξικό.

Ψάχνοντας στα hashtag του Twitter που αφορούσαν πολιτικά θέματα βρήκαμε κάποιες λέξεις που χρησιμοποιούνταν από τους χρήστες του Twitter περισσότερο όταν οι δημοσιεύσεις τους αφορούσαν τον λαϊκισμό και τον Ευρωσκεπτικισμό. Σε αυτό το πρώτο στάδιο τα κριτήρια που χρησιμοποιήσαμε για να επιλέξουμε τις συγκεκριμένες λέξεις ήταν σχετικά απλά. Για τον λαϊκισμό επιλέξαμε λέξεις που αναφέρονταν συχνά σε tweet που έβρισκαν είτε εσωτερικούς, είτε εξωτερικούς "εχθρούς" του λαού. Σε κάθε περίπτωση φαίνεται ότι με κάποιες λέξεις κλειδιά είναι πιο πιθανό να ανιχνευθούν tweets για το συγκεκριμένο φαινόμενο από ότι με άλλες.

Στον παρακάτω πίνακα μπορεί κανείς να δει τις λέξεις κλειδιά με βάση τις οποίες συγκεντρώνουμε τα tweets. Όπως αναφέρθηκε και νωριτέρα οι λέξεις κλειδιά προέρχονται από τέσσερις βασικές πηγές. Όλες οι λέξεις στην αναζήτηση έχουν μικρά γράμματα και δεν έχουν τόνους (και διαλυτικά) καθώς με αυτές τις ρυθμίσεις μπορούμε να εντοπίσουμε τις ίδιες λέξεις και όταν γράφονται με κεφαλαία και τόνους. Επίσης, υπάρχουν θέματα (π.χ. ευρωπ) τα οποία αρχικά χρησιμοποιήσαμε προσθέτοντας διάφορες παραλλαγές στην κατάληξη όπως ευρωπαικος, ευρωπαικα, ευρωπαικες κτλ.

Πίνακας 1. Λέξεις κλειδιά για την αναζήτηση των tweets

Λέξεις που προέκυψαν από την ανάλυση του Populismus	Λέξεις που προέκυψαν από το λεξικό του DataPopEU	Λέξεις που μπορεί να σχετίζονται με την εκφορά απόψεων για τον λαό και την Ευρώπη
ευρω	κομισιον	συνοδος κορυφης
ευρωπη/ς	εε	συνοδο κορυφης
ευρωζώνη	εκτ	ελιτ
ευρωπ	βρυξέλλες	μαζες
λαος/ου	μπρεξιτ	κατεστημενο
λαικισμος/ου/ο	brexit	διαθφορα/ας, διεθφαρμενοι
λαικιστης/η/ες/ων, λαικιστικη/ες/ων/ο/α, λαικιζει/ετε/ουν	μερκελ	διαπλοκη
(εθνικη) κυριαρχια	(συνοδος) κορυφης	υποχειρια
λιτοτητα	πολιτικη	καναλαρχες/ων
συστημα	συμφεροντΔΗΛΩΣΗ COVIDα	τραπεζιτες
δεξια		αποκλεισμενοι
		αδυναμοι
		γεωπολιτικα
		ελληνορωσικες σχεσεις

Η τρίτη στήλη του Πίνακα 1 αποτελείται από λέξεις που επιλέξαμε από πολλές και διαφορετικές πηγές όπως το έργο Team Populism[2] .

Καθώς το Twitter αποτελεί έναν ανοιχτό δίαυλο επικοινωνίας με απεριόριστη θεματολογία, η ανανοηματοδότηση διαφόρων λέξεων ανάλογα με τη μόδα, με αυτόν που τις χρησιμοποιεί και με τα γεγονότα της επικαιρότητας είναι κάτι που συμβαίνει συνέχεια. Άρα υπάρχει πιθανότητα, σε συγκεκριμένη χρονική στιγμή μια λέξη-κλειδί να χρησιμοποιείται για ένα συγκεκριμένο θέμα και να αφορά λαϊκιστικό λόγο, αλλά σε άλλη χρονική στιγμή να χρησιμοποιείται σε tweets που δεν έχουν καμία σχέση με τον λαϊκισμό. Έχοντας αυτό στο

2 Για περισσότερες πληροφορίες https://populism.byu.edu/

μυαλό επιλέξαμε τις παραπάνω λέξεις για να συγκεντρώσουμε tweets αλλά όταν τελειώσει η συλλογή των δεδομένων ίσως χρειαστεί να γίνει περαιτέρω καθαρισμός έτσι ώστε να κρατηθούν μόνο τα σχετικά tweets.

ΚΩΔΙΚΑΣ ΓΙΑ ΤΗ ΣΥΓΚΕΝΤΡΩΣΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER

ΔΗΜΟΣΙΕΥΣΕΙΣ ΥΠΟΨΗΦΙΩΝ ΒΟΥΛΕΥΤΩΝ

```
# load libraries
library("rtweet")
library("tidyverse")
# candidates' Twitter handles are divided in three objects
# because they are too many for getting their tweets in one request
# load canandidates1
load("candidates1")
# load canandidates2
load("candidates2")
# load canandidates3
load("candidates3")
# load maximum status_id
load("new_max_status_id")
# get tweets for candidates1
tweets_1 <- get_timelines(candidates1, since_id = new_max_status_id)
# get tweets for candidates2
tweets_2  <- get_timelines(candidates2, since_id = new_max_status_id)
# get tweets for candidates3
tweets_3  <- get_timelines(candidates3, since_id = new_max_status_id)
# bind all tweets mined
temp<-tweets_after_1 %>%
  bind_rows(tweets_after_2) %>%
```

```
  bind_rows(tweets_after_3)
# keep the new maximum status id
new_max_status_id<-max(temp$status_id)
# save new maximum status id
save(new_max_status_id, file = "new_max_status_id")
# when the code run first time
# temp was renamed and then saved
# tweets_of_candidates<-temp
# save(tweets_of_candidates, file = "tweets_of_candidates")
# after the first time
# load the tweets
load("tweets_of_candidates")
# bind new tweets to the old ones
tweets_of_candidates <- tweets_of_candidates %>% bind_rows(temp)
# save again the tweets
save(tweets_of_candidates, file = "tweets_of_candidates")
```

ΑΝΑΖΗΤΗΣΗ ΔΗΜΟΣΙΕΥΣΕΩΝ ΜΕ ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

```
# load libraries
library(purrr)
library(tidyverse)
library(rtweet)

# function to get the tweets
get_tweets <- function(x) {
  root_query <- search_tweets(x, n = 18000, type = "mixed", retryonratelimit = TRUE)
  return(root_query)
}

# get tweets
```

```
tweets <- map(keywords, get_tweets)
# name the lists
names(tweets) <-paste0(files_names, format(Sys.time(), "%Y-%m-%d-%H-%M"))
# save each tibble separately
map(.x = names(tweets), .f = function(x){
  assign(x, tweets[[x]])
   save(list = x, file = paste0("tweets_collected/", x, format(Sys.time(), "%Y-%m-%d-%H-%M"),
".RData"))
})
```