

**ΣΥΛΛΟΓΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ**

**ΑΝΑΦΟΡΑ ΕΡΓΑΣΙΑΣ BIG-DATA 2022-2023**

**TWITTER DATA MINING**

Finding the frequency of “mentions” of Greek political parties on twitter using python.

**Πρόλογος:**

Οι εκλογές αποτελούν μια κρίσιμη περίοδο για κάθε χώρα, και το Twitter έχει γίνει ένας σημαντικός χώρος για την έκφραση απόψεων και την

ανταλλαγή πληροφοριών σχετικά με τους υποψηφίους, τα κόμματα και τις εκλογικές διαδικασίες γενικότερα. Το data mining σε εκλογικά Twitter αξιοποιεί τον τεράστιο όγκο των δεδομένων που δημιουργούνται καθημερινά στην πλατφόρμα αυτή για να εξάγει συμπεράσματα και να ανακαλύψει τάσεις.

Στη παρακάτω αναφορά χρησιμοποιούνται εργαλεία για την εξαγωγή δεδομένων από τη πλατφόρμα twitter για τις Ελληνικές εκλογές του 2023.

Συγκεκριμένα για την εξαγωγή της συχνότητας αναφορών των 4ων μεγαλύτερων κομμάτων, προέδρων αυτών των κομμάτων, (αυτής της

στιγμής στην Ελλάδα) ξεχωριστά και οπτικοποίηση των αποτελεσμάτων στη συνέχεια με τη χρήση κατάλληλων εργαλείων στη γλώσσα προγραμματισμού Python.

Τα αποτελέσματα της ανάλυσης αυτής θα μπορούσαν να αξιοποιηθούν για την κατανόηση της δυναμικής των εκλογών, την απήχηση των

υποψηφίων και των πολιτικών μηνυμάτων, καθώς πιθανώς και την

πρόβλεψη αποτελεσμάτων εκλογών. Αναλύσεις όπως αυτή θα μπορούσε να βοηθήσει τους πολιτικούς να προσαρμόσουν τις εκλογικές τους

εκστρατείες, να απευθύνουν τις προσπάθειες τους και να επικοινωνήσουν

αποτελεσματικά τις ιδέες τους στους πολίτες. Παρότι το data mining στο

Twitter παρέχει πολύτιμες πληροφορίες για εκλογικά δεδομένα, είναι σημαντικό να αναφερθεί ότι απαιτείται προσεκτική διαχείριση και ερμηνεία των αποτελεσμάτων. Επίσης, πρέπει να λαμβάνονται υπόψη ζητήματα

απορρήτου, δεοντολογίας και προστασίας των δεδομένων, προκειμένου να διασφαλιστεί η ενδεδειγμένη χρήση των πληροφοριών και η προστασία των δικαιωμάτων των χρηστων.

**Εργαλεία που χρησιμοποιήθηκαν:**

Αρχικά για την αναπτυξη και τη διαχειριση αυτου του project χρησιμοποιήθηκε το Anaconda και Jupyter Notebook για μια πιο σωστή διαχείριση πακέτων και διευκόλυνση στην ανάπτυξη του προγράμματος

(**Python**). Τα πακετα και οι βιβλιοθηκες που χρησ/καν στον κωδικα ειναι η **tweepy**, η **pandas**, η **re** και η **matplot**.

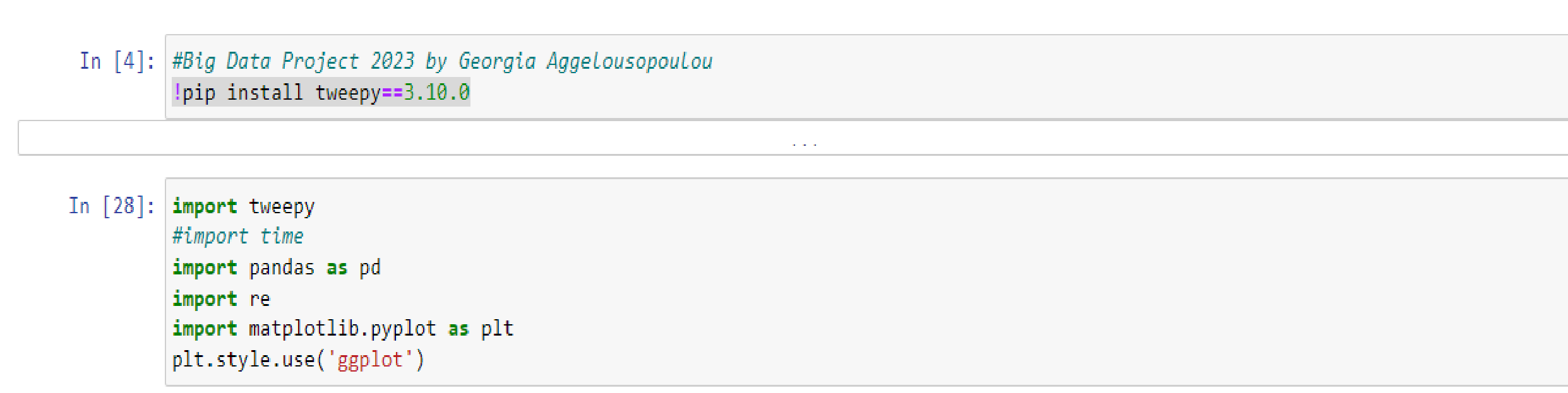
* Βιβλιοθήκη **tweepy**: παρέχει μια εύκολη διεπαφή για το API του Twitter μέσω της Python. Εδώ χρησιμο/ται για την αναζητηση tweets που πληρούν κάποιες προϋποθέσεις.
* Βιβλιοθήκη **pandas**: βιβλιοθήκη για την ανάλυση και τον χειρισμό δεδομένων. Παρέχει δομές δεδομένων όπως τα DataFrames, τα οποία επιτρέπουν την οργάνωση, το φιλτράρισμα, την ανάλυση και την επεξεργασία των δεδομένων με ευκολία. Σε αυτήν την περίπτωση, των tweets.
* Βιβλιοθήκη **re**: παρέχει υποστήριξη για τις κανονικές εκφράσεις. Επιτρέπει τον εντοπισμό και την επεξεργασία κειμένου, την αναζήτηση και την αντικατάσταση μοτίβων (tweets).
* Βιβλιοθήκη **MatPlot**: βιβλιοθήκη οπτικοποίησης δεδομένων. Το pyplot παρέχει μια διεπαφή για τη δημιουργία διαγραμμάτων και γραφικών παραστάσεων.

Παρακάτω θα γίνει μια αναλυτική επεξήγηση του κώδικα …

**Ανάλυση του κωδικα:**

Αρχίζουμε με την εντολή “ !pip install tweepy==3.10.0 “. Γίνεται η εγκαταστήσει της βιβλιοθήκης **Tweepy**, έκδοση 3.10.0. Κατά την εκτέλεση της εντολής θα ελεγχθεί αν η βιβλιοθήκη Tweepy στην έκδοση 3.10.0 δεν

είναι ήδη εγκατεστημένη στο περιβάλλον εκτέλεσης σας. Αν δεν είναι, τότε θα ξεκινήσει η διαδικασία λήψης και εγκατάστασης της συγκεκριμένης έκδοσης της βιβλιοθήκης.

Αμέσως μετα εισάγουμε τις απαραίτητες βιβλιοθήκες.

Σημειωση:

Με import pandas as pd, ανάθεση της βιβ/κης στη μεταβλητή pd

Με import matplotlib.pyplot as plt, ανάθεση της βιβ/κης στη μεταβλητή plt. Τέλος Η εντολή “ plt.style.use('ggplot') “

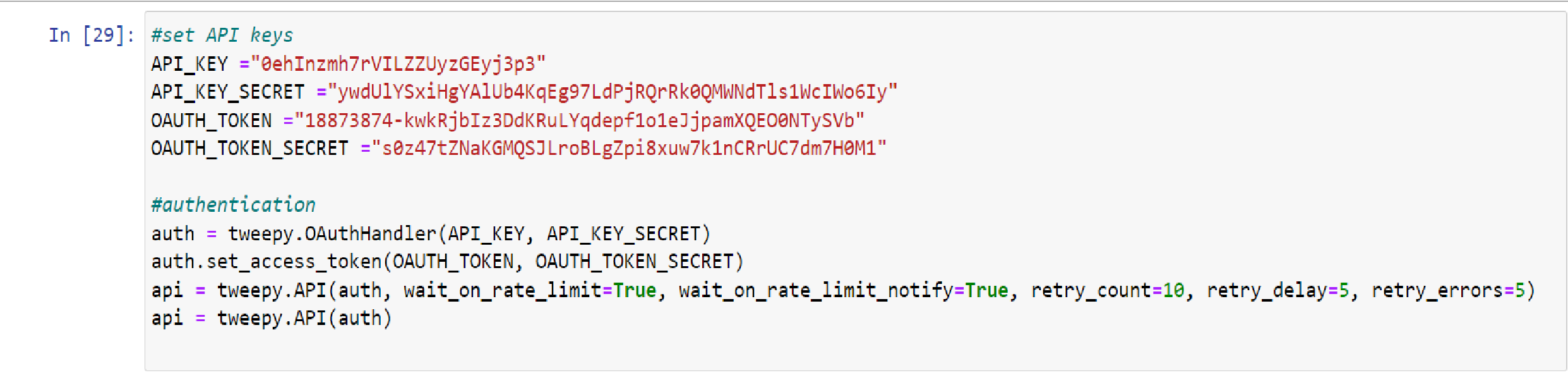
χρησιμοποιείται για να επιλέξει το στυλ της γραφικής αναπαράστασης στο Matplotlib. Συγκεκριμένα, η εντολή αυτή ορίζει το στυλ "ggplot".

Επειτα τη ρύθμιση και πιστοποίηση των κλειδιών του API του

Twitter και στη δημιουργία μιας σύνδεσης με το API χρησιμοποιώντας τη βιβλιοθήκη Tweepy.

Στην αρχή, δηλώνονται τα κλειδιά του API του Twitter:

Στη συνέχεια, γίνεται η διαδικασία πιστοποίησης και δημιουργίας μιας σύνδεσης με το API του Twitter με χρήση των παραπάνω κλειδιών:

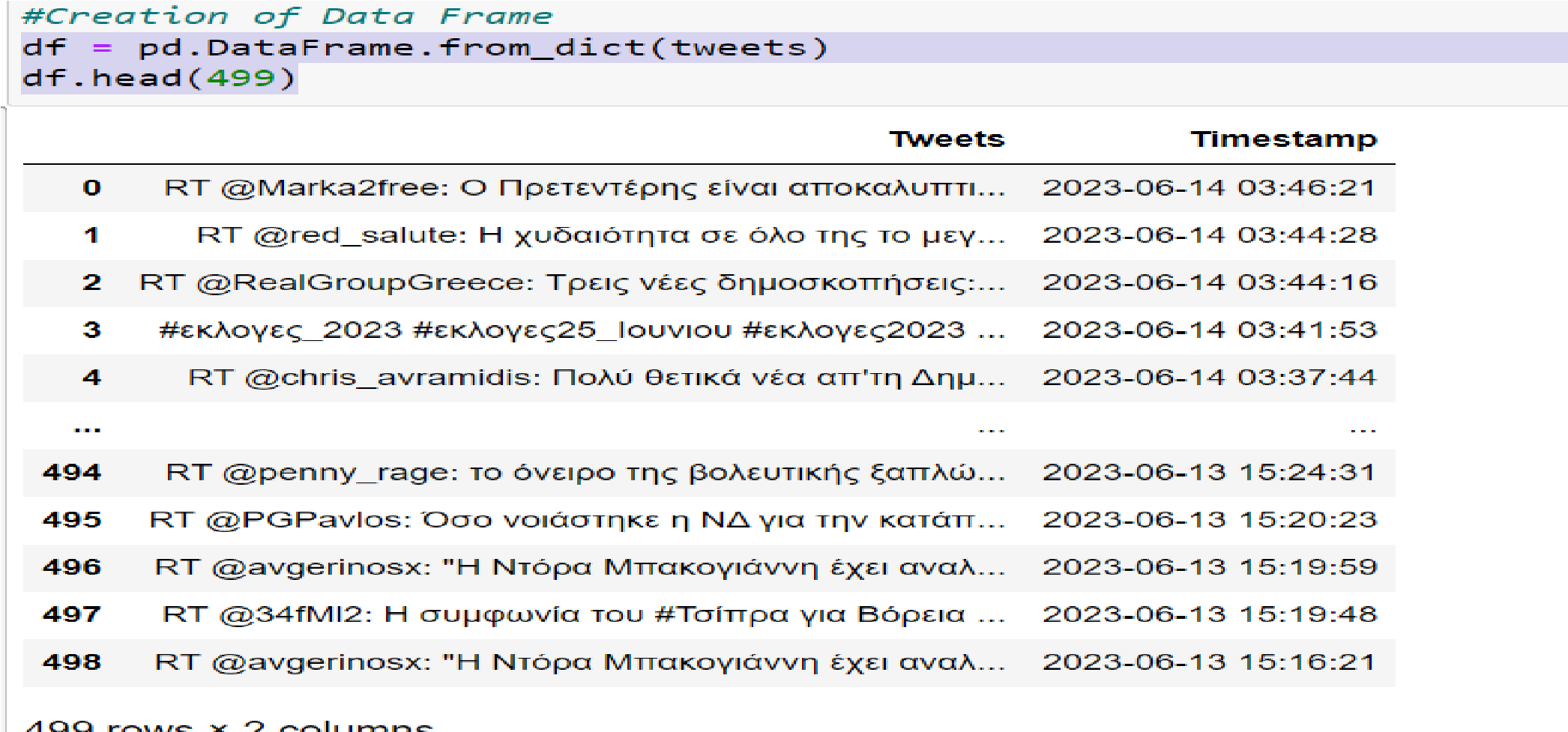
* Η εντολή auth = tweepy.OAuthHandler(API\_KEY, API\_KEY\_SECRET)` δημιουργεί ένα αντικείμενο πιστοποίησης OAuthHandler με τα κλειδιά πρόσβασης του API.
* Η εντολή `auth.set\_access\_token(OAUTH\_TOKEN, OAUTH\_TOKEN\_SECRET)` ορίζει τα τοκεν πρόσβασης στο αντικείμενο πιστοποίησης.
* Η εντολή `api = tweepy.API(auth)` δημιουργεί ένα αντικείμενο API χρησιμοποιώντας την πιστοποίηση που έχει οριστεί.

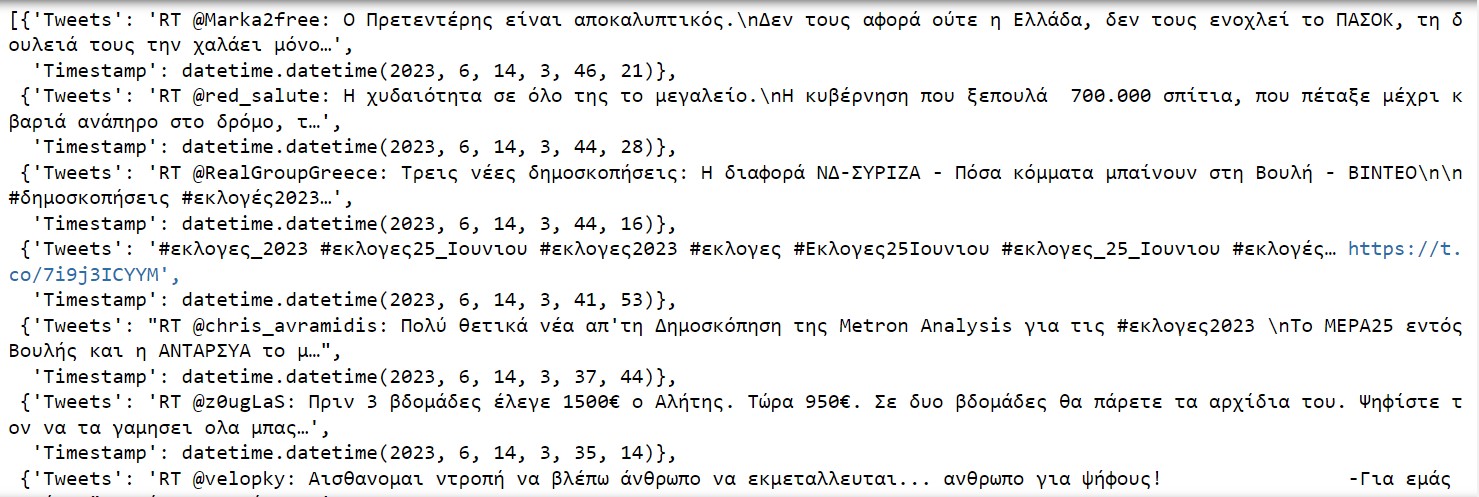
Συνεχίζουμε θέτοντας τις παραμέτρους για την συλλογή των επιθυμητών tweets.

* Ορίζεται η μεταβλητή hashtag που περιέχει το #ekloges2023.

Επειτα χρησιμοποιείται η κλαση ‘Cursor’ του Tweepy για τη δημιουργία ενος iterator για την αναζήτηση των tweets με τη χρήση της μεθόδου api.search βάση του συγκεκριμένου hashtag και των ορισμάτων τα οποία θέλουμε να είναι η γλώσσα, η ημερομηνία αναζήτησης των tweets και η

κωδικοποίηση τους. Επίσης, με τη μέθοδο items ζητάμε από το iterator 500 tweets.

* Τέλος δημιουργείται η λίστα tweets (τα tweets και το



timestamp) οπου και την εμφανιζουμε.

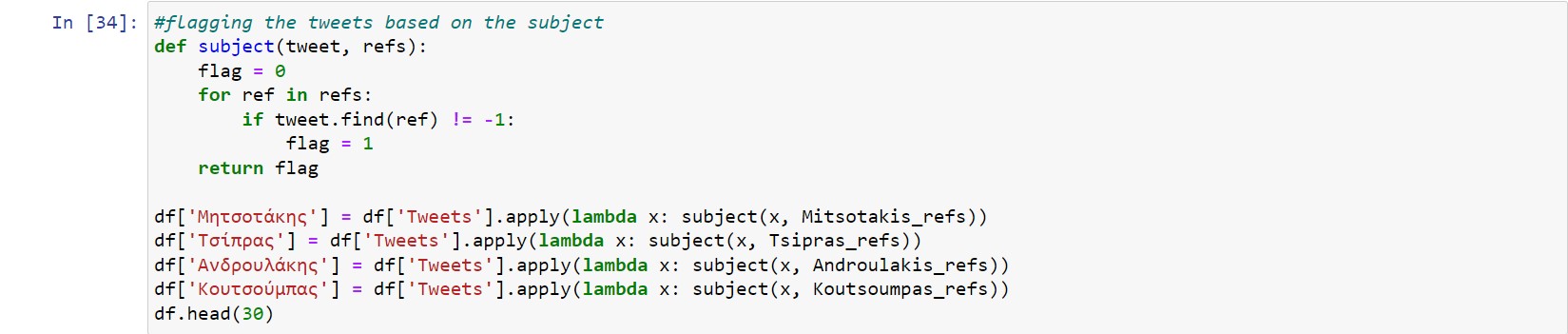
* Μετά δημιουργούμε ένα DataFrame στη βιβλιοθήκη Pandas από το λεξικό tweets που τα περιέχει. Στη συνέχεια επιστρέφουμε τα πρώτα 499.
* Στη συνέχεια κάνουμε refreshing δηλαδή αναθέτουμε σε μεταβλητές λίστες αναφοράς. Συγκεκριμένα:
  + Mitsotakis\_refs: Αναφορές που σχετίζονται με τον Κυριάκο Μητσοτάκη και το κόμμα Νέα Δημοκρατία (ΝΔ).
  + Tsipras\_refs: Αναφορές που σχετίζονται με τον Αλέξη Τσίπρα και το κόμμα ΣΥΡΙΖΑ.
  + Androulakis\_refs: Αναφορές που σχετίζονται με τον Νίκο Ανδρουλάκη και το κόμμα Πανελλήνιο Σοσιαλιστικό Κίνημα (ΠΑΣΟΚ).
  + Koutsoumpas\_refs: Αναφορές που σχετίζονται με τον Δημήτρη Κουτσούμπα και το Κομμουνιστικό Κόμμα Ελλάδας (ΚΚΕ).
* Στο επόμενο τμήμα του κώδικα, δημιουργούνται νέες στήλες στο

DataFrame με τα ονόματα "Μητσοτάκης", "Τσίπρας",

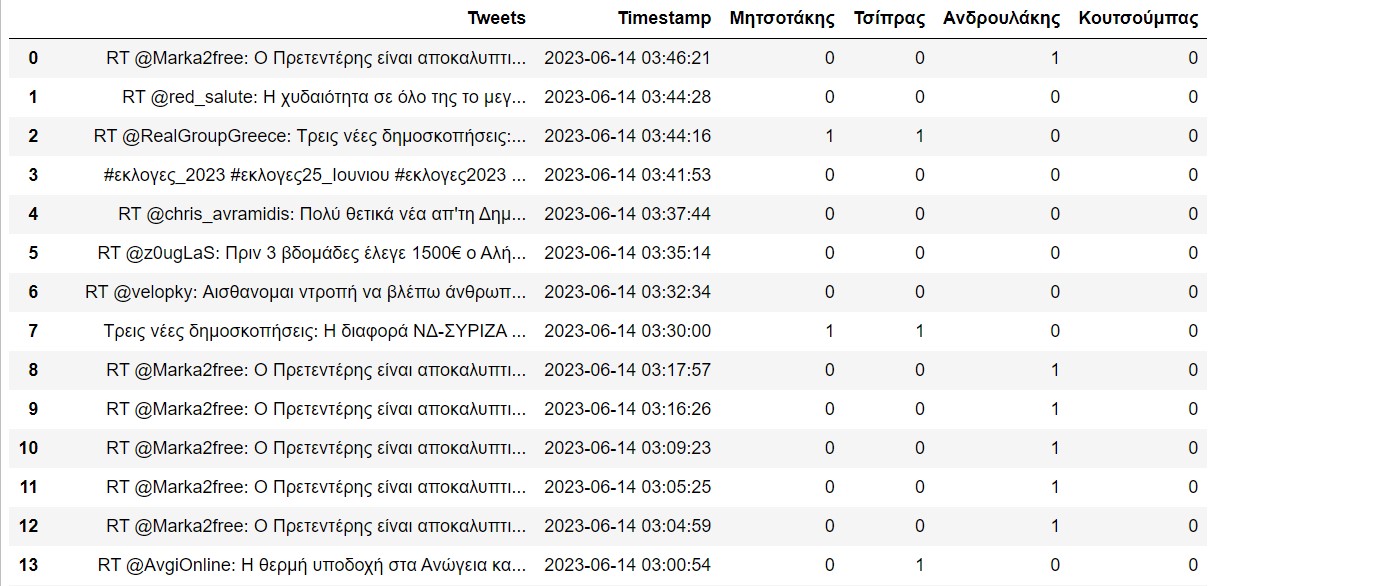
"Ανδρουλάκης" και "Κουτσούμπας". Οι στήλες αυτές χρησιμοποιούνται για να σημειώσουν αν ένα tweet περιέχει αναφορές στον αντίστοιχο πολιτικό ή κόμμα.

Η συνάρτηση “ subject(tweet, refs) “ ελέγχει αν ένα tweet περιέχει κάποια από τις αναφορές που περιέχονται στη λίστα `refs`. Εάν βρεθεί κάποια αναφορά στο tweet, η σημαία (`flag`) τίθεται σε 1, διαφορετικά παραμένει 0.

Όπως φαίνεται παρακάτω, εφαρμόζεται η συνάρτηση `subject()` σε κάθε tweet στη στήλη "Tweets" και το αποτέλεσμα αποθηκεύεται στις αντίστοιχες στήλες "Μητσοτάκης", "Τσίπρας", "Ανδρουλάκης" και "Κουτσούμπας".



Τέλος, με την εντολή `df.head(30)` εμφανίζονται τα πρώτα 30 στοιχεία του DataFrame με τις νέες στήλες που προστέθηκαν.



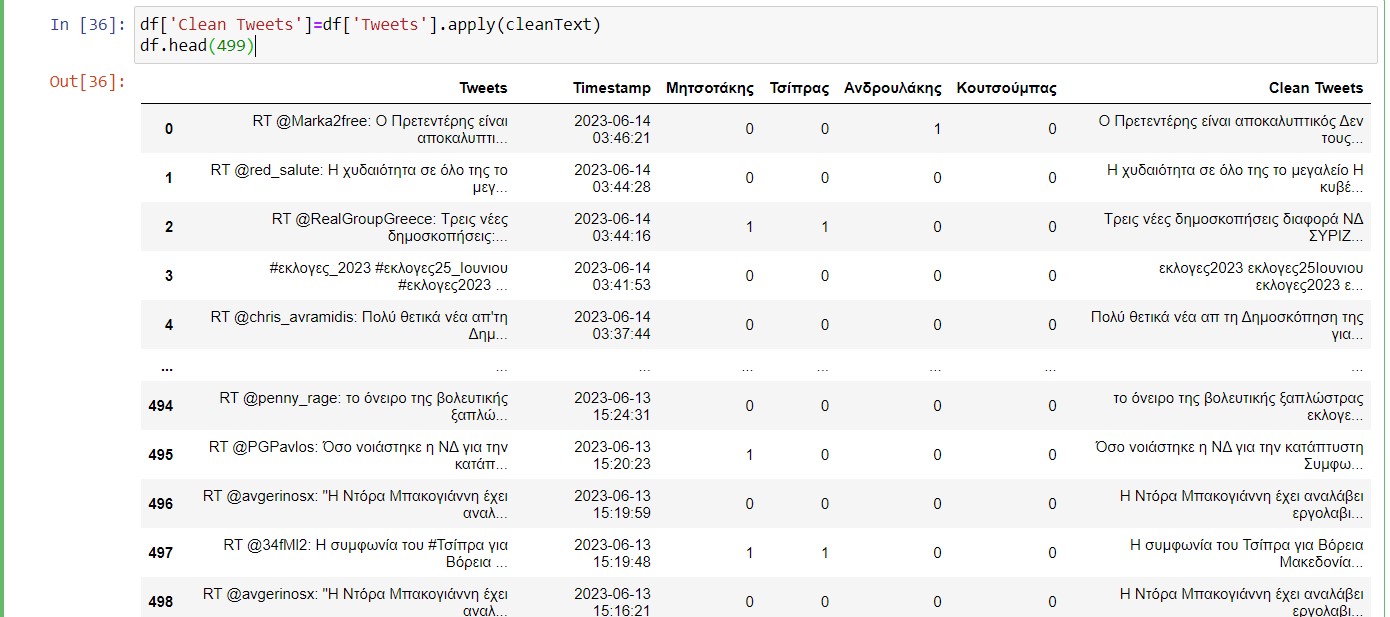
Έπειτα, “καθαρίζουμε” τα tweets με την **βιβλιοθήκη re**. Η συνάρτηση cleanText(txt) χρησιμοποιείται για τον καθαρισμό του κειμένου των

tweets. Εφαρμόζονται πολλές κανονικές εκφράσεις για την αντικατάσταση

συγκεκριμένων προτύπων με κενά. Αντικαθίστανται τα @usernames, αγγλικές λέξεις, τα RT (retweets), όλους τους μη αλφαριθμητικούς χαρακτήρες, συγκεκριμενους χαρακτηρες, emoji οπως και αλλαγές γραμμής με κενό.



Τέλος, δημιουργείται μια νέα στήλη με τα clean tweets με τη μέθοδο **apply**.



Ετσι δημιουργούμε 5 καινούρια Data Frames για τον κάθε πρόεδρο ή κόμμα ξεχωριστά, όπου κάθε tweet που περιλαμβάνει κάποια από τις

αναφορές που ορίσαμε παραπάνω για τον πρόεδρο ή το κόμμα ισούται

με 1 για το κάθε πρόεδρο του κόμματος ξεχωριστά. Δημιουργειται και ενα ολικο Data Frame για να δουμε αν αναφερονται ολοι σε ενα tweet. Με τη χρηση & δηλωνουμε πως θελουμε απο ολους να αναφέρονται.

* ND\_df=df[df['Μητσοτάκης']==1]
* SYR\_df=df[df['Τσίπρας']==1]
* PASOK\_df=df[df['Ανδρουλάκης']==1]
* KKE\_df=df[df['Κουτσούμπας']==1]
* ALL df=df[(df['Τσίπρας']==1) & (df['Μητσοτάκης']==1) &

(df['Μητσοτάκης']==1) & (df['Ανδρουλάκης']==1) &

(df['Κουτσούμπας']==1)]

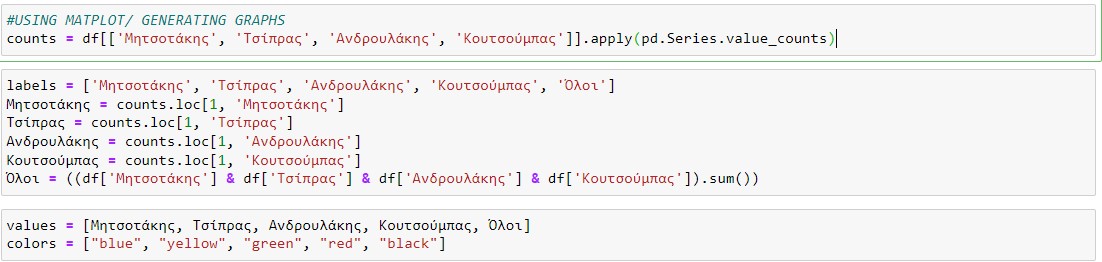
Επίσης βλέπουμε σε πόσα tweets γίνεται η αναφορά βλέποντας το μεγεθος του καθε Data Frame με τη χρηση της len().

* Για τη ΝΔ έχουμε 83 tweets
* Για το ΣΥΡΙΖΑ έχουμε 80 tweets
* Για το ΠΑΣΟΚ έχουμε 44 tweets
* Για το ΚΚΕ έχουμε 23 tweets
* Για ολους έχουμε 7 tweets



Μπορείτε να δείτε τους πίνακες με τα tweets στο αρχείο με τον κώδικα.

Τέλος, γίνεται η **οπτικοποίηση** με τη χρήση της βιβλιοθήκης matplot.



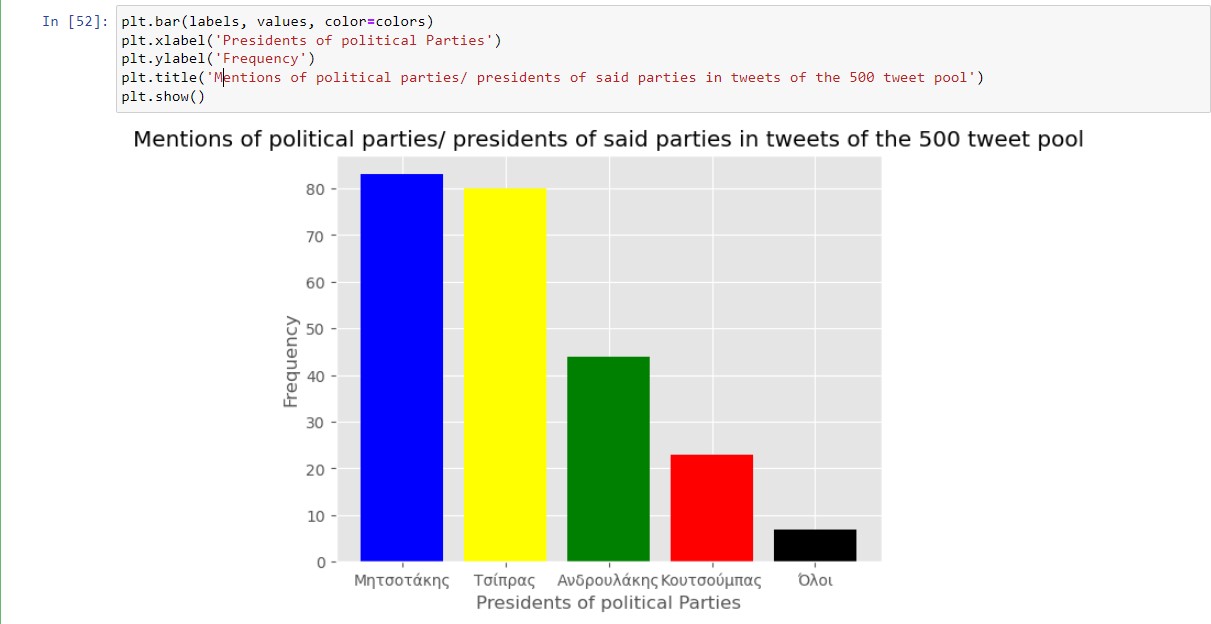
Γίνεται υπολογισμός του αριθμού των tweets που αναφέρονται σε κάθε πολιτικό, καθώς και τον αριθμό των tweets που αναφέρονται σε όλους τους πολιτικούς με βάση τα δεδομένα του DataFrame "df".

Στη συνέχεια, δημιουργούνται λίστες με τις ετικέτες των πολιτικών και τις αντίστοιχες τιμές για τον αριθμό των tweets που αναφέρονται σε κάθε πολιτικό.

Τέλος, ορίζονται τα χρώματα για κάθε πολιτικό, με τη χρήση της μεταβλητής "**colors**".

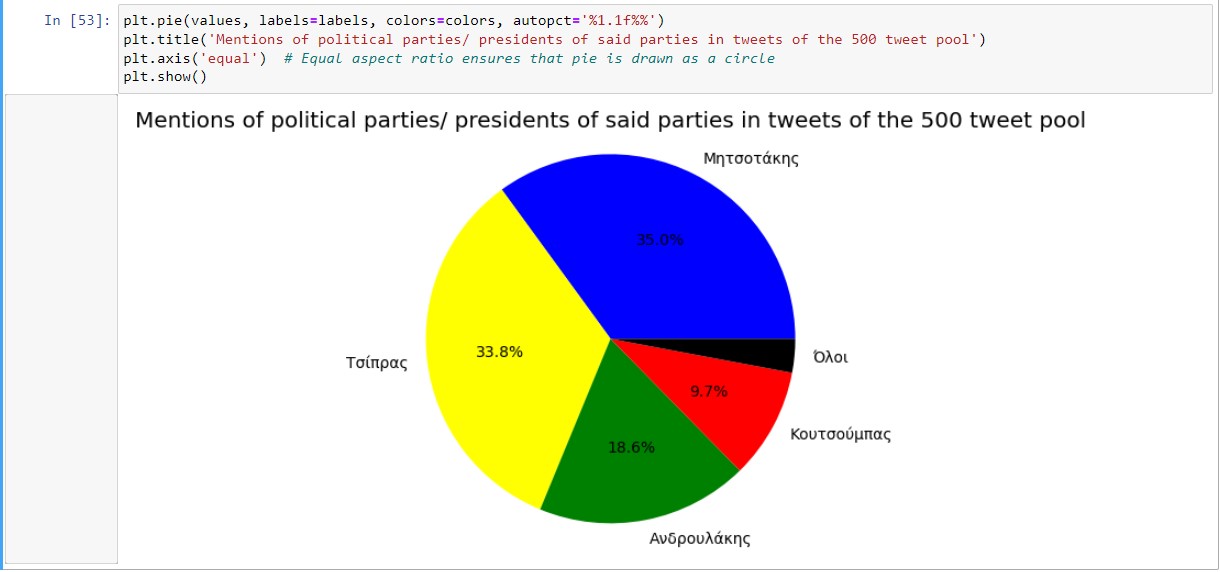
Παράγουμε λοιπόν 3 graphs με τίτλο “ Mentions of political parties/ presidents of said parties in tweets of the 500 tweet pool ”.

1. Πρωτο γράφημα με μπάρες βάζοντας στον άξονα **x** τους προέδρους των κομμάτων και στον **y** τη συχνότητα αναφοράς αυτών στα tweets.

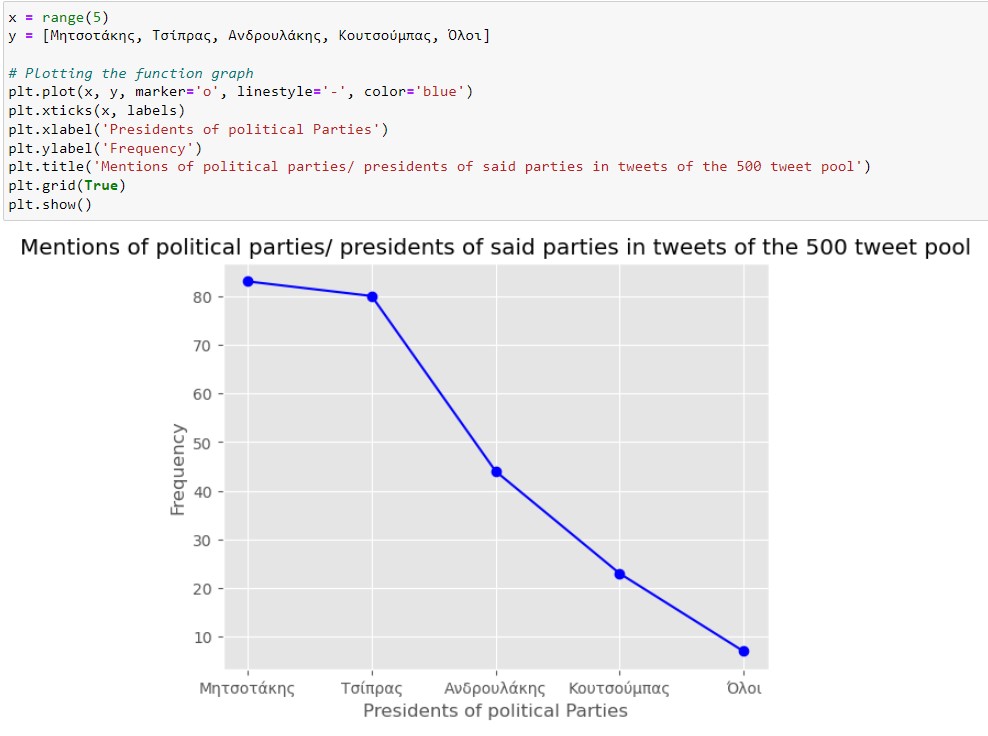


1. Δεύτερη οπτικοποίηση με **pie chart** μπορούμε να δούμε τις ποσοστιαίες αναφορές για τους πολιτικούς. (autopct='%1.1f%%': Ο τρόπος μορφοποίησης των ποσοστών που εμφανίζονται στο

διάγραμμα πίτας.), [axis('equal'): Εξασφαλίζει ότι το διάγραμμα πίτας θα είναι σχεδιασμένο ως κύκλος.].



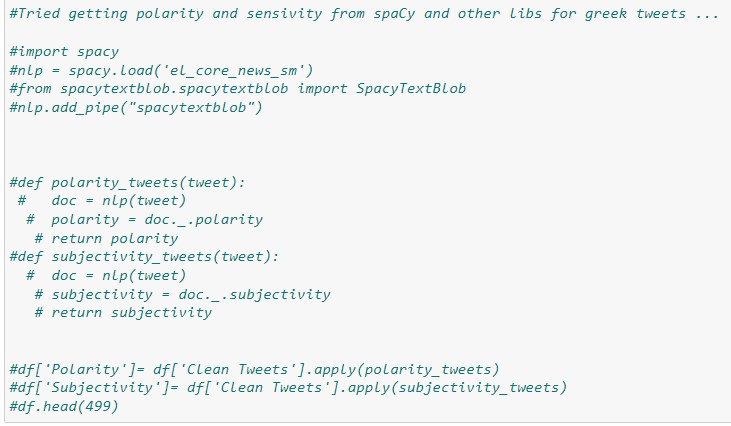
1. Το τελευταίο είναι ένα γράφημα γραμμής. Έχουμε την οπτικοποίηση της πτώσης των αποτελεσμάτων των αναφορών καθώς βλέπουμε την κάθοδο του γράφου από τη ΝΔ σε ολες τις αναφορές. Ίδιους άξονες **x**, **y** οπως στο 1.



**Σημειωση:**

Έγινε προσπάθεια για μετάφραση των ελληνικών clean tweets σε αγγλικων με τη χρήση της βιβλιοθήκης google translate (αποθηκεύοντας τη λίστα tweets σε ενα csv αρχειο). Έπειτα με τη χρήση της **spacy Τext Βlob** θα μπορούσαμε να κάνουμε ένα **sentiment analysis** των tweets όμως η μετάφραση των 500 tweet δεν θα ήταν ακριβής άρα θα είχαμε ανακριβή αποτελέσματα. Ο κώδικας όμως υπάρχει στο αρχείο python σαν σχόλια για το sentiment analysis.

Το προγραμμα επισης δεν τρεχει αυτη τη στιγμη λογω καποιου θεματος του tweepy. Αξιζει να σημειωθει πως σε αλλα runs του προγραμματος τα αποτελεσματα των αναφορων στα tweets και των αποτελεσματων των οπτικοποιησεων ηταν παρομοια.



**ΣΥΜΠΕΡΑΣΜΑΤΑ**

Συμπερασματικά, η εργασία είχε ως στόχο να αναλύσει tweets με το ελληνικό hashtag #εκλογες2023, προκειμένου να προκύψουν

συμπεράσματα σχετικά με τη δημοτικότητα των προέδρων των

Ελληνικών πολιτικών κομμάτων. Με προσεκτική ανάλυση δεδομένων, διαπιστώθηκε ότι οι πρόεδροι με τις περισσότερες αναφορές σε αυτά τα tweets ήταν με φθίνουσα σειρά, ο Κυριάκος Μητσοτάκης, ο Αλέξης Τσίπρας, ο Νίκος Ανδρουλάκης και ο Δημήτρης Κουτσούμπας.

Μεταξύ των tweets που αναλύθηκαν, ο Κυριάκος Μητσοτάκης αναδείχθηκε ως ο πρόεδρος του κόμματος με τις περισσότερες αναφορές, δηλαδή με 83 αναφορές από τα 500 tweets. Αυτό δείχνει ένα σημαντικό επίπεδο δημόσιας δέσμευσης και συζήτησης γύρω από τον

Μητσοτάκη και το πολιτικό του κόμμα. Ο Τσίπρας, ακολούθησε με 80 αναφορές, τονίζοντας το συνεχές ενδιαφέρον για την πολιτική του παρουσία.

Επιπλέον, τα tweets αποκάλυψαν ότι ο Νίκος Ανδρουλάκης, συγκέντρωσε 44 αναφορές στο σύνολο δεδομένων που αναλύθηκε. Αυτό υποδηλώνει ότι ο ίδιος και το κόμμα του, εξακολουθεί να κατέχει

σημαντική θέση στήν πολιτική σκηνή πριν από τις εκλογές του 2023. Τέλος, ο Κουτσούμπας, πρόεδρος άλλου πολιτικού κόμματος, έλαβε 23 αναφορές, υποδηλώνοντας σχετικά χαμηλότερο επίπεδο προβολής και εμπλοκής στα αναλυόμενα tweets.

Τα ευρήματα αυτού του έργου παρέχουν πολύτιμες γνώσεις για το τρέχον πολιτικό κλίμα και το δημόσιο αίσθημα γύρω από τους προέδρους των Ελλήνων κομμάτων στο πλαίσιο των επερχόμενων εκλογών. Είναι προφανές ότι ο Κυριάκος Μητσοτάκης και ο Τσίπρας, ως αρχηγοί

μεγάλων πολιτικών κομμάτων, έχουν ουσιαστική παρουσία στη διαδικτυακή συζήτηση, υποδηλώνοντας τη συνεχιζόμενη επιρροή και τη σημασία τους στο εκλογικό τοπίο.

Οι πολιτικοί αναλυτές μπορούν να αξιοποιήσουν αυτά τα δεδομένα για να αποκτήσουν μια βαθύτερη κατανόηση της αντίληψης του κοινού για τους ηγέτες των κομμάτων και τον πιθανό αντίκτυπό τους στα εκλογικά αποτελέσματα. Οι υπεύθυνοι στρατηγικής εκλογών μπορούν να

χρησιμοποιήσουν αυτά τα ευρήματα για να βελτιώσουν τις στρατηγικές της εκστρατείας τους και να προσαρμόσουν τα μηνύματά τους ώστε να

έχουν απήχηση στους ψηφοφόρους. Οι οργανισμοί μέσων ενημέρωσης μπορούν να χρησιμοποιήσουν τα δεδομένα για να παρέχουν ακριβή και βασισμένη σε δεδομένα κάλυψη των εκλογών, ενισχύοντας την

ευαισθητοποίηση και τον λόγο του κοινού. Τέλος, το ευρύ κοινό μπορεί να επωφεληθεί από μια πιο ενημερωμένη κατανόηση του πολιτικού κλίματος, δίνοντάς του τη δυνατότητα να λαμβάνει μορφωμένες αποφάσεις και να συμμετέχει ενεργά στη δημοκρατική διαδικασία.

**ΤΕΛΟΣ**