Περιβάλλοντα Επίλυσης Προβλημάτων για Εφαρμογές στην Επιστήμη Δεδομένων ECE525

Report

Πέμπτη 17 Νοεμβρίου 2022

Χειμερινό Εξάμηνο

Twitter Sentiment Analysis

Χρήστος Αρσενίου 2730

Επιβλέπων καθηγητής: Ηλίας Χούστης

1 Εισαγωγή

Η εργασία, επικεντρώνεται στην ανάκτηση tweets από το twitter που θα αφορούν τον Covid19, μέσω του tweepy API της γλώσσας προγραμματισμού Python. Στην συνέχεια, με το framework nltk, γίνεται η sentiment analysis και το tokenization των δεδομένων.

2 Σύνδεση με το Developer Portal του Twitter

Αρχικά, φτιάξαμε ένα twitter developer account, κάνοντας upgrade σε Elevated Access. Έπειτα, φτίαξαμε ένα project στο dashboard του twitter, και εντοπίσαμε τους consumer και access κλειδάριθμους, επιτυγχάνοντας την σύνδεση με το πρόγραμμά μας. Πλέον, είμαστε στην θέση να αναζητήσουμε tweets με βάση το επιθυμητό hashtag.

3 Scrapping Data από το Twitter

Η ανάκτηση των tweets, έγινε με την συνάρτηση Cursor του tweepy framework, επισημαίνονται το hashtag #Covid19, και επιλέγοντας την αγγλική γλώσσα. Επίσης, δηλώνουμε ως True την παράμετρο wait_on_rate_limit, ώστε να πάρουμε τον επιθυμό αριθμό tweets, χωρίς να έχουμε προβλήματα με το όριο.

4 Sentiment Analysis

Για την υλοποίηση του Sentiment Analysis έγινε χρήση του pretrained sentiment analyzer, VADER. Ένα πλεονέκτημα αυτού του analyzer είναι ότι παράγει αποτελέσματα γρηγορότερα σε σύγκριση με άλλους analyzers. Παράλληλα, ο VADER παράγει πιο ακριβή αποτελέσματα αν χρησιμοποιείται για μικρές προτάσεις, αργκό γλώσσα καθώς και συντομογραφίες. Για αυτό το λόγο, είναι ιδανικός για την ανάλυση δεδομένων από το tweeter.
Όσον αφορά την υλοποίηση που κάναμε, αρχικά δημιουργείται ένα instance του συγκεκριμένου analyzer και στη συνέχεια αναλύεται το κάθε tweet. Αυτό επιτυγχάνεται με τη χρήση της συνάρτησης polarity_scores. Η συνάρτηση αυτή παράγει ένα λεξικό που περιέχει διαφορετικά scores, αρνητικό, θετικό ουδέτερο.
Όλη αυτή η διαδικασία γίνεται σε μια επανάληψη για κάθε σχόλιο και στη συνέχεια ανάλογα με το ποιο score υπερισχύει το κάθε σχόλιο κατανέμεται

σε λίστα. Έχουμε τη 'θετική' λίστα, στην οποία κατανέμονται τα σχόλια που έχουν θετικό score μεγαλύτερο του αρνητικού, την 'αρνητική' λίστα, στην οποία κατανέμονται τα σχόλια που έχουν αρνητικό score μεγαλύτερο του θετικού και τέλος την 'ουδέτερη' λίστα, όπου κατανέμονται τα σχόλια που έχουν θετικό score ίσο με το αρνητικό.

5 Web App

Η Web εφαρμογή μας, ζητά το hashtag και τον αριθμό των tweets που θα κάνει scap. Πατώντας το κουμπί Sentiment Analysis, αρχικά αντλούνται τα δεδομένα, και γίνεται η ανάθεση των positive, negative και neutral τιμών για κάθε tweet με το SentimentIntensityAnalyzer, εφόσον εντοπίσουμε το clean text.

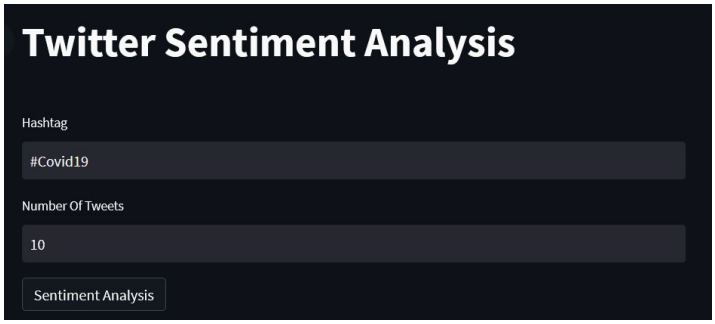

Figure 1: User Input

Απεικόνιζουμε ένα διάγραμμα πίτας, αναγράφοντας το ποσοστό των negative, positive και neutral tweets.

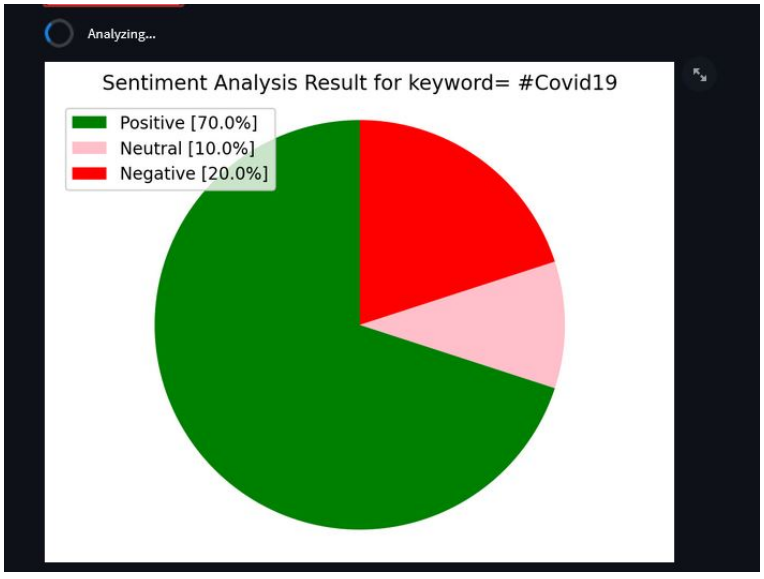

Figure 2: Pie Chart

Τέλος, φτιάχνουμε 4 wordclouds, ένα για όλα τα tweets, και ένα για κάθε υποσύνολο, που περιέχει τα positive ή τα negative ή τα neutral tweets.

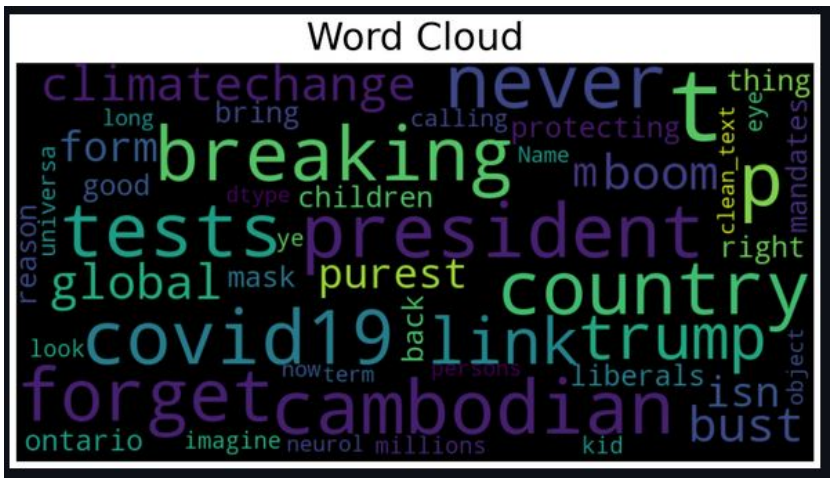

Figure 3: Wordcloud για όλα τα tweets.