Ανάπτυξη Εφαρμογής για τη Συλλογή και την Ανάλυση Δεδομένων του Twitter

Γιατσίδης Γεώργιος, mai:18010
Επιβλέπων: Δρ. Μαργαρίτης Κωνσταντίνος

Περιεχόμενα

1. Εισαγωγή

Σημαντικότητα του Θέματος, Σκοποί - Στόχοι

2. Τεχνολογικό Υπόβαθρο

Big Data, Twitter, Apache Spark, Scala, Apache Kafka, JS, Node.js, Redis, MySQL, HTTP, WebSocket, MVC

3. Σχεδίαση

Καταγραφή Απαιτήσεων, Σχεδίαση Βάσης Δεδομένων, Εφαρμογή Τεχνολογιών

4. Υλοποίηση

Ανάλυση Κώδικα, Συγκριτικά Αποτελέσματα, Παρουσίαση της Εφαρμογής

5. Επίλογος

Συμπεράσματα, Όρια - Περιορισμοι, Μελλοντικές επεκτάσεις

Εισαγωγή

Πρόβλημα - Σημαντικότητα του Θέματος:

- ⇨ Σημαντική αύξηση στη χρήση και στον όγκο δεδομένων των κοινωνικών δικτύων.
- ⇨ Αύξηση του ενδιαφέροντος της συλλογής και ανάλυσης δεδομένων που προέρχονται από τα κοινωνικά δίκτυα.
- ⇨ Ο σκοπός αυτής της ανάλυσης είναι:
  - Λήψη σωστών αποφάσεων
  - Παροχή αξιόπιστων πληροφοριών
  - Εξαγωγή συμπερασμάτων για την άποψη της κοινής γνώμης

Σκοποί - Στόχοι:

- ⇨ Υλοποίηση ενός συστήματος που χρησιμοποιεί τεχνολογίες αιχμής και μεγάλων δεδομένων.
  - Συλλογή και επεξεργασία των δεδομένων του Twitter σε πραγματικό χρόνο, χρησιμοποιώντας τεχνικές κατανεμημένης επεξεργασίας δεδομένων.
- ⇨ Ανάλυση, Κατηγοριοποίηση και απεικόνιση των δεδομένων που συλλέγονται μέσω του κοινωνικού δικτύου Twitter.
- ⇨ Αναπαράσταση σε διαδικτυακό γραφικό περιβάλλον φιλικό προς το χρήστη.

Big Data

- Ανάλυση και επεξεργασία τεράστιων συνόλων δεδομένων.
- Οι **επιχειρήσεις** και οι **οργανισμοί** επιτυγχάνουν καλύτερα οικονομικά αποτελέσματα ενσωματώνοντας τα μεγάλα δεδομένα στη διαδικασία λήψης αποφάσεων.
- Χαρακτηριστικά (3V):
  - **Volume**: αναφέρεται στον όγκο των δεδομένων που πρέπει να συλλεχθούν και να υποβληθούν σε επεξεργασία
  - **Variety**: αναφέρεται στους διαφορετικούς τύπους και πηγές δεδομένων.
  - **Velocity**: αναφέρονται στην ταχύτητα με την οποία συσσωρεύονται τα μεγάλα δεδομένα
- Νέα Χαρακτηριστικά (5V)
  - **Veracity**: αναφέρεται στην ακρίβεια των δεδομένων
  - **Value**:αναφέρεται στην ικανότητα μετατροπής ενός τεράστιου συνόλου δεδομένων σε επιχειρηματική ευφυΐα, μέσα από διαδικασίες επεξεργασίας και ανάλυσης, καθώς η συλλογή πολλών δεδομένων δεν δίνει αξία.

Τεχνολογίες

Τεχνολογίες για τη συλλογή και την ανάλυση δεδομένων

- **Scala, Apache Spark**
- **Apache Kafka**
- **Apache Avro**
  - σύστημα δυαδικής σειριοποίησης που συμβάλλει στην ανταλλαγή δεδομένων μεταξύ γλωσσών προγραμματισμού.
- **Redis**
  - Μετάδοση δεδομένων σε πραγματικό χρόνο χρησιμοποιώντας το μοντέλο Δημοσίευσης/Συνδρομής
  - In-memory βάση δεδομένων
- **MySQL**
  - Σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων

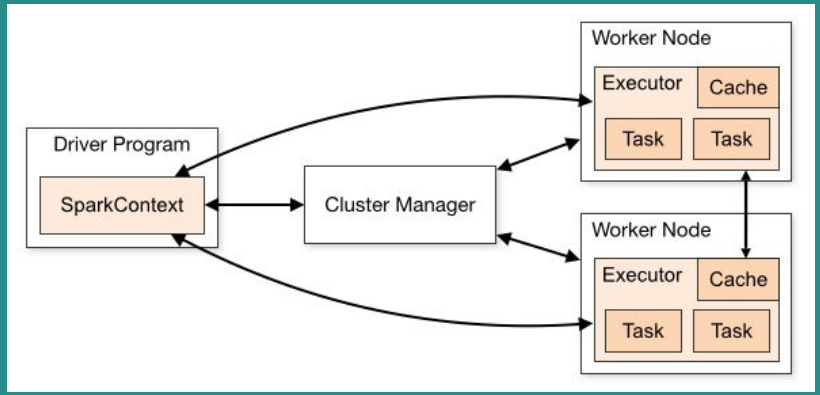
Τεχνολογίες Ιστού

- **HTML**
  - γλώσσα σήμανσης (Markup Language).
  - Αποτελείται από ένα σύνολο εικέτων.
- **CSS** (Cascading Style Sheets)
  - γλώσσα μορφοποίησης
  - καθορίζει τη διαρρύθμιση των HTML στοιχείων.
- **JavaScript / Node.JS**
  - Socket.IO (WebSocket), Express.JS (HTTP)

Apache Spark

Το Spark είναι προγραμματιστικό **πλαίσιο** επεξεργασίας δεδομένων που χρησιμοποιείται για αποθήκευση και ανάλυση μεγάλων δεδομένων:

1. Υψηλού Επιπέδου API (**Scala**, Java, Python)
2. Εκτελείται σε τοπική ή κατανεμημένη λειτουργία
3. Οικοσύστημα με μια γκάμα βιβλιοθηκών για ένα ευρύ φάσμα φόρτου εργασιών
   - ⇨ Spark Core **RDD**
   - ⇨ Spark Streaming
   - ⇨ Spark MLlib (Machine Learning)
     - Classifiers
   - ⇨ Spark SQL



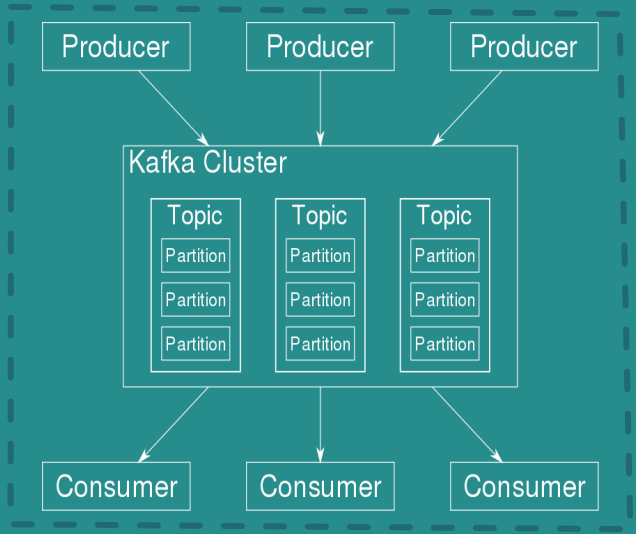
Apache Kafka

Το Kafka είναι μία πλατφόρμα διαχείρισης ροών δεδομένων.

- Μετάδοση δεδομένων σε πραγματικό χρόνο χρησιμοποιώντας το μοντέλο **Δημοσίευσης/Συνδρομής.**
- **Γρήγορο** και **επεκτάσιμο** λόγω της κατανεμημένης φύσης του.

Αρχιτεκτονική:

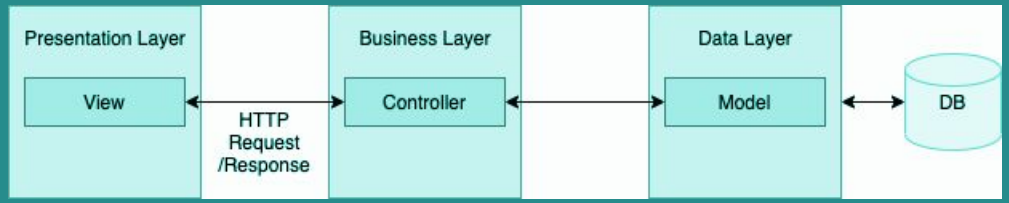
- **Topic** (Θέμα)
  - Μία ομάδα μηνυμάτων που χαρακτηρίζεται από ένα μοναδικό όνομα.
- **Partition** (Διαμέρισμα)
  - Τα δεδομένα κάθε θέματος μπορούν να χωριστούν σε διαφορετικά διαμερίσματα (partitions).
- **Producer** (Παραγωγός)
  - δημοσιεύει μηνύματα σε ένα ή περισσότερα θέματα.
- **Consumer** (Καταναλωτής)
  - Εγγράφεται σε ένα ή περισσότερα θέματα. Είναι υπεύθυνος για την ανάγνωση των μηνυμάτων.
- **Kafka Cluster**
  - Το Kafka μπορεί να εκτελεστεί σε cluster υπολογιστών. Τα διαμερίσματα όλων των θεμάτων κατανέμονται στους κόμβους του cluster.



Node.js

- Εκτελεί JavaScript στην πλευρά του **διακομιστή**.
- Μπορεί να χειριστεί μεγάλο αριθμό ταυτόχρονων συνδέσεων
- Δεν απαιτείται η γνώση **2** γλωσσών για την υλοποίηση της πλευράς πελάτη και της πλευράς διακομιστή
- Διαχείριση πακέτων μέσω του **NPM (Node Package Manager).**.
  - Παρέχει διεπαφή γραμμής εντολών που διευκολύνει την εγκατάσταση και την αφαίρεση πακέτων.
- Εύκολη ανάπτυξη **HTTP** διακομιστή (Express.js) και **WebSocket** διακομιστή (Socket.IO).

MVC / Αρχιτεκτονική 3-tier

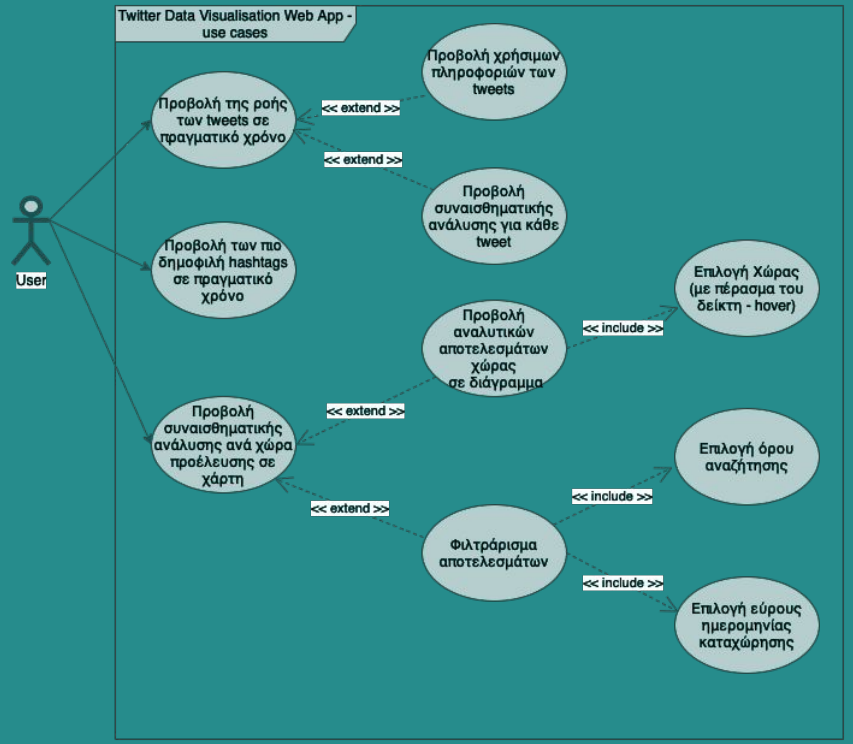


- Αρχιτεκτονικό Πρότυπο Λογισμικού

- **Model** (Μοντέλο)
  - Αντιπροσωπεύει τα δεδομένα της βάσης
- **View** (Προβολή)
  - Εμφάνιση των δεδομένων στον χρήστη
- **Controller** (Ελεγκτής)
  - Χειρισμός Δεδομένων, Μεσολαβητής μεταξύ Μοντέλου και Προβολής

Καταγραφή Απαιτήσεων

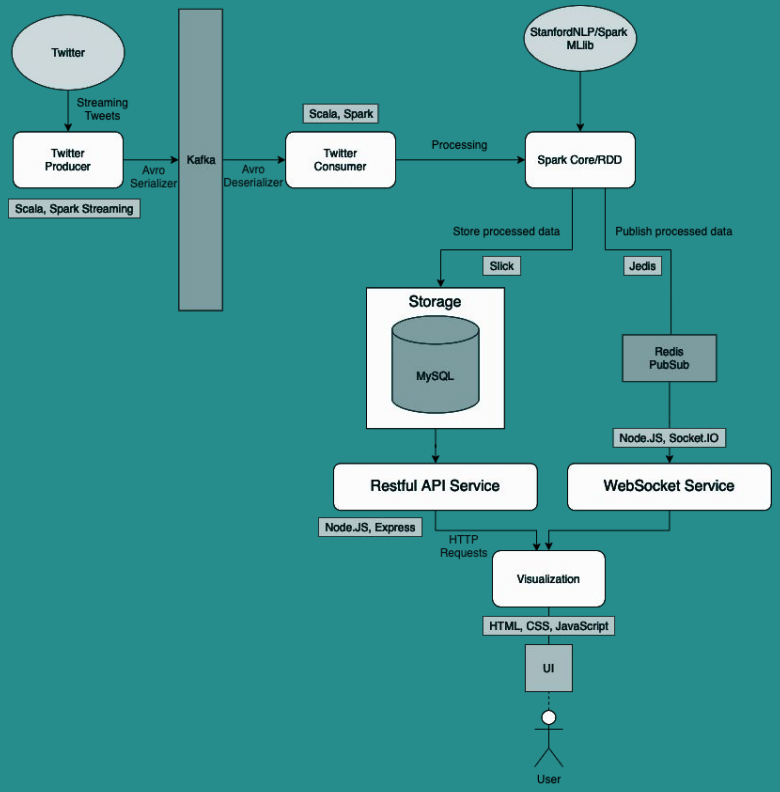
- Εμφάνιση των συλλεχθέντων tweets σε πραγματικό χρόνο συνοδευόμενα με πληροφορίες που σχετίζονται με αυτά.
- Εμφάνιση των **δέκα** πιο δημοφιλή θεμάτων συζήτησης **(hashtags)** σε πραγματικό χρόνο.
- Προβολή συναισθηματικής ανάλυσης ανά χώρα προέλευσης σε διαδραστικό χάρτη.
  - Προβολή επικρατέστερου συναισθήματος ανά **χώρα** προέλευσης.
  - Προβολή αναλυτικής συναισθηματικής ανάλυσης όλων των tweets ανά χώρα προέλευσης σε διάγραμμα πίτας.
  - Υποστήριξη σύνθετης αναζήτησης με βάση τον όρο αναζήτησης και το εύρος της ημερομηνίας καταχώρησης.



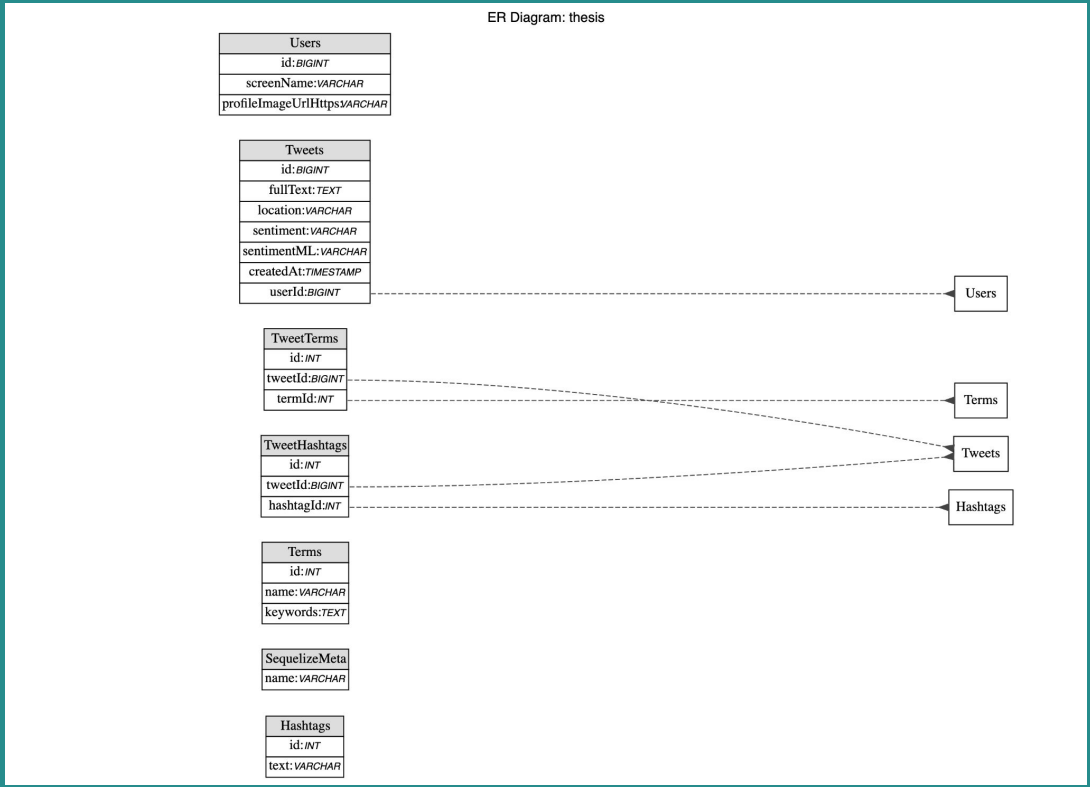
Αρχιτεκτονική

Scala, Spark, Spark MLlib, Stanford NLP	Node.js, Express.js	Node.js, Express, Socket.IO, Redis	HTML, CSS, JavaScript, Google Maps API
Kafka, Avro, MySQL, Redis	MySQL - Sequelize	Redis	Third-party JS libraries

Αρχιτεκτονική



Βάση Δεδομένων



Συγκριτική Μελέτη Stanford NLP - Spark MLlib Naive Bayes

- ⇨ Spark MLlib
  - Naive Bayes Classifier
  - Εκπαίδευση μοντέλου
  - Σύνολο Δεδομένων Sentiment140
- ⇨ Stanford Core NLP
  - Πρόβλεψη Συναισθήματος

Πρόβλεψη Stanford - MLlib

Αρχικό κείμενο του tweet	Stanford	MLlib
Probably the best buy this month	POSITIVE	POSITIVE
Mr. Wonderful is smart, be like Mr. Wonderful and diversify your portfolio with some #bitcoin	VERY POSITIVE	POSITIVE
wow very good, good luck on the project	VERY POSITIVE	POSITIVE
so the 8 million people who die from air pollution each year due to fossil fuels, and the environmental damage from oil spills were not enough to catalyze a renewable energy revolution. but #bitcoin will raise	VERY NEGATIVE	NEGATIVE
The current price of #eth is $1,590.90, an increase of +1.2%, a 24 hr volume of $29.41B, and a market cap of $182.88B	NEGATIVE	NEGATIVE
Bro that's a life changing amount of money. Thanks for doing something like this and best of luck to whoever wins!	POSITIVE	POSITIVE
I wanted to report that I just got the BTC! I am happy right now!	POSITIVE	NEGATIVE
Just got those BTC, fast and easy. Nice!	POSITIVE	NEGATIVE
Your ETH is gonna grow to 30k each in 5 years, that's what's gonna happen	NEGATIVE	NEGATIVE
I think if oil etc is traded in BTC vs USD it will boom. I think that happens before Central Bank holds any.	NEGATIVE	NEGATIVE
Bitcoin has been unstoppable. But it's just he beginning of a whole new industry	NEUTRAL	POSITIVE

Ακρίβεια Συναισθ. Ανάλυσης

61%

Stanford NLP

81%

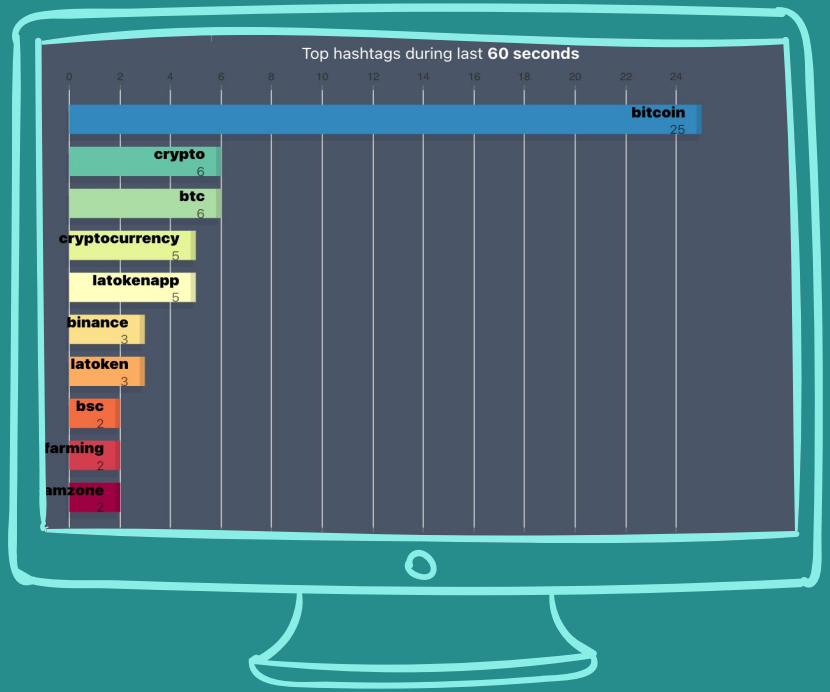
Spark MLlib

Ζωντανή Ροή Twitter, Συναισθ. Ανάλυση

- ⇨ Προβολή ζωντανής ροής του Twitter
- ⇨ Προβολή αποτελεσμάτων συναισθηματικής ανάλυσης σε πραγματικό χρόνο
- ⇨ Προβολή πληροφοριών χρήστη

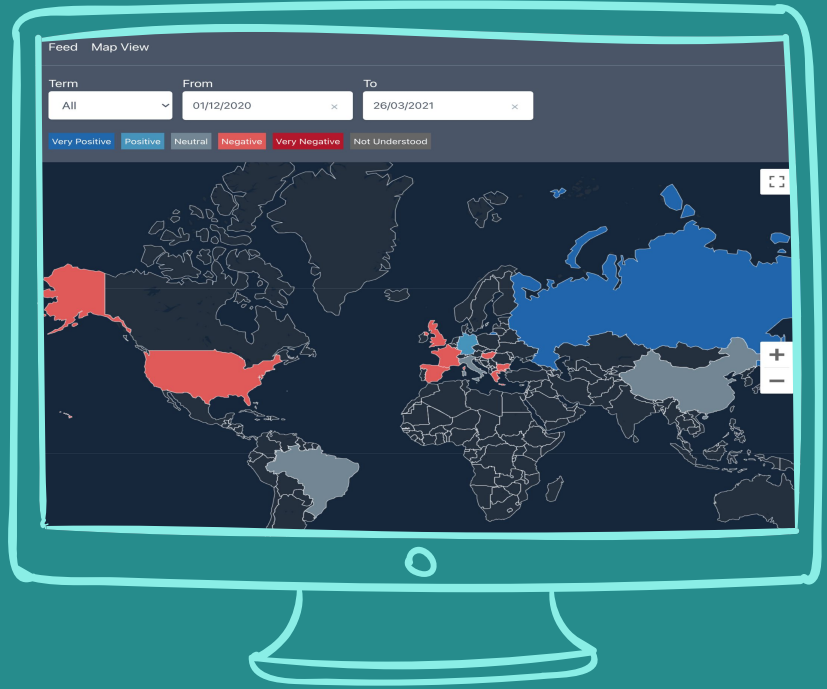
Προβολή Κορυφαίων Hashtag

- Κορυφαία 10 hashtag
- Ανανέωση ανά 60 δευτερόλεπτα



Προβολή Χάρτη

- Προβολή Συναισθηματικής Ανάλυσης σε διαδραστικό χάρτη
- Ομαδοποίηση ανά χώρα
- Υποστήριξη σύνθετης αναζήτησης
  - Όρος αναζήτησης μέσω λίστας επιλογών
  - Επιλογή εύρους ημερομηνίας (από, μέχρι)



Προβολή Χάρτη

Αναλυτικά
αποτελέσματα χώρας

Επίλογος

Συμπεράσματα

- ⇨ Σημαντικότητα της ανάλυσης δεδομένων που προέρχονται από τα κοινωνικά δίκτυα.
- ⇨ Υλοποίηση μιας εφαρμογής που βασίζεται σε **σύγχρονες τεχνολογίες ιστού** και **μεγάλων δεδομένων**.
  - Αποθήκευση σε σχεσιακή βάση δεδομένων.
  - Απεικόνιση αποτελεσμάτων σε **πραγματικό χρόνο** με τη βοήθεια διαδραστικού χάρτη και γραφημάτων.
- ⇨ Το Twitter είναι μια πολύ καλή πηγή εξαγωγής δεδομένων.
  - Άντληση δεδομένων σε πραγματικό χρόνο.
  - Απόκτηση γνώσης έγκαιρα (π.χ. Λανσάρισμα ενός προϊόντος)

Όρια - Περιορισμοί

- ⇨ Συναισθηματική Ανάλυση σε tweets που είναι μόνο στην Αγγλική γλώσσα
- ⇨ Αδυναμία Εντόπισης ειρωνικών σχολίων.

Μελλοντικές Επεκτάσεις

- ⇨ Μεταφορά σε απομακρυσμένο περιβάλλον
- ⇨ Προσθήκη νέων δυνατοτήτων
  - Επιλογή θεμάτων αναζήτησης από τον τελικό χρήστη
  - Υλοποίηση διαχειριστικού συστήματος

Thank you!

Do you have any questions?