ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΕΙΣ. ΠΑΡΟΥΣΙΑΣΗΣ: ΤΑΚΤΙΚΟΥ ΕΛΕΝΗ
ΕΠ. ΚΑΘΗΓΗΤΗΣ: ΚΩΝΣΤΑΝΤΙΝΟΣ ΨΑΝΝΗΣ

Η ΕΠΑΝΑΣΤΑΣΗ ΤΩΝ BIG DATA:

Ανιχνεύουμε συναισθήματα μέσω κοινωνικών δικτύων!

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ ΦΕΒΡΟΥΑΡΙΟΣ 2021

1. Εισαγωγή

Τι είναι τα Big Data;

*Τα **Big Data** είναι μεγάλοι **όγκοι, σύνθετων και μεταβλητών δεδομένων, υψηλής ταχύτητας**.*

Απαιτούν προηγμένες τεχνικές και τεχνολογίες για τη **σύλληψη**, **αποθήκευση**, **διανομή**, **διαχείριση** και **ανάλυση τους**.

..Ανταπόκριση στις επιθυμίες του κοινού πριν καν αυτό συνειδητοποιήσει τις ανάγκες του..

Το twitter είναι μια από τις μεγαλύτερες πηγές big data εξαιτίας των πολλών tweets καθημερινά. Είναι πλούσια πηγή για την ανάλυση της κοινής γνώμης, πολύ δύσκολη στο να επεξεργαστούν τα δεδομένα με παραδοσιακούς τρόπους γιατί παράγονται σε πραγματικό χρόνο.

Οι κοινωνικές σχέσεις στα social media επηρεάζουν τις συμπεριφορές σε δημόσιο επίπεδο και συμβάλλουν στη δημιουργία γνώσης.
Η εξαγωγή πληροφοριών από αυτά τα δεδομένα έχει γίνει μια πολύ διευρυμένη πολυεπιστημονική περιοχή που απαιτεί τη συνεργασία πολλών σύγχρονων εργαλείων.

Τα big data αναλύονται τόσο πολύ, που δημιουργούνται αμφιβολίες σχετικά με το κατά πόσο χρησιμοποιούνται με ηθικούς τρόπους.

Θα δούμε...

- Πως τα big data επηρεάζουν τα κοινωνικά δίκτυα;
- Μεθόδοι ανάλυσης τους και σύγκριση τους
- Βασικά εργαλεία
- Apache Spark (βασικά στοιχεία, εγκατάσταση προγράμματος)
- Παραδείγματα ανάλυσης δεδομένων σε πραγματικό χρόνο με το Spark Streaming.

2. Οι τρόποι που τα big data επηρεάζουν τον κόσμο μας.

The Rise of the IoT

Big Data, επιχειρήσεις και διαδίκτυο των πραγμάτων..

Η εμφάνιση νέων τεχνολογιών και του Διαδικτύου των πραγμάτων (IoT) οδήγησαν σε μια εκρηκτική αύξηση δεδομένων.

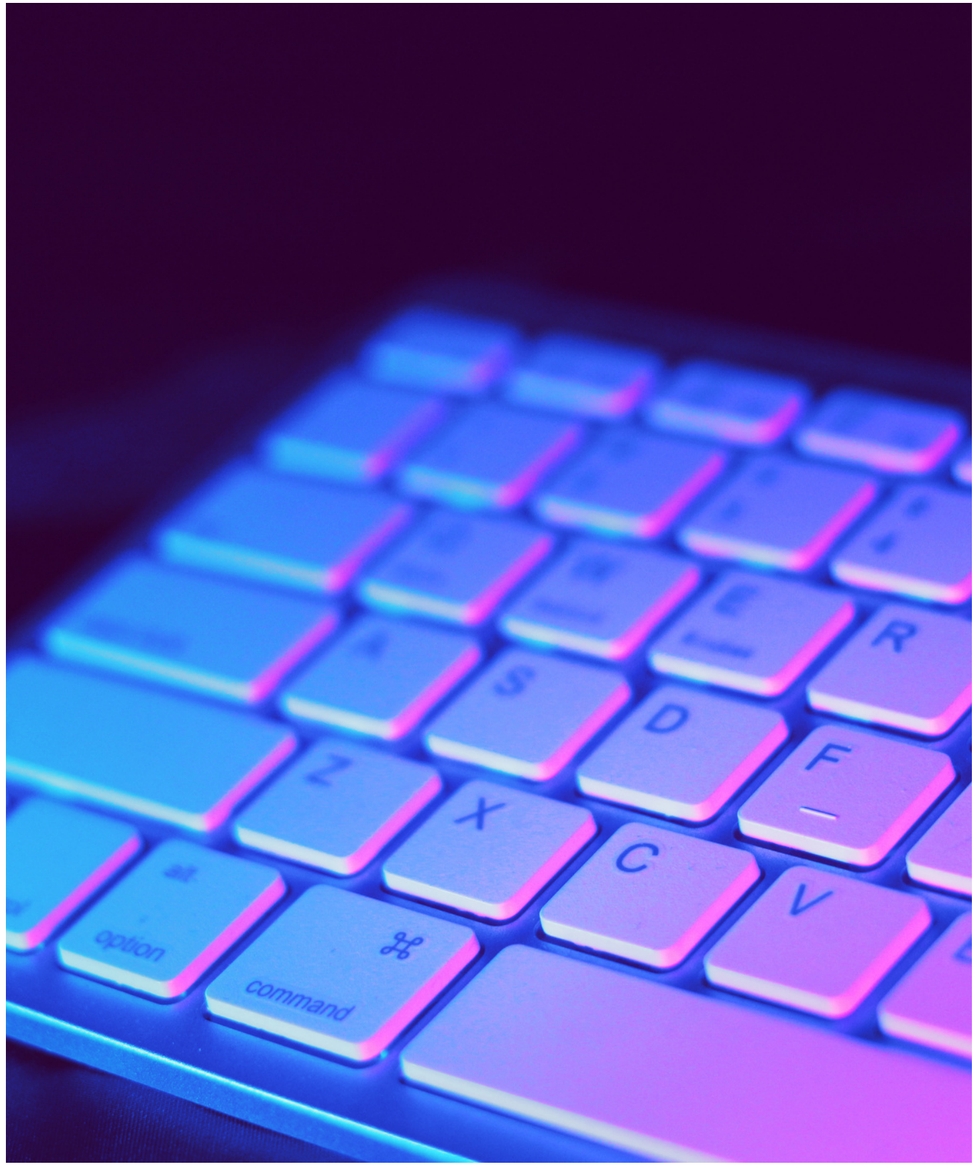
Ένας τεράστιος αριθμός δικτυωμένων συσκευών σε όλο τον κόσμο συλλέγουν διαφορετικούς τύπους δεδομένων (περιβαλλοντικά, γεωγραφικά, λογιστικά κ.λ.π.).

Στη συνέχεια, οι συσκευές IoT μεταδίδουν τα συλλεγόμενα δεδομένα ώστε να μπορούν να αποθηκευτούν, να υποβληθούν σε επεξεργασία και να αναλυθούν.

Στον επιχειρηματικό κόσμο, η πρόσβαση σε άπειρα δεδομένα, χάρις στο Διαδίκτυο των πραγμάτων είναι μεγάλη πρόκληση και ευκαιρία!

Η ακρίβεια στην επεξεργασία των Big Data μπορεί να οδηγήσει σε :

(1)μείωση του κόστους
(2) μείωση του χρόνου
(3) ανάπτυξη νέων προϊόντων και βελτιστοποίηση των προσφορών τους και
(4) λήψη πιο έξυπνων αποφάσεων.



4 βασικοί τρόποι με τους οποίους τα big data επηρεάζουν το Social Media Marketing:

"Τα Social Media είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων οργανωμένα σε μια πλατφόρμα στο διαδίκτυο. Επιτρέπουν την διεπαφή ανάμεσα στους χρήστες τους π.χ. με σχόλια, φωτογραφίες, μηνύματα κ.α"

Οι ιστότοποι αυτοί αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνούν και να αναπτύσσουν επαφές.

Εξατομίκευση

Τα big data επιτρέπουν την εξατομίκευση, και δίνουν τη δυνατότητα στα brands να προσεγγίζουν τους πελάτες τους με προσωποποιημένο τρόπο.

Αυτό συμβάλει στο να δημιουργήσουν προσαρμοσμένη επικοινωνία για να βελτιώσουν την εμπιστοσύνη προς το πρόσωπό τους.

Οι διαφημίσεις θα στοχεύουν με βάση τις αναρτήσεις κοινωνικών μέσων των χρηστών. Έτσι η μετατροπή ενός ακολούθου σε πελάτη είναι πιο εύκολη από ποτέ.

Λήψη αποφάσεων

Τα big data επιτρέπουν στις επιχειρήσεις να:

- εντοπίζουν τις τάσεις των κοινωνικών μέσων
- αποκτούν πληροφορίες που μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων.
- παρακολουθούν δημογραφικά στοιχεία για να αποφασίζουν σε ποιά πλατφόρμα κοινωνικών μέσων να στοχεύσουν.
- κατανοούν εύκολα τα συναισθήματα των καταναλωτών
- αναπτύσσουν στρατηγικές νίκης.
- στοχεύουν στην πρόβλεψη μελλοντικών αναγκών.

Αποτελεσματικότητα εκστρατειών

Τα εργαλεία ανάλυσης δίνουν τη δυνατότητα στις επιχειρήσεις να λαμβάνουν αποφάσεις σχετικά με το πότε πρέπει να σταματήσει η εκστρατεία για να αποφευχθεί απώλεια χρημάτων.
Αποκτώνται ευαίσθητες πληροφορίες όπως:

- οι ώρες αιχμής των πελατών,
- οι προτιμήσεις τους,
- η συμπεριφορά τους

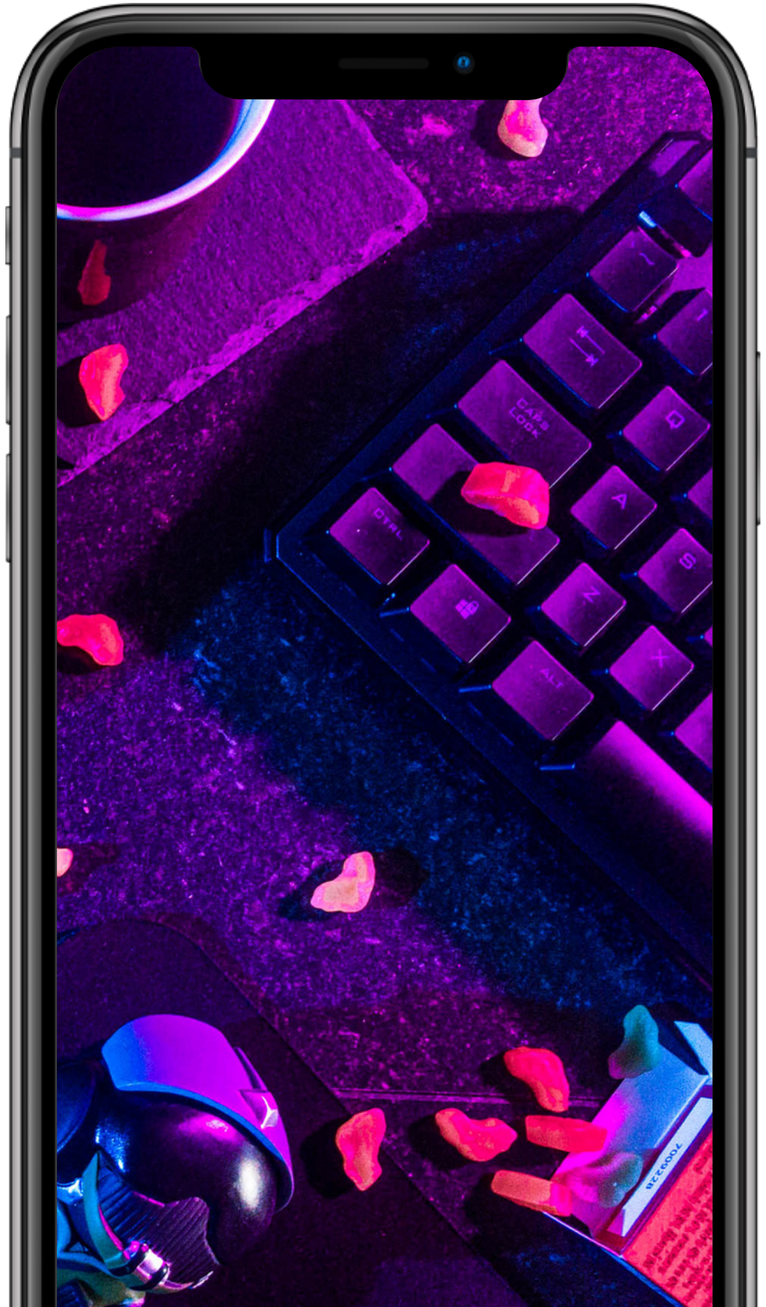
κάτι που οδηγεί σε αυξημένη αποτελεσματικότητα εκστρατείας.

Στατιστικά στοιχεία προϊόντος

"Τι θέλουν οι καταναλωτές, πότε το θέλουν και πως το θέλουν. "

Έτσι δημιουργούνται νέα προιόντα ή βελτιώνονται ήδη υπάρχοντα:

Αναλύονται οι επιλογές των ανθρώπων, τα παράπονά τους, τα προϊόντα που λείπουν απ την αγορά, ελαττώματα σε προϊόντα.



Κανόνες που πρέπει να τηρούνται:

- Τα προσωπικά δεδομένα πρέπει να υποβάλλονται σε επεξεργασία με νόμιμο δίκαιο και διαφανή τρόπο.
- Πρέπει να υπάρχουν συγκεκριμένοι σκοποί για την επεξεργασία των δεδομένων και η εταιρεία / οργανισμός πρέπει να τους αναφέρει.
- Πρέπει να συλλέγονται και να επεξεργάζονται μόνο τα προσωπικά δεδομένα που είναι απαραίτητα για την εκπλήρωση του εκάστοτε σκοπού («ελαχιστοποίηση δεδομένων»).
- Πρέπει να διασφαλίζεται ότι τα προσωπικά δεδομένα είναι ακριβή και ενημερωμένα.
- Δεν μπορούν να χρησιμοποιηθούν περαιτέρω τα προσωπικά δεδομένα για άλλους σκοπούς που δεν είναι συμβατοί με τον αρχικό σκοπό και δεν πρέπει να αποθηκεύονται μετά το πέρας της διαδικασίας
- Πρέπει να χρησιμοποιούνται τεχνικές και οργανωτικές διασφαλίσεις που υπόσχονται ασφάλεια προσωπικών δεδομένων.

..και τελικά η αξία της ανάλυσης κοινωνικών δικτύων συνοψίζεται στα:

- Καλύτερη αξιολόγηση των αναγκών των καταναλωτών.
- Διαμόρφωση target group και ανίχνευση κρυφών πελατών.
- Ανάλυση της απόκρισης του κοινού σχετικά με τα μέτρα που λαμβάνονται κατά τη διάρκεια μιας κρίσης.
- Γνώση αντίδρασης κοινού σε νέες ιδέες ή προϊόντα.
- Κατά την περίοδο των εκλογών, βοηθά στην πρόβλεψη της πρόθεσης των ψήφων.
- Παρακολούθηση τάσεων.
- Επίβλεψη του πόσο "υγιής" είναι μια επιχείρηση ή ένας οργανισμός.
- Προσδιορισμός ατόμων που λειτουργούν ως "influencers".
- Εξοικονόμηση χρόνου για τη λήψη κρίσιμων αποφάσεων.
- Παροχή συγκριτικού πλεονεκτήματος.
- Βελτίωση της απόδοσης της επένδυσης.
- Εντοπισμός προκλήσεων ανταγωνιστών.
- Δημιουργία αποτελεσματικής στρατηγικής.
- Έγκαιρη λήψη και αξιολόγηση ανατροφοδότησης.
- Πρόβλεψη πορείας ενός προϊόντος ή μιας υπηρεσίας.
- Εύρεση δημοτικότητας προσώπων και προϊόντων.
- Ανάλυση εγκλημάτων/ εντοπισμός απάτης/ χρηματικής διαφθοράς.

Προκλήσεις Big Data:

- Ο τρόπος παρουσίασής τους
- Η εύρεση κατάλληλης ποιότητας και ποικιλίας
- Η διαχείριση του κύκλου ζωής τους
- Το απόρρητο και η ασφάλεια
- Η δυνατότητές τους
- Η ετερογένεια τους
- Η ταχύτητα που δημιουργούνται
- Η ακρίβειά τους
- Η αποθήκευσή τους
- Η εξαγόμενη γνώση από αυτά
- Η δημιουργία και η ανάπτυξη εργαλείων και αλγορίθμων ανάλυσης τους

3.Το φάσμα των εδραιωμένων πρακτικών ανάλυσης δεδομένων

Υποδομές που απαιτούνται για την επεξεργασία big data:

Η επεξεργασία απαιτεί:

- ασφάλεια
- ταχύτητα, και
- μικρότερη χρήση πόρων

Η λήψη δεδομένων γίνεται κυρίως πλέον μέσα του Διαδικτύου των Πραγμάτων και με την υποδομή της Υπολογιστικής Νέφους.

Και οι δύο υποδομές αυτές συμβάλλον σε μεγάλο βαθμό στην δημιουργία ενός ασφαλούς περιβάλλοντος για την ανάλυση των Big Data.

Οι τεχνικές που χρησιμοποιούνται για την επεξεργασία δεδομένων

Μηχανική Μάθηση - Μηχανική Μάθηση σε βάθος

Η Μηχανική μάθηση είναι μία τεχνική που παρέχει στο σύστημα τη δυνατότητα να μαθαίνει αυτόματα μέσω δεδομένων και να εξαγάγει σημαντικές πληροφορίες χωρίς να έχει προγραμματιστεί ρητά γι αυτό. Η εξαγωγή πληροφοριών βασίζεται:

- στην εύρεση προτύπων στα δεδομένα και
- στη δημιουργία προβλέψεων βάση των δεδομένων αυτών.

Μηχανική Μάθηση - Μηχανική μάθηση σε βάθος

Η μηχανή μαθαίνει με τη βοήθεια τριών τεχνικών:

- Εποπτευόμενη μάθηση
- Μη εποπτευόμενη μάθηση
- Ενισχυμένη μάθηση

Τα κοινωνικά δίκτυα ως επί το πλείστον αναλύονται με τη μη εποπτευόμενη μάθηση γιατί τα δεδομένα τους συνήθως δεν είναι δομημένα. Οι τεχνικές που χρησιμοποιούνται έχουν κυρίως τον σκοπό δημιουργίας προβλέψεων.

Στον πίνακα παρουσιάζονται τα εργαλεία ανάλυσεις big data

Εργαλεία	Περιγραφή
Apache Hadoop	εργαλείο ανοιχτού κώδικα, αξιόπιστο και ισχυρό. Αποτελείται από στοιχεία γνωστά ως Map Reduce Framework & Hadoop Distrbuted File System.
Apache Storm και Apache Spark	χρησιμοποιούνται κυρίως για κατανεμημένους υπολογισμούς σε πραγματικό χρόνο.
Apache Hive	χρησιμοποιείται για ανάλυση, έρευνα και σύνοψη δεδομένων.
Jaql	χρήση για συναρτησιακή επεξεργασία δεδομένων.
Nosql	κυρίως χρησιμοποιείται για ανάκτηση και αποθήκευση δεδομένων.
Hama and Spark Tools	δημοτικότητα στην έρευνα ανάλυσης κοινωνικών δικτύων.

Στον πίνακα παρουσιάζονται οι κατηγορίες ανάλυσεις big data

Τεχνικές	Περιγραφή
Prescriptive Analytics	Βοηθητικά για την επιλογή της ενέργειας που θα λάβει χώρα.
Predictive Analytics	Προβλεπτικά που σκοπό έχουν την πρόβλεψη του μέλλοντος.
Diagnostic Analytics	Διαγνωστικά, για ανάλυση της προηγούμενης κατάστασης, γιατί συνέβη και πώς θα ξεπεραστεί.
Descriptive Analytics	Περιγραφικά για ανάλυση της τρέχουσας κατάστασης και πρόβλεψη του εγγύς μέλλοντος.

Οι αλγόριθμοι μηχανικής μάθησης που διαδραματίζουν σημαντικό ρόλο στην ανάλυση κοινωνικών δικτύων περιλαμβάνουν τους:

Naive Bayes, Learning tree, Μέγιστη μέθοδο εντροπίας, ταξινομητής Nearest Neighbor, Dynamic Language Model classifier, Μηχανή διανυσμάτων υποστήριξης, NPL, γραμμική παλινδρόμηση & λογιστική παλινδρόμηση, Πολυστρωματικές Αντιλήψεις και Bayes Net.

Αλγόριθμοι μηχανικής μάθησης

Τεχνική	Μέθοδος	Σκοπός Χρήσης
Naive Bayes	Εποπτευόμενη και μη εποπτευόμενη μάθηση	είναι ένας ταξινομητής που χρησιμοποιείται κυρίως για την ταξινόμηση κειμένου με σκοπό να βρει την κατηγορία του εκάστοτε κειμένου/εγγράφου, π.χ. Για ανίχνευση σεξουαλικού ακατάλληλου περιεχομένου ή ανεπιθύμητων μηνυμάτων
Δέντρα Αποφάσεων	Εποπτευόμενη και μη εποπτευόμενη μάθηση	χρησιμοποιούνται για ταξινόμηση κειμένου.
Ταξινομητής - Ο κοντινότερος γείτονας	Εποπτευόμενη μάθηση	χρήση για την αναγνώριση προτύπων από δεδομένα. Η τεχνική χρησιμοποιείται στην ανάλυση κοινωνικών δικτύων όταν δεν υπάρχουν καθόλου ή υπάρχουν ελάχιστες πληροφορίες σχετικά με τα κοινόχρηστα δεδομένα
Μηχανή διανυσμάτων υποστήριξης- SVM	Εποπτευόμενη μάθηση	αναλύει τα δεδομένα που θα χρησιμοποιηθούν για την ανάλυση παλινδρόμησης και την ταξινόμησης
NPL- Αλγόριθμος φυσικής γλώσσας	Εποπτευόμενη και μη εποπτευόμενη μάθηση	χρησιμοποιείται για την ταξινόμηση δεδομένων σε δύο κλάσεις και είναι επέκταση του Naïve Bayes. Και οι δύο χρησιμοποιούν το ίδιο υποθετικό πλαίσιο.
Γραμμική παλινδρόμηση & λογιστική παλινδρόμηση	Εποπτευόμενη μάθηση	χρησιμοποιούνται για προβλέψεις. Η γραμμική παλινδρόμηση βρίσκει τη μοναδική τιμή εξόδου ενώ η λογιστική παλινδρόμηση χρησιμοποιεί πιθανότητες για να δείξει την τιμή εξόδου.
Πολυστρωματικές Αντιλήψεις	Εποπτευόμενη μάθηση	χρησιμοποιείται για την ταξινόμηση δεδομένων αλλά έχει πολλαπλά επίπεδα, στα οποία κάθε προηγούμενο επίπεδο συνδέεται πλήρως με το επόμενο.

Μηχανική μάθηση σε βάθος

Η βαθια μηχανική μάθηση είναι μια εξέλιξη της μάθησης μηχανής που βασίζεται σε τεχνητά νευρωνικά δίκτυα και είναι πολύ δημοφιλής για:

- μοντελοποίηση,
- ταξινόμηση και
- αναγνώριση πολύπλοκων δεδομένων όπως εικόνες, ομιλία και κείμενο.

Διαφορά της με την μηχανική μάθηση:

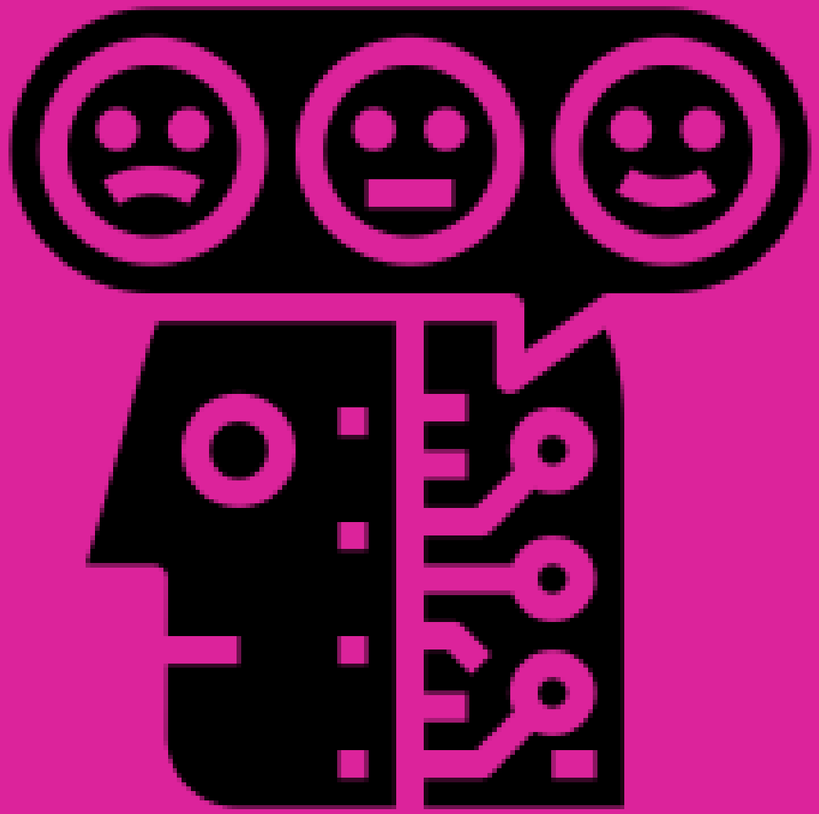
η μηχανική μάθηση αναλύει συνεχώς δεδομένα με σκοπό να μάθει και παίρνει αποφάσεις με βάση αυτό, ενώ η βαθιά μηχανική μάθηση δημιουργεί ένα τεχνητό νευρωνικό δίκτυο το οποίο μπορεί να μάθει και να αποφασίζει αυτόνομα.

Η λογική που λειτουργούν οι αλγόριθμοι βαθιάς μηχανικής μάθησης είναι σαν του ανθρώπινου εγκεφάλου, γι αυτό το λόγο χρησιμοποιούν τεχνητά νευρωνικά δίκτυα.

Η συλλογή των δεδομένων πριν την ανάλυση

Υπάρχουν πολλές διαθέσιμες εφαρμογές για τη συλλογή δεδομένων από Facebook, Twitter, Linked-in ή οποιοδήποτε άλλο κοινωνικό δίκτυο όπως το Twitter API και το Facebook API.
Χρησιμοποιούνται επίσης συνδρομητικές εφαρμογές συλλογής δεδομένων όπως το Gardenhose και το Firehose για τη λήψη δεδομένων από το twitter.

4. Ανίχνευση συναισθημάτων από τα Social Media



Ανάλυση Συναισθημάτων: Ανακαλύπτοντας συναισθήματα από τα Social Media

Η ανάλυση συναισθημάτων (εξόρυξη γνώμης) είναι η διαδικασία του ορισμού και της κατηγοριοποίησης ενός κειμένου με άποψη ανάλογα με το συναίσθημα που περικλείει.
Στα μέσα κοινωνικής δικτύωσης παρέχει πληροφορίες για το πώς αισθάνονται οι χρήστες για ένα brand στο διαδίκτυο.
Αντί για ένα απλό πλήθος αναφορών ή σχολίων, λαμβάνει υπόψη τα συναισθήματα και τις απόψεις.

Η μέτρηση του συναισθήματος είναι πάρα πολύ σημαντική σε κάθε πλάνο ελέγχου των κοινωνικών δικτύων.

Χρειάζεται ανάλυση συναισθημάτων γιατί:

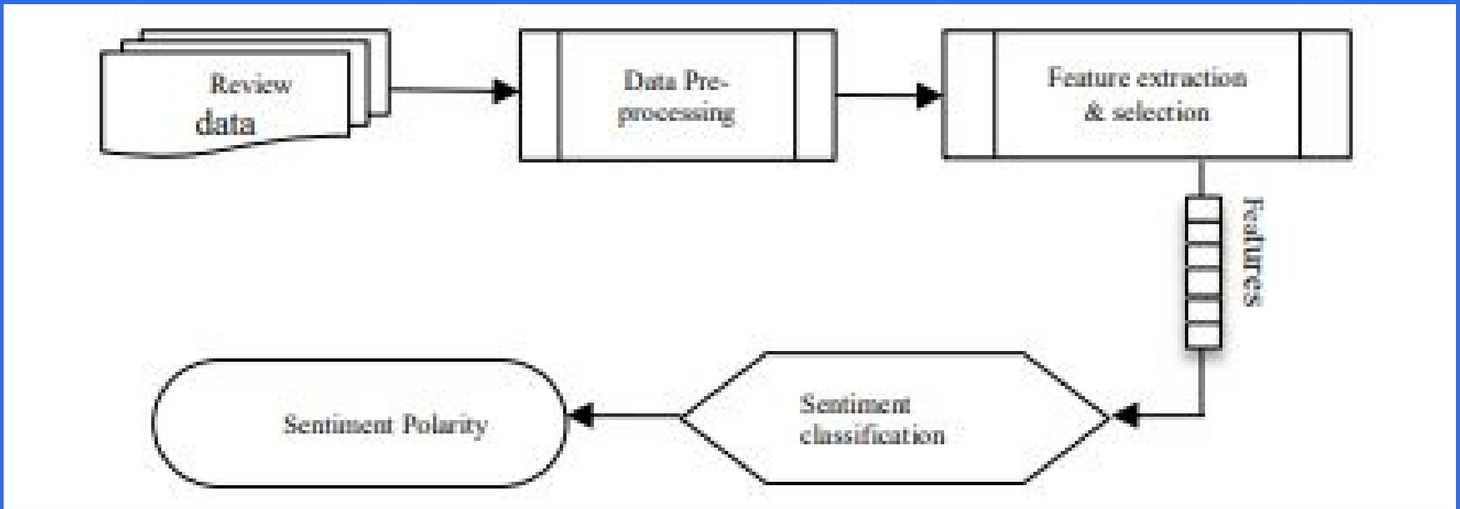
- Συμβάλει στην κατανόηση του κοινού
- Βελτιώνει την εξυπηρέτηση πελατών
- Δημιουργεί στοχευμένα μηνύματα και ανάπτυξη προϊόντων
- Βοηθάει στο να θέτονται ρεαλιστικοί στόχοι
- Αντιμετωπίζει τις κρίσεις στα αρχικά στάδια

Κυρίως για την ανάλυση συναισθημάτων χρησιμοποιείται ένας από τους παρακάτω αλγορίθμους:

- Naive Bayes
- **Δέντρα Αποφάσεων**
- **Μηχανές διανυσμάτων υποστήριξης**

Η διαδικασία της ανάλυσης συναισθημάτων

Η ανάλυση συναισθημάτων χρησιμοποιεί τρεις όρους για να καθορίσει το συναίσθημα. Το αντικείμενο για το οποίο γίνεται η ανάλυση, τα χαρακτηριστικά του αντικειμένου και αυτός που δίνει τη γνώμη του για το αντικείμενο.



Η ταξινόμηση στην ανάλυση συναισθημάτων γίνεται:

Επίπεδο εγγράφου -> εύρεση συνολικής πολικότητας ενός θέματος. (πχ κριτική ταινίας)
Επίπεδο πρότασης -> προϋποθέτει κάθε μία πρόταση να έχει μια ενιαία άποψη.
Επίπεδο χαρακτηριστικών ή επίπεδο άποψης -> η ανάλυση εκτελείται για τα διάφορα χαρακτηριστικά ενός αντικειμένου.

Τα ψυχογραφικά στοιχεία έχει αποδειχθεί ότι αποτελούν πολύτιμο εργαλείο για την τμηματοποίηση της αγοράς και για την κατανόηση των προτιμήσεων των καταναλωτών.

Η προσωπικότητα των καταναλωτών έχει αποδειχθεί
αποτελεσματική για την πρόβλεψη των προτιμήσεων τους.

Οι πληροφορίες που αντλούνται από ψυχολογικές πτυχές της
συμπεριφοράς των χρηστών, είναι κρίσιμες για την κατανόηση και όχι απλώς την
πρόβλεψη των προτιμήσεων του**w.**

Συμβάλλουν τελικά στην έξυπνη λήψη αποφάσεων
στον τομέα του **marketing**.

οι αξίες και η προσωπικότητα των χρηστών, είναι τα κύρια εργαλεία για
την ψυχογραφική τμηματοποίηση τους,

Κατηγορίες τμηματοποίησης Big Five Factor

Big Five Factors	σχετικές λέξεις με θετική πολικότητα	σχετικές λέξεις με αρνητική πολικότητα
Νευρωτισμός	terrible, irony,lazy	Optimism,up, Family
Εξωστρέφια	drinks, dancing, bars, pools, friends, haha	computer, terrific ,not
Ανοικτότητα	humans, art,films, Music, poets	Home,diaper
Ευχαρίστηση	wonderful, morning, visiting, beautiful, spring,together. share, joy	Porn,cost, Anger, Fuck
Ευσυνειδησία	Achievement, Motion, Time	Swearing, shame,idiot, desperate

5.Big Data και Spark Streaming - Μελέτη των δεδομένων σε πραγματικό χρόνο

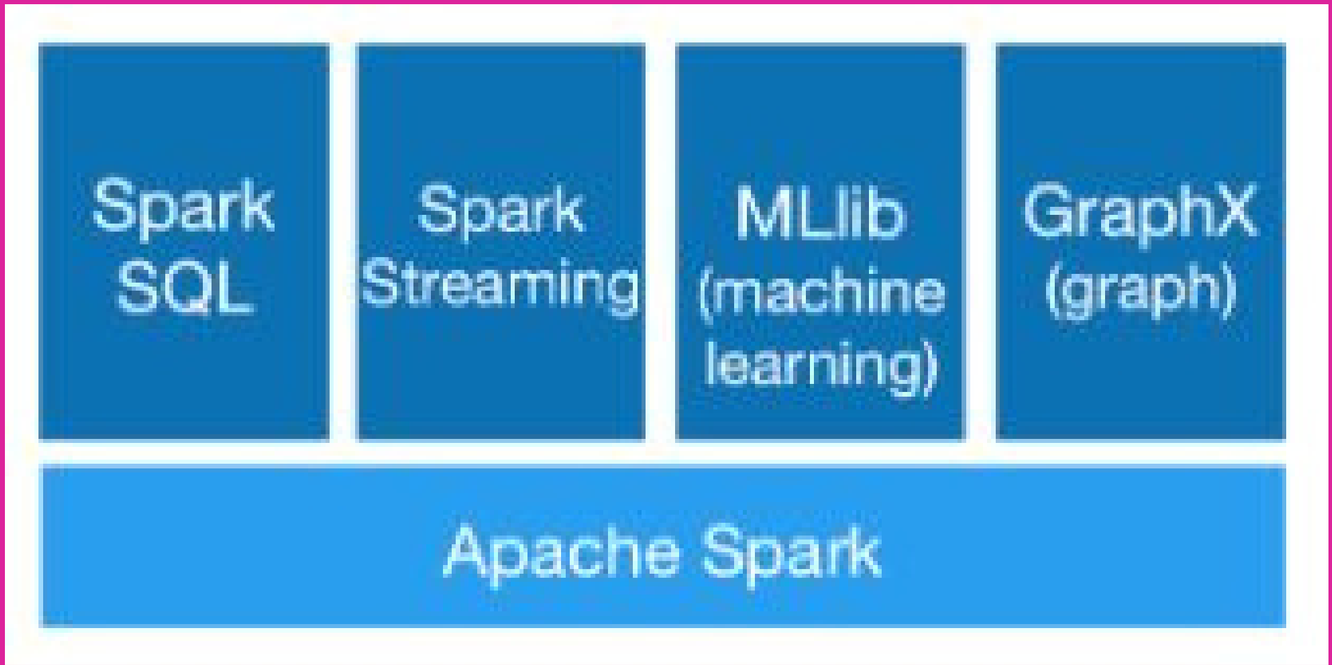
APACHE SPARK

Το Apache Spark ™ είναι μια μηχανή ανάλυσης ανοιχτού κώδικα που δημιουργήθηκε το 2009, για επεξεργασία δεδομένων μεγάλης κλίμακας που επιτυγχάνει υψηλή απόδοση και ταχύτητα τόσο για δεδομένα δέσμης, όσο και για δεδομένα ροής.

Η δημιουργία εφαρμογών είναι αρκετά εύκολη και μπορεί να γίνει σε
Java, Scala, Python, R και SQL.

...Πάνω από 80 λειτουργίες υψηλού επιπέδου που διευκολύνουν την κατασκευή παράλληλων εφαρμογών.

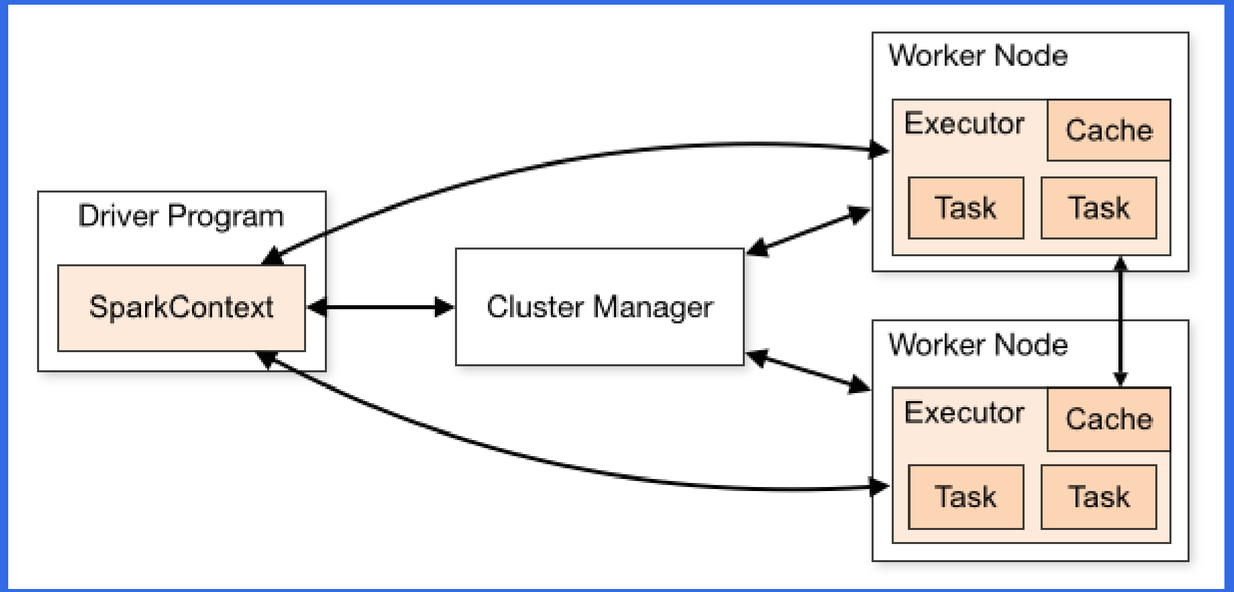
Συνδυάζει SQL, Streamming και σύνθετα εργαλεία ανάλυσης.
Ενεργοποιεί μια στοίβα βιβλιοθηκών, συμπεριλαμβανομένων των SQL και DataFrames, τις MLlib για μηχανική εκμάθηση, και των GraphX και Spark Streaming.
Οι βιβλιοθήκες αυτές μπορούν να χρησιμοποιηθούν και στην ίδια εφαρμογή.



Το Spark Streaming τρέχει παντού!

Στο Hadoop, Apache Mesos, Kubernetes, αυτόνομα ή σε cloud. Μπορεί να έχει πρόσβαση σε διαφορετικές πηγές δεδομένων όπως: HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive και εκατοντάδες άλλες πηγές δεδομένων.

Πως λειτουργεί;



Δομή Δεδομένων RDD

Τα RDDs είναι ανθεκτικά σε σφάλματα, διανεμημένα, σύνολα δεδομένων. Ένα από τα βασικότερα στοιχεία ενός RDD είναι πως είναι αμετάβλητο.
Ένα RDD είναι ένα τεράστιο, ογκώδες σύνολο δεδομένων, που δεν μπορεί να χωρέσει σε καμία μηχανή.

Οι λειτουργίες που μπορούν να πραγματοποιηθούν σε ένα RDD είναι:
map,
flatmap filter,
reduce,
distinct,
sample,
count,
collect κ.α.

Στο Spark όλες οι ενέργειες που γίνονται έχουν να κάνουν είτε με δημιουργία νέων RDDs, είτε με την μετατροπή υπαρχόντων RDDs, είτε τέλος με την εφαρμογή πράξεων σε ένα RDD για τον υπολογισμό ενός αποτελέσματος.

Οι δύο πιο βασικές εντολές είναι η Map (στο στάδιο των μετατροπών) και η Reduce (στο στάδιο των πράξεων). Σαν το MapReduce του Hadoop.

Στον προγραμματισμό με Scala **στο** Spark **δημιουργούμε ένα αντικείμενο** Spark Context **το οποίο είναι αυτό που δημιουργεί τα** RDDs.

Κάτι τέτοιο χρειαζόμαστε και στο SparkStreaming **γι αυτό χρησιμοποιούμε την εντολή**:

```
val ssc = new StreamingContext("local[*]", "PrintTweets", Seconds(1))
```

Apache
Spark vs Hadoop

Είναι ο αντικαταστάτης της διαδικασίας Map Reduce και όχι του Hadoop. Η MapReduce είναι ανεπαρκής για επαναληπτικές και διαδραστικές υπολογιστικές εργασίες.

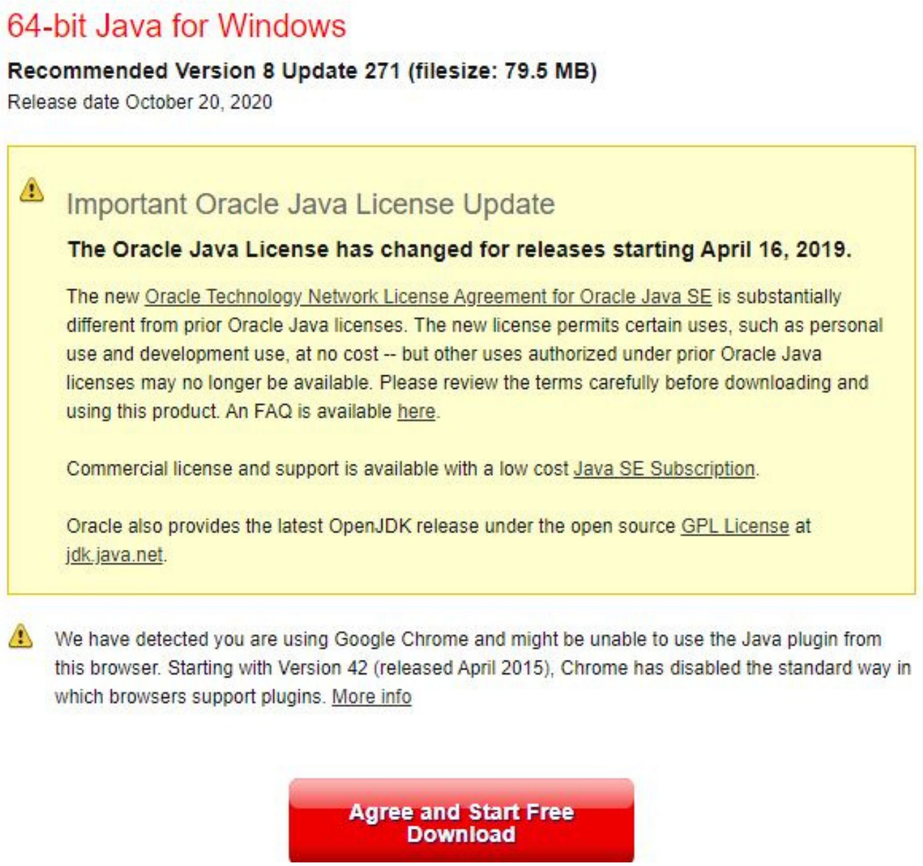
Είναι 100 φορές πιο γρήγορο στη μνήμη και 10 φορές πιο γρήγορο στο δίσκο.

Τίποτα δεν συμβαίνει μέχρι να δοθεί εντολή.

Μπορεί να χρησιμοποιηθεί συνδυαστικά με το Hadoop (σύμπλεγμα YARN) και να γίνει εκμετάλλευση της επεκτασιμότητας και ανθεκτικότητας του Hadoop.

Κάποιες από τις εταιρείες που χρησιμοποιούν το Spark είναι η Amazon, η Yahoo, το Ebay, Tripadvisor.

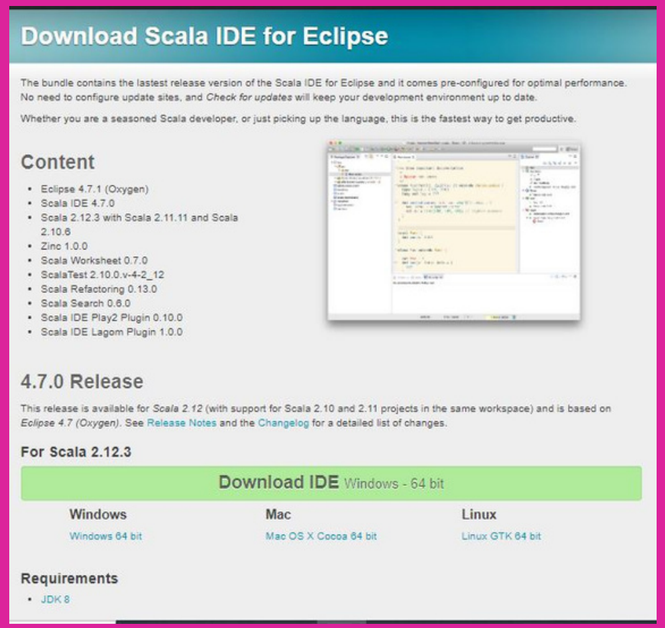
Εγκατάσταση του Spark:



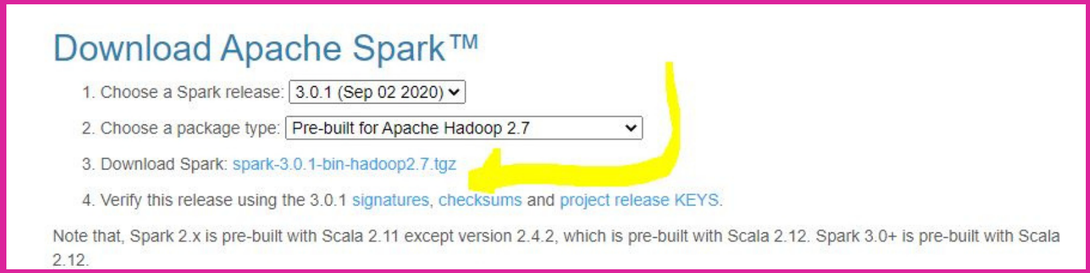
Εγκατάσταση της Java. (Για το παρόν εγχείρημα κατέβηκε η έκδοση 8 από το site java.com)

Προσοχή. Πρέπει η έκδοση της Java να είναι συμβατή με την έκδοση της Scala με την οποία θα πραγματοποιηθεί ο προγραμματισμός του κώδικα στο Eclipse. Για την παρούσα εργασία εγκαταστάθηκε η Scala IDE 4.7.1.. Παρατηρούμε πως στα προαπαιτούμενα είναι η JDK 8 την οποία και κατεβάσαμε. Μόλις κατέβουν οι εφαρμογές πρέπει να εγκατασταθούν τοπικά στο σύστημα.

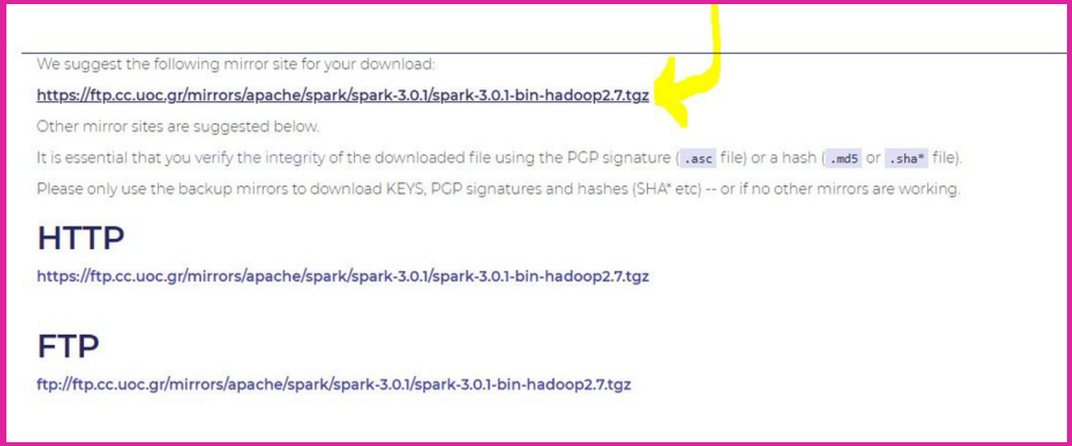
http://scala-ide.org/download/sdk.html



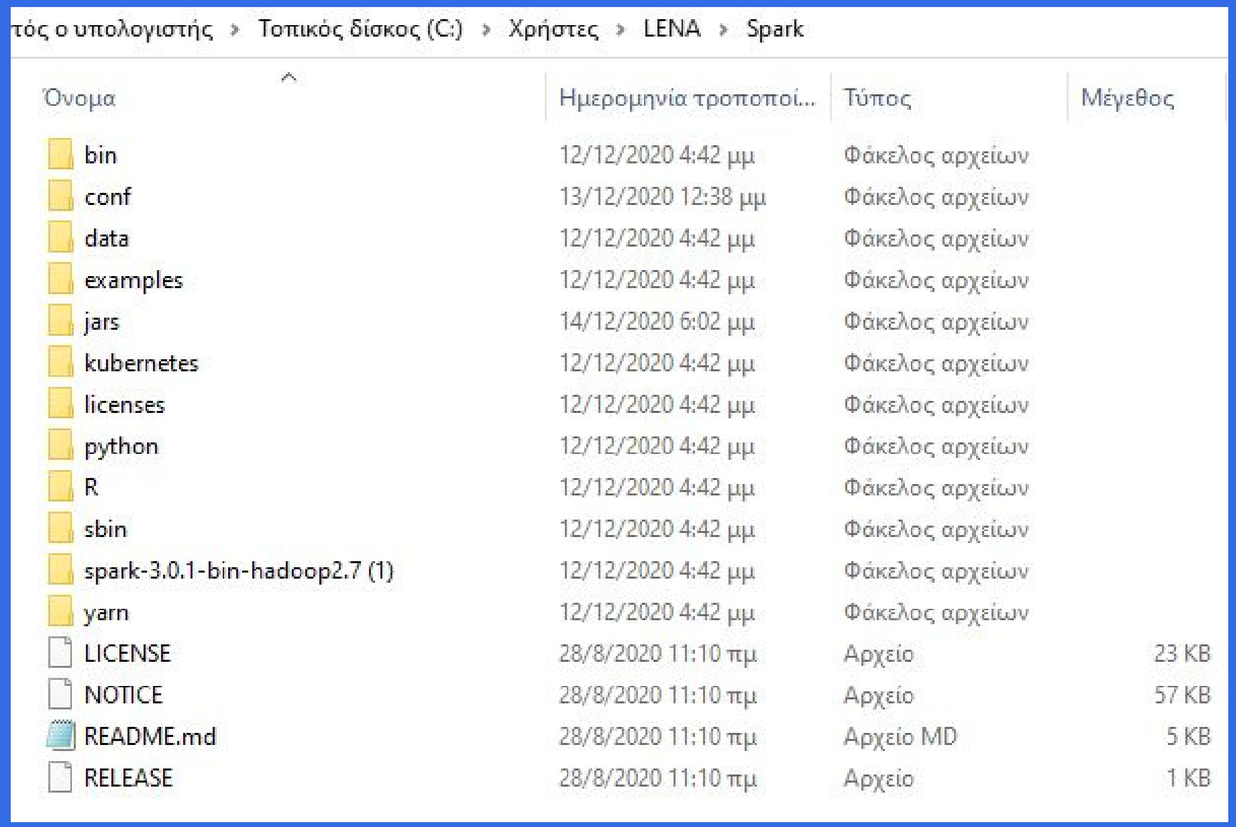
Μετά την εγκατάσταση της Java και της Scala προχωράμε στην εγκατασταση του Spark το οποίο υπάρχει εδώ: https://spark.apache.org/downloads.html .



Επιλέγεται η τελευταία έκδοση και μια εκδοχή pre-built για το Apache Hadoop.

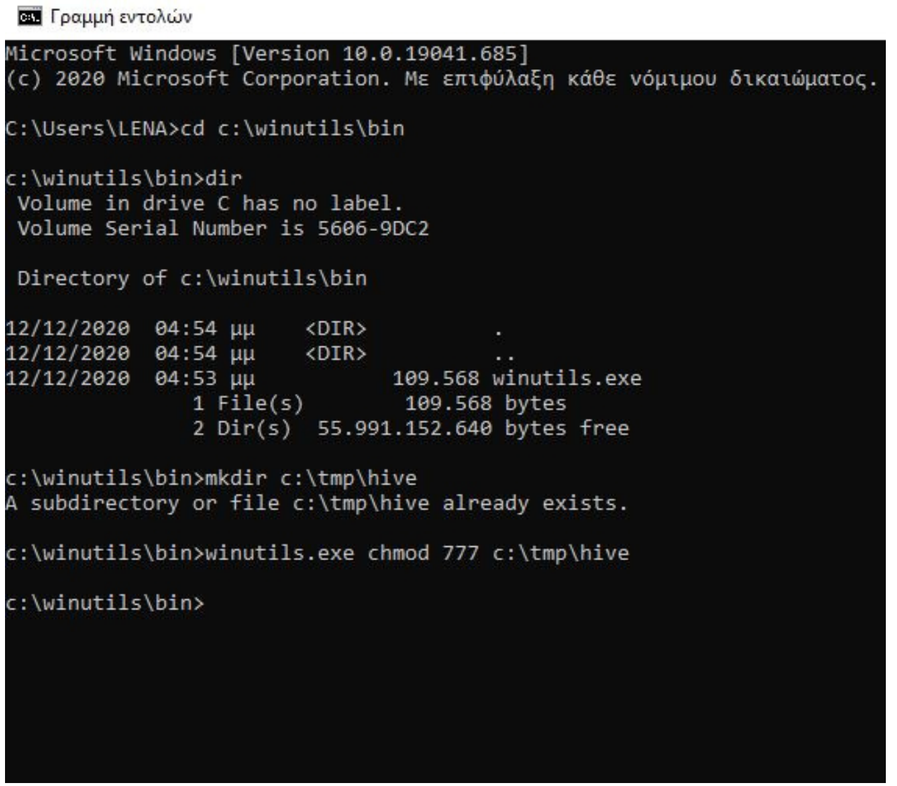


Η διαδρομή τοποθέτησης και τα περιεχόμενα τπυ προγράμματος φαίνονται στην παρακάτω εικόνα:

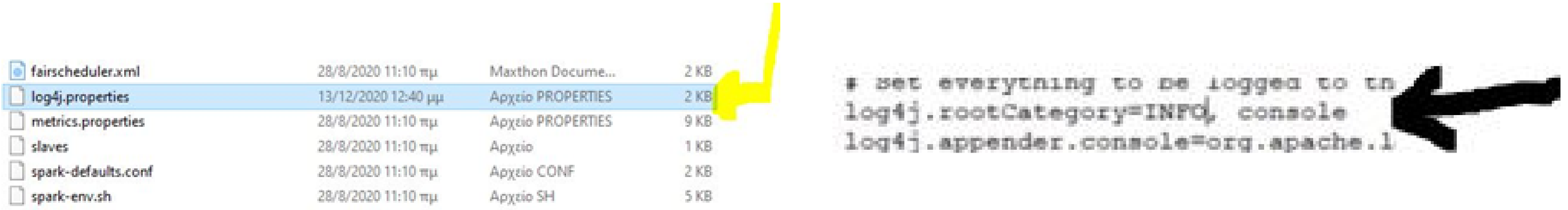


Με το αρχείο winutils.exe τα Windows θα «πιστεύουν» πως υπάρχει το Hadoop. Απαιτείται δημιουργία του φακέλου winutils\bin στον τοπικό δίσκο C και εκεί να γίνει τοποθέτηση του αρχείου winutils.exe. Σκοπός είναι η δημιουργία ενός προσωρινού hive directory ώστε να φαίνεται στα windows πως τρέχουμε το πρόγραμμα hadoop μέσω linux, χωρίς αυτό στη πραγματικότητα να συμβαίνει

Ανοίγουμε το παράθυρο της γραμμής εντολών και ακολουθούμε τα εξής βήματα:

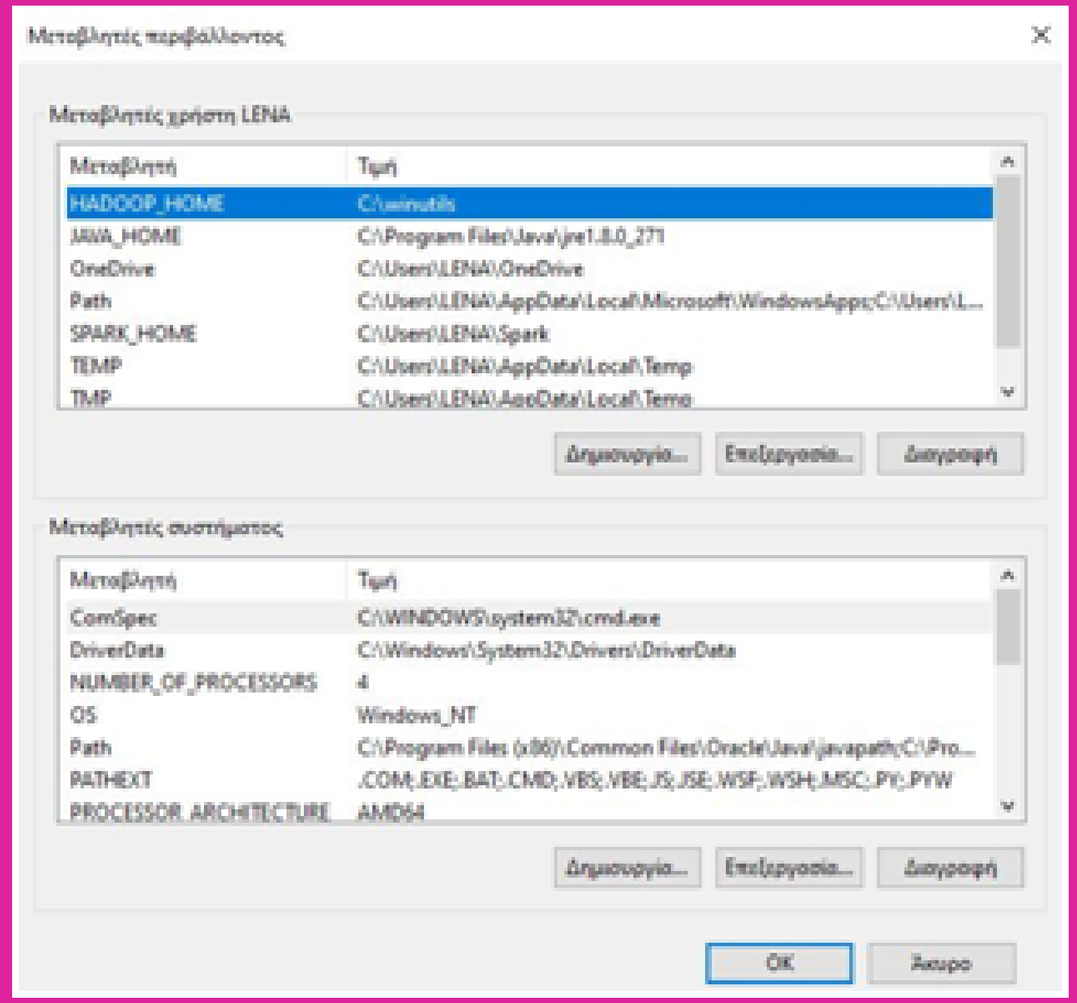


Το Spark **όσο εκτελούνται εργασίες είναι ρυθμισμένο να εμφανίζει πολλά μηνύματα με πληροφορίες. Καλό θα ήταν να μην γίνεται. Για να λυθεί το εξής θέμα πάμε στο φάκελο της εγκατάστασης του προγράμματος, στον υποφάκελο** conf **και στο αρχείο που φαίνεται στην εικόνα:**



Βρίσκουμε την παραπάνω σειρά και αλλάζουμε το INFO **σε** ERROR.

Τέλος ρυθμίζουμε κάποιες από τις μεταβλητές περιβάλλοντος ώστε να ενημερωθούν τα Windows πως το Spark έχει εγκατασταθεί.



Επιλέγουμε δημιουργία και προσθέτουμε τις παρακάτω μεταβλητές:
Όνομα μεταβλητής: SPARK_HOME
Τιμή μεταβλητής: C:\Users\LENA\Spark

Όνομα μεταβλητής: JAVA_HOME
Τιμή μεταβλητής: C:\Program Files\Java\jre1.8.0_271

Όνομα μεταβλητής: HADOOP_HOME
Τιμή μεταβλητής: C:\winutils

Έπειτα πρέπει να τροποποιηθεί και το path directory (**επεξεργασία-> νέο**) **και προσθέτουμε τα εξής:**
%SPARK_HOME%\bin
%HADOOP_HOME%\bin
C:\Program Files\Java\jdk1.8.0_271

```
C:\WINDOWS\system32>cd c:\Users\LENA\Spark

c:\Users\LENA\Spark>dir
 Volume in drive C has no label.
 Volume Serial Number is 5606-9DC2

 Directory of c:\Users\LENA\Spark

13/12/2020  12:37 μμ    <DIR>          .
13/12/2020  12:37 μμ    <DIR>          ..
12/12/2020  04:42 μμ    <DIR>          bin
13/12/2020  12:38 μμ    <DIR>          conf
12/12/2020  04:42 μμ    <DIR>          data
12/12/2020  04:42 μμ    <DIR>          examples
14/12/2020  06:02 μμ    <DIR>          jars
12/12/2020  04:42 μμ    <DIR>          kubernetes
28/08/2020  10:10 πμ            23.312 LICENSE
12/12/2020  04:42 μμ    <DIR>          licenses
28/08/2020  10:10 πμ            57.677 NOTICE
12/12/2020  04:42 μμ    <DIR>          python
12/12/2020  04:42 μμ    <DIR>          R
28/08/2020  10:10 πμ             4.488 README.md
28/08/2020  10:10 πμ               183 RELEASE
12/12/2020  04:42 μμ    <DIR>          sbin
12/12/2020  04:42 μμ    <DIR>          spark-3.0.1-bin-hadoop2.7 (1)
12/12/2020  04:42 μμ    <DIR>          yarn
               4 File(s)         85.660 bytes
              14 Dir(s)  55.992.107.008 bytes free

c:\Users\LENA\Spark>
```

```
c:\Users\LENA\Spark>spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://DESKTOP-U2NSK9D:4040
Spark context available as 'sc' (master = local[*], app id = local-1609763062654).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.1
      /_/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_271)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

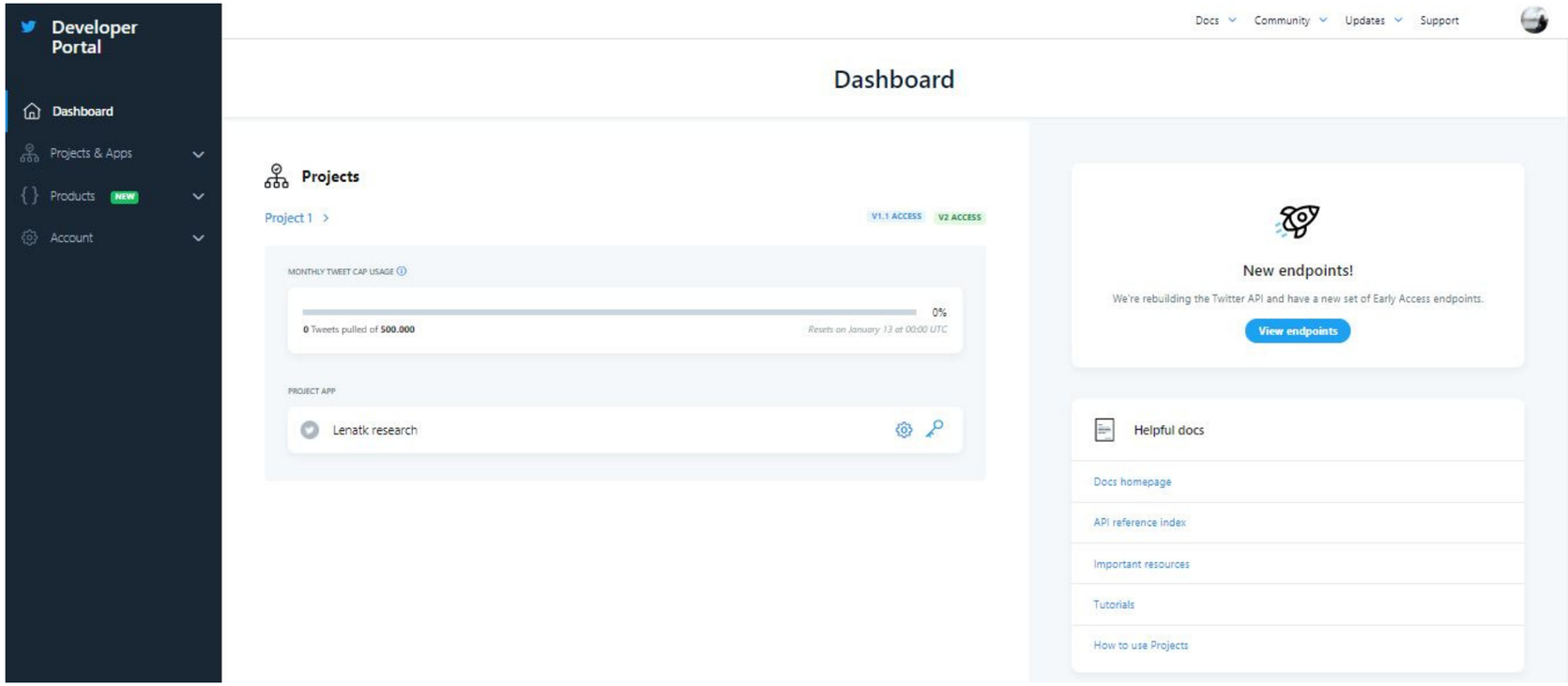
Ανοίγουμε και πάλι την γραμμή εντολών, αυτή την φορά ως διαχειριστής, για να ελέγξουμε αν όλα έχουν πάει καλά.

Πρόσβαση στο Twitter for Developers

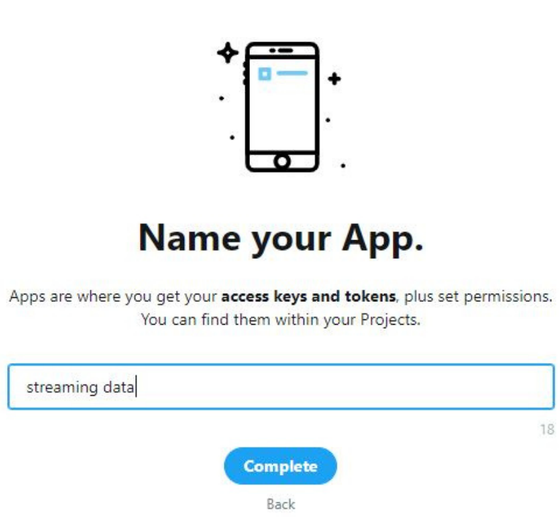
Για να αντλήσουμε δεδομένα από το Twitter απαιτείται η δημιουργία λογαριασμού developer. https://developer.twitter.com/en

Εκεί απαιτείται να δηλωθεί η χρήση που θα γίνει στα δεδομένα που ζητάμε πρόσβαση, ποιοι θα τα δουν και τι ακριβώς είναι τα δεδομένα αυτά (tweets, προσωπικά στοιχεία κ.α.)

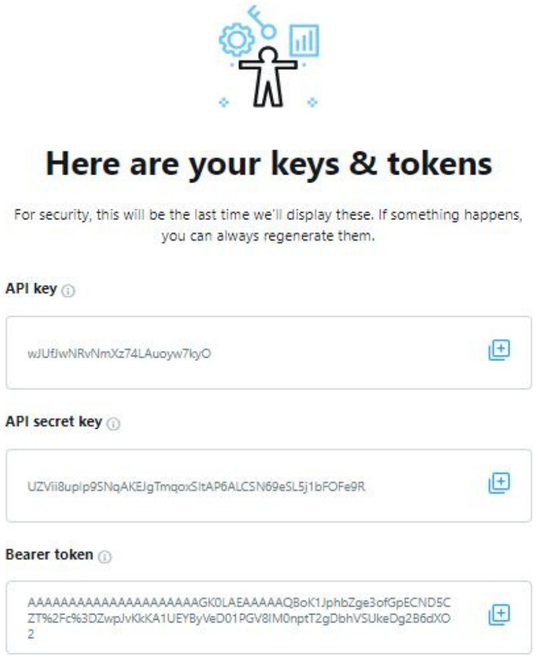
Εφόσον δημιουργήσουμε τον λογαριασμό έχουμε πρόσβαση στη δημιουργία εφαρμογών.



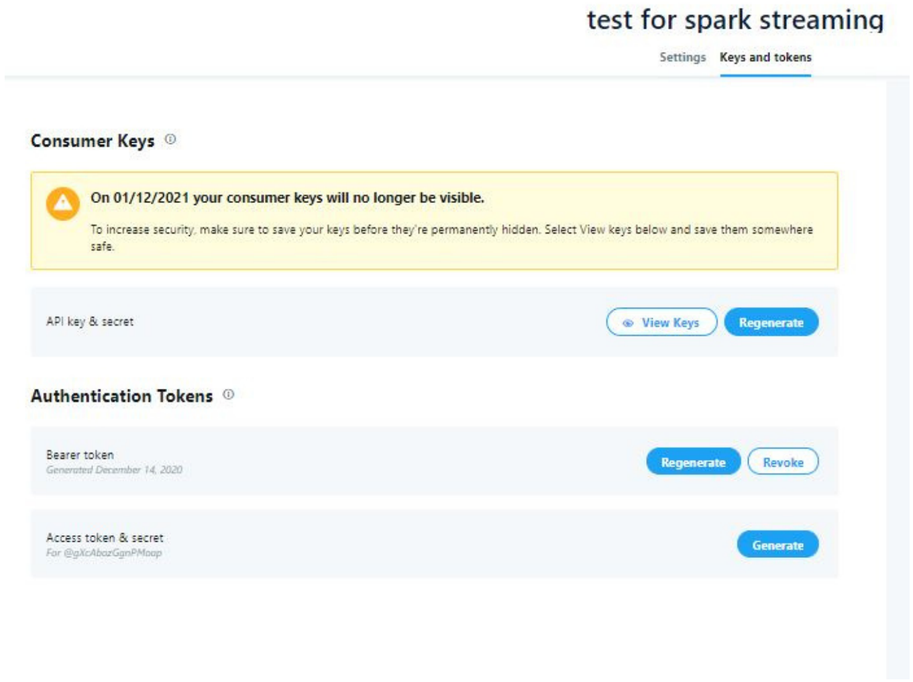
Από την επισκόπηση επιλέγουμε create an app και δίνουμε όνομα.

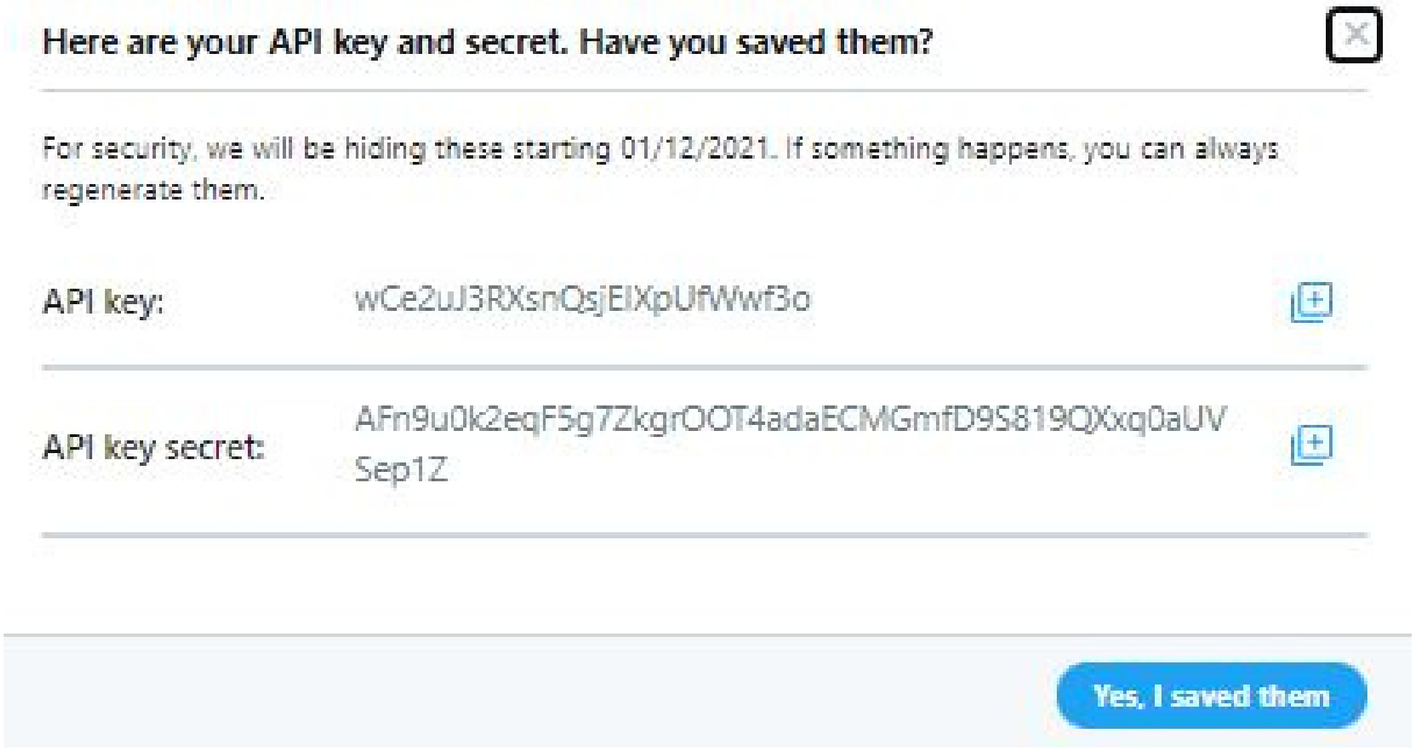


Αυτά που χρειαζόμαστε για το streaming είναι τα keys και τα access tokens. Για παράδειγμα:



Έπειτα επιλέγουμε ρυθμισεις -->Keys and Tokens και εκεί υπάρχουν και τα υπόλοιπα στοιχεία που χρειαζόμαστε.





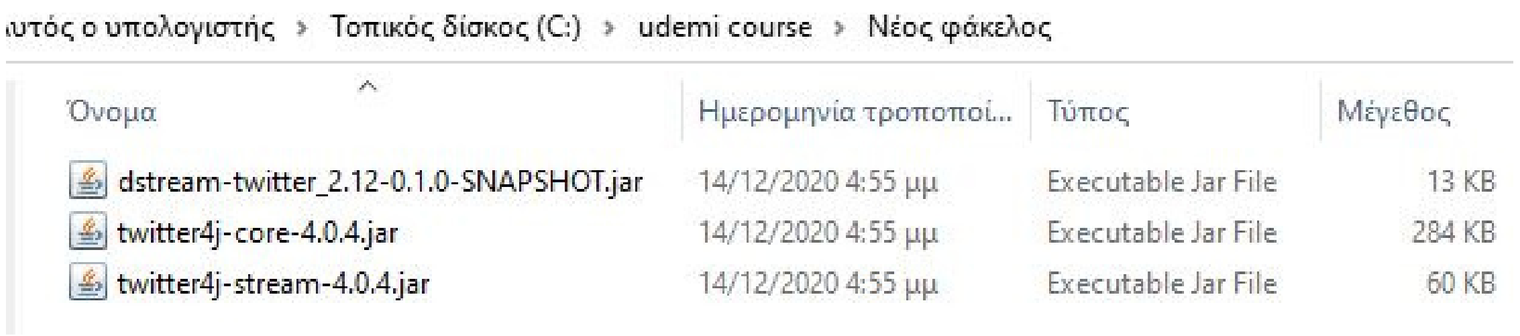
Δημιουργούμε ένα txt αρχείο twitter.txt που μέσα περιέχει όλα τα κλειδιά και τις άδειες. Έπειτα θα το προσθέσουμε στον κώδικα ώστε να μπορέσουμε να κάνουμε streaming. Για παράδειγμα: (Τα συγκεκριμένα κλειδιά δεν είναι πλέον ενεργά)

accessToken 1338077958070820875-E0Q932XaA1aCikb1k5IkzKLXBfDBC3
accessTokenSecret CR13ktWTagpdxWPw3UJuCpiuFBSEl8WP7daoXdLJYHQfP
consumerKey QP2Bve5nVJqNqs15Wf1Qkqz6D
consumerSecret 2BnEP3EzDPN31hvnIpwPlQ5sdQrH9aATGwyYRTHIfq5sYcmcMg

*Το Consumer Key είναι το Api Key

Μέσα στο project που δημιουργήθηκε στο eclipse προστέθηκαν όλες οι απαραίτητες βιβλιοθήκες (Spark Libraries) στο Java Build Path, για να μπορέσουν να εκτελεστούν όλα τα παραδείγματα. Η εξωτερικές αυτές βιβλιοθήκες είναι τα αρχεία .jar που υπάρχουν μέσα στον φάκελο της εγκατάστασης του προγράμματος.

Πέρα από τις βιβλιοθήκες για το Spark χρειαζόμαστε και τις βιβλιοθήκες για το Twitter Streaming.



Με την παρακάτω εντολή, συνδέουμε τα διεπιστευτήρια που έχουμε από το Twiiter στο αρχείο Twitter.txt.

```
setupTwitter()
```

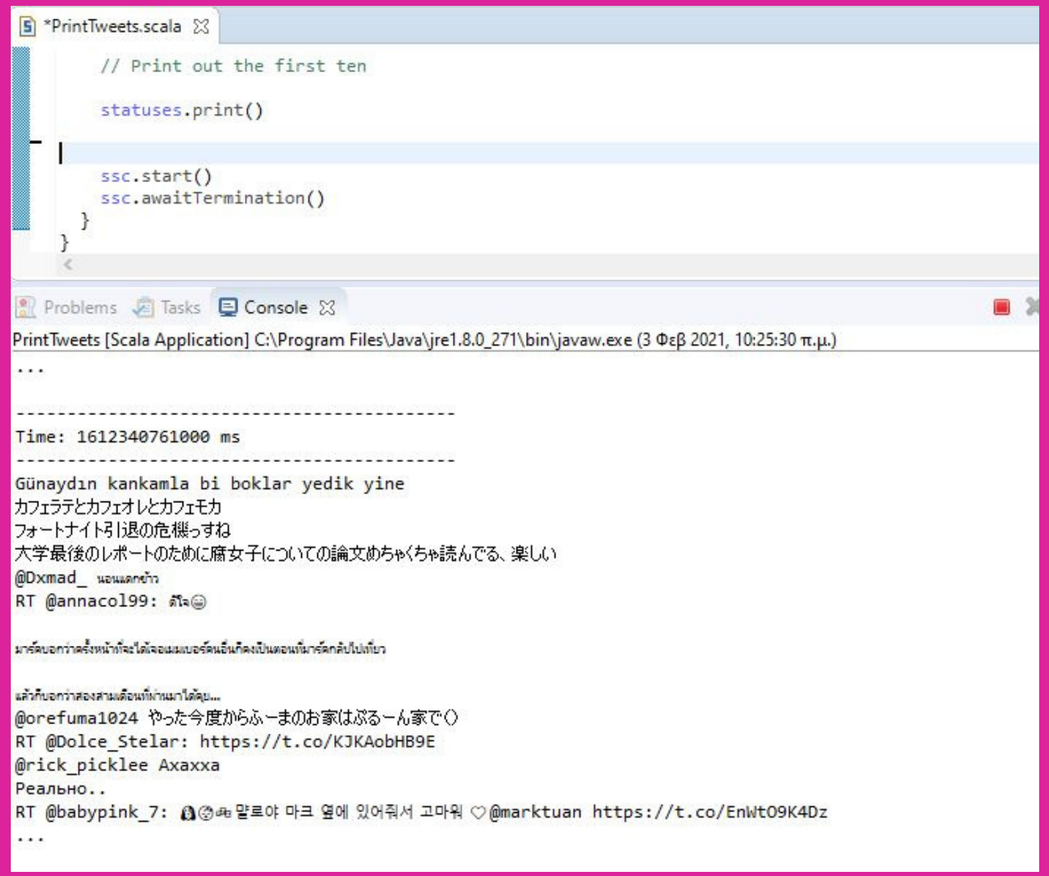
Παραδείγματα με χρήση του Eclipse

1.Εμφάνιση των tweets τοπικά στο δίσκο

```
import org.apache.spark._

object PrintTweets {

  def main(args: Array[String]) {

      setupTwitter()

    val ssc = new StreamingContext("local[*]", "PrintTweets", Seconds(1))

    setupLogging()


    val tweets = TwitterUtils.createStream(ssc, None)


    val statuses = tweets.map(status => status.getText())


    statuses.print()

    ssc.start()
    ssc.awaitTermination()
  }
}
```

1.Εμφάνιση των tweets τοπικά στο δίσκο

Τυχαία στιγμιότυπα με τα tweets που εκτυπώνουμε:



Για να σταματήσει η διαδικασία πατάμε το terminate .

2.Ποιος είναι ο μέσος όρος του μεγέθους των tweets και ποιό το μεγαλύτερο σε χαρακτήρες;

```
import org.apache.spark._

object AverageTweetLength {

  def main(args: Array[String]) {

    setupTwitter()

    val ssc = new StreamingContext("local[*]", "AverageTweetLength", Seconds(1))

    setupLogging()

    val tweets = TwitterUtils.createStream(ssc, None)

    val statuses = tweets.map(status => status.getText())

    val lengths = statuses.map(status => status.length())


    var totalTweets = new AtomicLong(0)
    var totalChars = new AtomicLong(0)
    var mostchars = new AtomicLong(0)
```

2.Ποιος είναι ο μέσος όρος του μεγέθους των tweets και ποιό το μεγαλύτερο σε χαρακτήρες;

```
    lengths.foreachRDD((rdd, time) => {

      var count = rdd.count()

   if (count > 0) {
       totalTweets.getAndAdd(count)

    totalChars.getAndAdd(rdd.reduce((x,y) => x + y))


var maxlength = rdd.reduce((x, y) => Math.max(x, y))

        println("Total tweets: " + totalTweets.get() +
            " Total characters: " + totalChars.get() +
            " Average: " + totalChars.get() / totalTweets.get() +
            " μακρύτερο: " + maxlength
            )

}
    })


    ssc.checkpoint("C:/checkpoint/")
    ssc.start()
    ssc.awaitTermination()
  }
}
```

Problems Tasks Console

AverageTweetLength [Scala Application] C:\Program Files\Java\jre1.8.0_271\bin\javaw.exe (22 Ιαν 2021, 5:30:03 μ.μ.)

```
Total tweets: 17168 Total characters: 1397406 Average: 81 μακρύτερο: 154
Total tweets: 17233 Total characters: 1402566 Average: 81 μακρύτερο: 141
Total tweets: 17272 Total characters: 1405521 Average: 81 μακρύτερο: 144
Total tweets: 17337 Total characters: 1410847 Average: 81 μακρύτερο: 142
Total tweets: 17402 Total characters: 1416027 Average: 81 μακρύτερο: 140
Total tweets: 17472 Total characters: 1421885 Average: 81 μακρύτερο: 144
Total tweets: 17534 Total characters: 1427077 Average: 81 μακρύτερο: 146
Total tweets: 17597 Total characters: 1432856 Average: 81 μακρύτερο: 167
Total tweets: 17666 Total characters: 1438490 Average: 81 μακρύτερο: 141
Total tweets: 17743 Total characters: 1444529 Average: 81 μακρύτερο: 147
Total tweets: 17805 Total characters: 1448970 Average: 81 μακρύτερο: 145
Total tweets: 17868 Total characters: 1454423 Average: 81 μακρύτερο: 140
```

3.Τα 10 πιο δημοφιλή hashtags που χρησιμοποιούν οι χρήστες

```
import org.apache.spark._

object PopularHashtags {

  def main(args: Array[String]) {

    setupTwitter()

    val ssc = new StreamingContext("local[*]", "PopularHashtags", Seconds(1))

    setupLogging()

    val tweets = TwitterUtils.createStream(ssc, None)
    val statuses = tweets.map(status => status.getText())

    val tweetwords = statuses.flatMap(tweetText => tweetText.split(" "))

    val hashtags = tweetwords.filter(word => word.startsWith("#"))

    val hashtagKeyValues = hashtags.map(hashtag => (hashtag, 1))

    val hashtagCounts = hashtagKeyValues.reduceByKeyAndWindow( (x,y) => x + y,
        (x,y) => x - y, Seconds(300), Seconds(1))

    val sortedResults = hashtagCounts.transform(rdd => rdd.sortBy(x => x._2, false))

    sortedResults.print
    ssc.checkpoint("C:/checkpoint/")
    ssc.start()
    ssc.awaitTermination()
  }
}
```

3.Τα 5 πιο δημοφιλή hashtags που χρησιμοποιούν οι χρήστες

```
PopularHashtags [Scala Application] C:\Program Files\Java\jre1.8.0_271\bin\javaw.exe (23 Ιαν 2021, 6:26:00 μ.μ.)
(#세븐틴,3)
(#SVT_IN_COMPLETE,3)
...

-------------------------------------------
Time: 1611419243000 ms
-------------------------------------------
(#güzelgünlere,7)
(#Master,5)
(#edabolat,5)
(#SEVENTEEN,4)
(#โมแม้ไบร์ทวันแห่งวัดจนเสียงแหบ,4)
(#ファミマのいちご狩り,3)
(#fashion…,3)
(#goLOUD,3)
(#세븐틴,3)
(#SVT_IN_COMPLETE,3)
...
```

```
PopularHashtags [Scala Application] C:\Program Files\Java\jre1.8.0_271\bin\javaw.exe (23 Ιαν 2021, 6:26:00 μ.μ.)
(#세븐틴,7)
(#imgxnct,7)
...

-------------------------------------------
Time: 1611419354000 ms
-------------------------------------------
(#güzelgünlere,25)
(#โมแม้ไบร์ทวันแห่งวัดจนเสียงแหบ,20)
(#edabolat,19)
(#SEHUN,11)
(#23ETriunfoDelPueblo,8)
(#SEVENTEEN,8)
(#ファミマのいちご狩り,8)
(#23Ene,7)
(#세븐틴,7)
(#imgxnct,7)
...
```

Εδώ βλέπουμε και πάλι 2 τυχαία στιγμιότυπα

Συμπερασματικά..

Χωρίς τα big data το social media marketing δεν θα ήταν στο σημείο που είναι σήμερα. Τα big data ανοίγουν το δρόμο για μια συναρπαστική νέα εμπειρία για επαγγελματίες και πελάτες.
Με την πρόσβαση σε big data κοινωνικών δικτύων και με τη χρήση των κατάλληλων εργαλείων για την ανάλυση τους, η σκηνή ψηφιακού marketing θα αλλάξει πάρα πολύ γρήγορα.

Είναι ένα ζωτικό εργαλείο για σχεδόν κάθε πτυχή της ζωής μας. Ακόμα και αν σε κάποια βιομηχανία δεν χρησιμοποιούνται σήμερα, δεν είναι απίθανο να χρησιμοποιηθούν στο μέλλον. Με την άνθιση της τεχνολογίας, που φαίνεται από τα έξυπνα σπίτια και τις συσκευές που συλλέγουν στοιχεία και γίνονται όλο και μικρότερες και πιο αποτελεσματικές, είναι θέμα χρόνου να δούμε αλλαγή στην αντίληψή μας για τον κόσμο.

Όλες οι έρευνες καταλήγουν πως η μεγαλύτερη πρόκληση είναι ο όγκος των δεδομένων που προκύπτει κυρίως εξαιτίας της φύσης τους να δημιουργούνται σε πραγματικό χρόνο. Γι αυτό με τη σειρά του είναι πολύ σημαντικός και ο καθαρισμός τους γιατί συνήθως έχουν αδόμητο και αβέβαιο χαρακτήρα.

Απαιτείται περαιτέρω εργασία και μελέτη για τον προσδιορισμό των τεχνικών μηχανικής μάθησης που απαιτούν τροποποιήσεις για την ανάλυση κοινωνικών δικτύων και για τον συνδυασμό τους με τεχνικές ανάλυσης συναισθημάτων για καλύτερη απόδοση στην συνολική ανάλυση των κοινωνικών μέσων.

Συμπερασματικά..

Το Spark προσφέρει μια πληθώρα δυνατοτήτων για την ανάλυση και ερμηνεία των Big Data. Το Spark Streaming και η επεξεργασία δεδομένων σε πραγματικό χρόνο όμως, λύνει τα χέρια στους ερευνητές και έρχεται να κουμπώσει τέλεια στην ίδια την φύση των δεδομένων που δεν είναι άλλη, από το να δημιουργούνται σε πραγματικό χρόνο. Το θέμα σίγουρα απαιτεί περεταίρω διερεύνηση.

Στην παρούσα εργασία ασχοληθήκαμε μόνο με κάποια παραδείγματα. Θα μπορούσαμε να βρούμε και να εμφανίσουμε τα σχόλια των χρηστών που περιέχουν κάποιο συγκεκριμένο hashtag. Για παράδειγμα να βρούμε όλα τα σχόλια των χρηστών που περιέχουν το hashtag #covid19. Επίσης θα μπορούσαμε να αναζητήσουμε σχόλια με βάση λέξεις που εκδηλώνουν συναισθήματα χρηστών σε σχέση με ένα επίκαιρο θέμα όπως, ο Κορωνοιός ώστε να κάνουμε ανάλυση συναισθημάτων. Οι λέξεις μπορούν να βρεθούν από κατάλληλο λεξικό συναισθημάτων και μπορεί να υποδηλώνουν συναισθήματα όπως θυμό, φόβο, αγωνία, αμφισβήτηση, ελπίδα κ.α.

ΕΥΧΑΡΙΣΤΩ ΓΙΑ ΤΗΝ ΠΡΟΣΟΧΗ ΣΑΣ!