Ανάλυση Συναισθημάτων Κειμένου σε Hadoop & Spark

Διπλωματική Εργασία

Αλέξανδρος Μήλας (Α.Μ.: 711141293)

Επιβλέπων Καθηγητής: **Βασίλειος Μάμαλης**

Περίγραμμα Παρουσίασης

- **Θεωρητικό Υπόβαθρο**
  - Κατηγοριοποίηση, Naive Bayes, SVM, Ανάλυση Συναισθημάτων, MapReduce
- **Ανάπτυξη Υλοποιήσεων Ανάλυσης Συναισθημάτων Κειμένου**
  - Hadoop, Spark
- **Πειραματικά Αποτελέσματα και Αξιολόγηση**
  - Ορθότητα, Μέτρο F1, Χρόνος Εκτέλεσης, Κλιμακωσιμότητα, Επιτάχυνση
- **Συμπεράσματα και Επεκτάσεις**

Εισαγωγή

- Ο όγκος της αξιοποιήσιμης πληροφορίας ολοένα και αυξάνεται
- Τα πλούσια σε πληροφορία δεδομένα του διαδικτύου δημιουργούν έντονο ενδιαφέρον για την συνεχή μελέτη νέων δυνατοτήτων
- Η μερίδα του λέοντος στο πολυμεσικό περιεχόμενο του διαδικτύου ανήκει ακόμα στο κείμενο, που βρίσκεται σε μεγάλες ποσότητες και σε άτακτη ή ημιδομημένη μορφή σε πλατφόρμες κοινωνικής δικτύωσης όπως το *Twitter* ή το *Facebook*
- Η επεξεργασία αυτού του όγκου πληροφορίας προβληματίζει ως προς την αναζήτηση αποδοτικών τρόπων επεξεργασίας του και τον εξοπλισμό που χρειάζεται

Εξόρυξη Κειμένου & Κατηγοριοποίηση

Βήματα Εξόρυξης Κειμένου
1. Προεπεξεργασία Κειμένου
2. Μετατροπή Κειμένου
3. Επιλογή Χαρακτηριστικών
4. Εξόρυξη Κειμένου
5. Αξιολόγηση Αποτελεσμάτων

Μήτρα Σύγχυσης

		Προβλεφθείσα Κατηγορία	
		+	-
Πραγματική Κατηγορία	+	True Positive	False Negative
	-	False Positive	True Negative

Σχηματική Απεικόνιση Κατηγοριοποίησης

Naïve Bayes

- Πιθανοτική μέθοδος κατηγοριοποίησης όπου για κάθε έγγραφο προς έλεγχο υπολογίζεται η πιθανότητα αυτό να ανήκει σε κάθε κλάση, με την μέγιστη αυτών να είναι εκείνη που στην οποία θα ταξινομηθεί
- Συχνή εφαρμογή σε προβλεπτικά και περιγραφικά μοντέλα λόγω της διαφάνειας επί της ταξινόμησης που προσφέρει
- Η ύπαρξη μηδενικών τιμών των πιθανοτήτων υπό συνθήκη για χαρακτηριστικά που εμφανίζονται μόνο σε κάποιες κλάσεις του συνόλου εκπαίδευσης αντιμετωπίζεται με την **εξομάλυνση Laplace**

υπό συνθήκη πιθανότητα το έγγραφο x να ανήκει στην κλάση y

εκ των προτέρων πιθανότητα το έγγραφο να ανήκει στην κλάση y

υπό συνθήκη πιθανότητα το χαρακτηριστικό xi να ανήκει σε έγγραφο της κλάσης y

$$P(x_i|y) = \frac{\sum x_i, y + 1}{\sum x, y \cdot |V|}$$

$$P(y|x) \propto P(y) \cdot \prod_{i=1}^{n_x} P(x_i|y)$$

Σχέση Πιθανότητας Ταξινόμησης Εγγράφου x στην κλάση y με Naïve Bayes

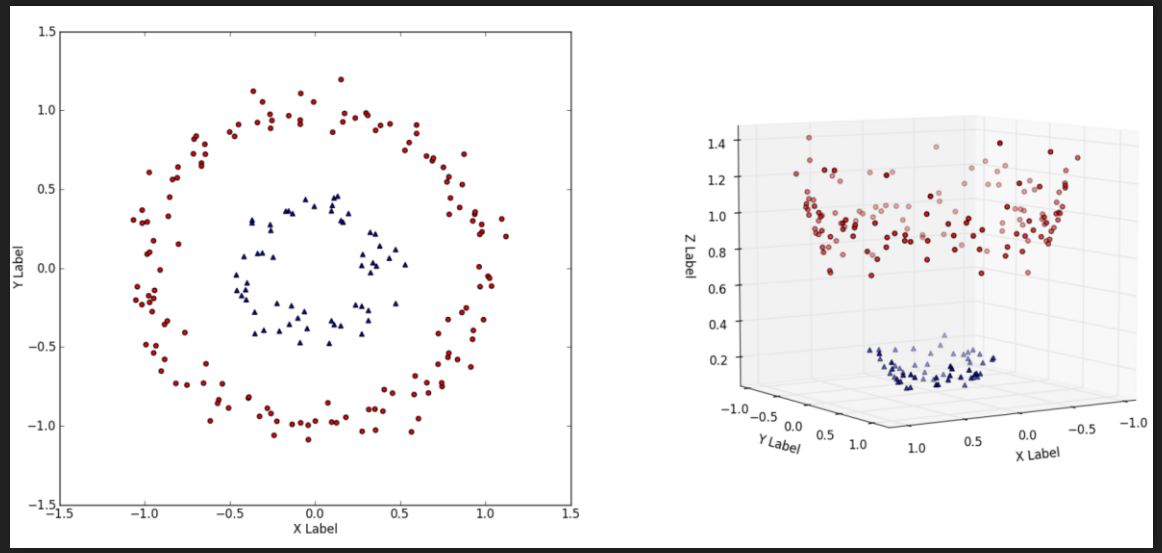
$$y_{map} = argmax(P(y|x)) = argmax(P(y) \cdot \prod_{i=1}^{n_x} P(x_i|y))$$

Σχέση Βέλτιστης Πιθανότητας Ταξινόμησης Εγγράφου x στην Κλάση y

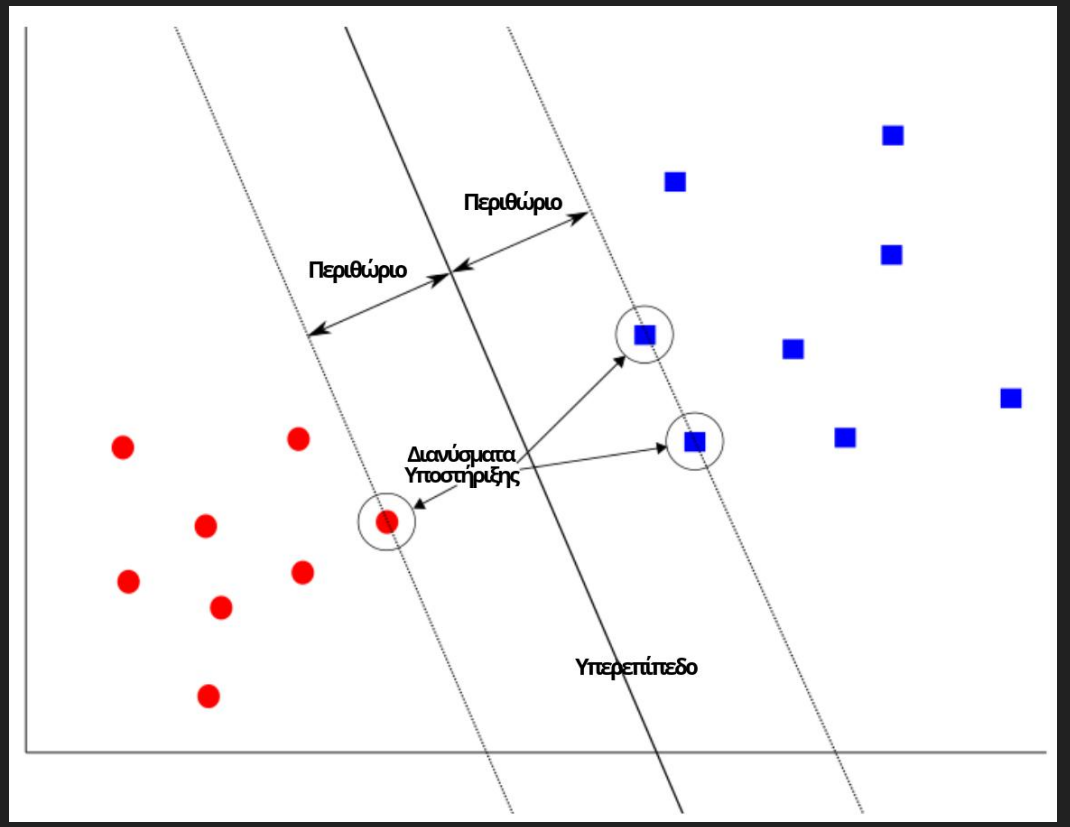
Support Vector Machines

$$w^T \cdot x + b = 0$$

Γενική Εξίσωση Διαχωριστικού Υπερεπιπέδου SVM



Μετασχηματισμός Δεδομένων από 2 σε 3 Διαστάσεις για Πλήρη Διαχωρισμό



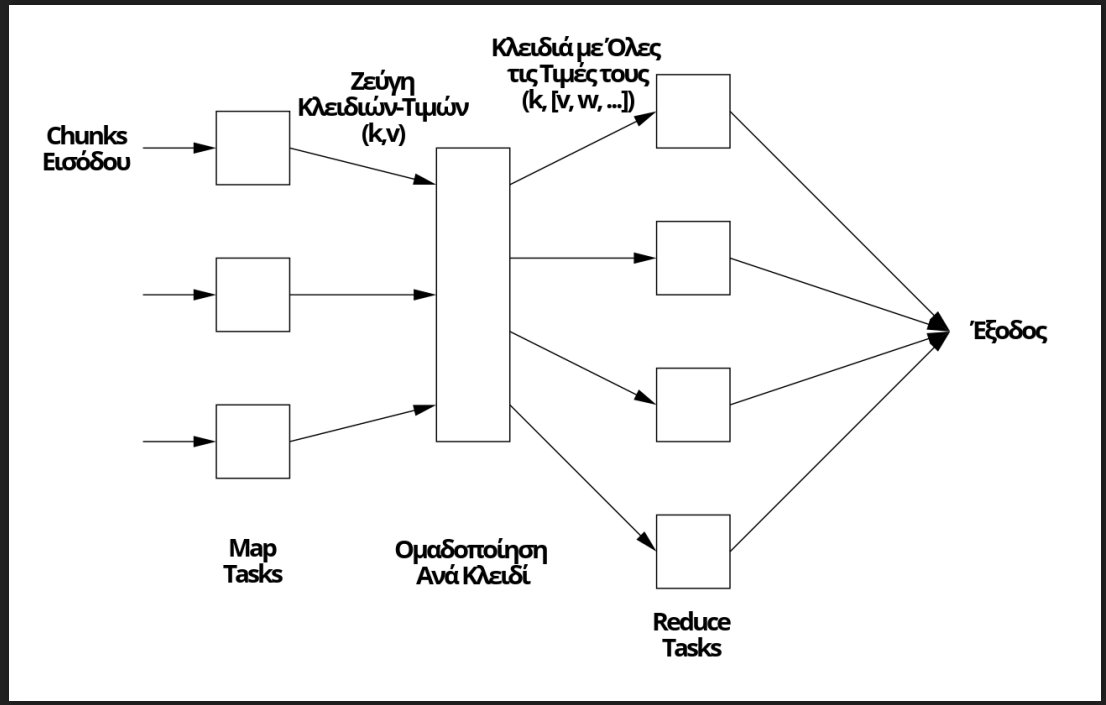
Διανύσματα Υποστήριξης, Περιθώρια, και Υπερεπίπεδο SVM

Ανάλυση Συναισθημάτων

- Η ανάλυση συναισθημάτων αναφέρεται στην μελέτη της συναισθηματικής πολικότητας που εκφράζει μια οντότητα (πρόσωπο, έγγραφο, κλπ.) απέναντι σε κάποια άλλη οντότητα (π.χ. ένα δημόσιο πρόσωπο, κάποιο προϊόν, ένα γενικό θέμα προς συζήτηση κ.α.)
- Θεωρείται γενικότερα ένα πρόβλημα κατηγοριοποίησης με επίλυση μέσω μεθόδων μηχανικής μάθησης και βρίσκεται στην θεματική περιοχή που συνδυάζεται η επεξεργασία φυσικής γλώσσας με την ανάκτηση πληροφορίας και την εξόρυξη κειμένου
- Βρίσκει πρόσφορο έδαφος στη σύγχρονη πραγματικότητα του διαδικτύου και κυρίως στις αναρτήσεις των κοινωνικών δικτύων
- Παρόλα αυτά, τα δεδομένα από τέτοιες πηγές ενδέχεται να περιέχουν αρκετό θόρυβο

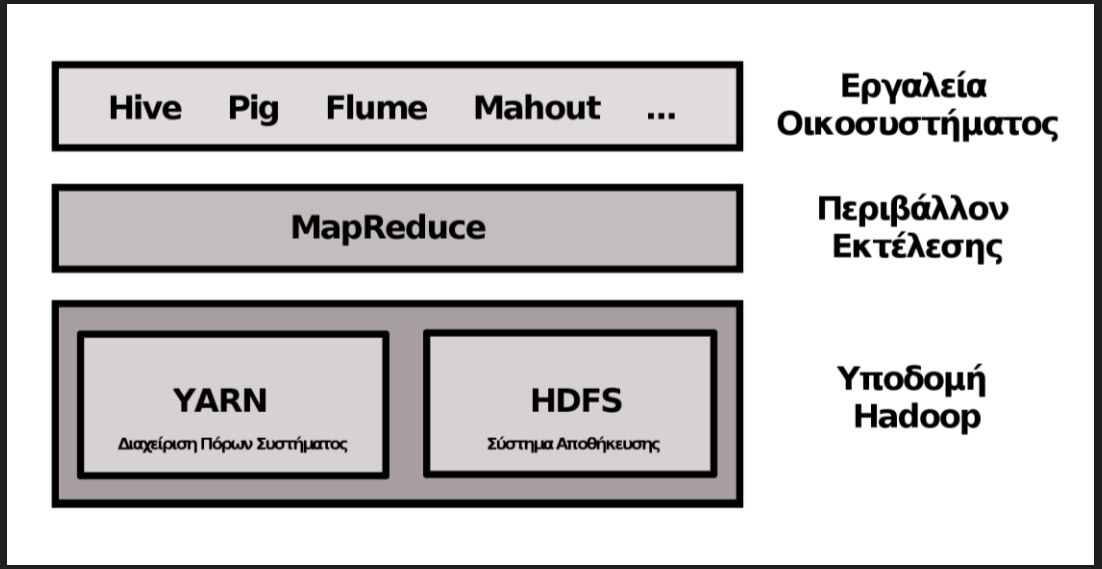
MapReduce

- Ένας αριθμός Map tasks δέχονται ως είσοδο έναν αριθμό chunks από το κατανεμημένο σύστημα αρχείων, τα οποία μετατρέπουν σε μια ακολουθία ζευγαριών κλειδιού-τιμής (k,v)
- Τα ζεύγη του προηγούμενου βήματος συλλέγονται και ταξινομούνται ανά κλειδί πριν μοιραστούν σε έναν αριθμό Reduce tasks
- Κάθε Reduce task επεξεργάζεται κάθε κλειδί ξεχωριστά συγκεντρώνοντας όλες τις τιμές του και τα αποτελέσματα που προκύπτουν αποθηκεύονται στο κατανεμημένο σύστημα αρχείων.

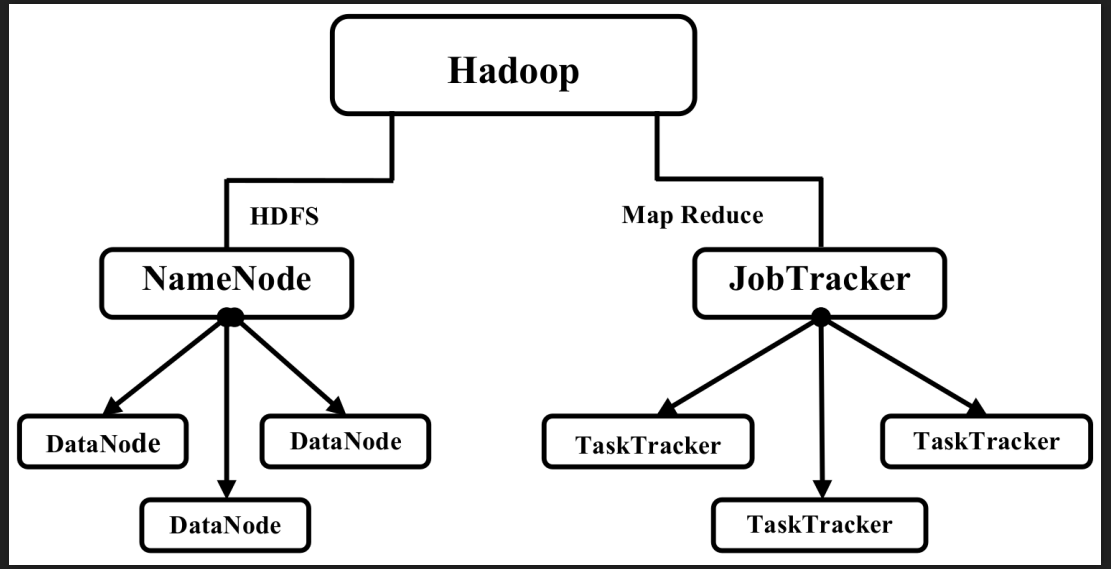


Σχηματική Απεικόνιση Υπολογισμού σε MapReduce

Apache Hadoop

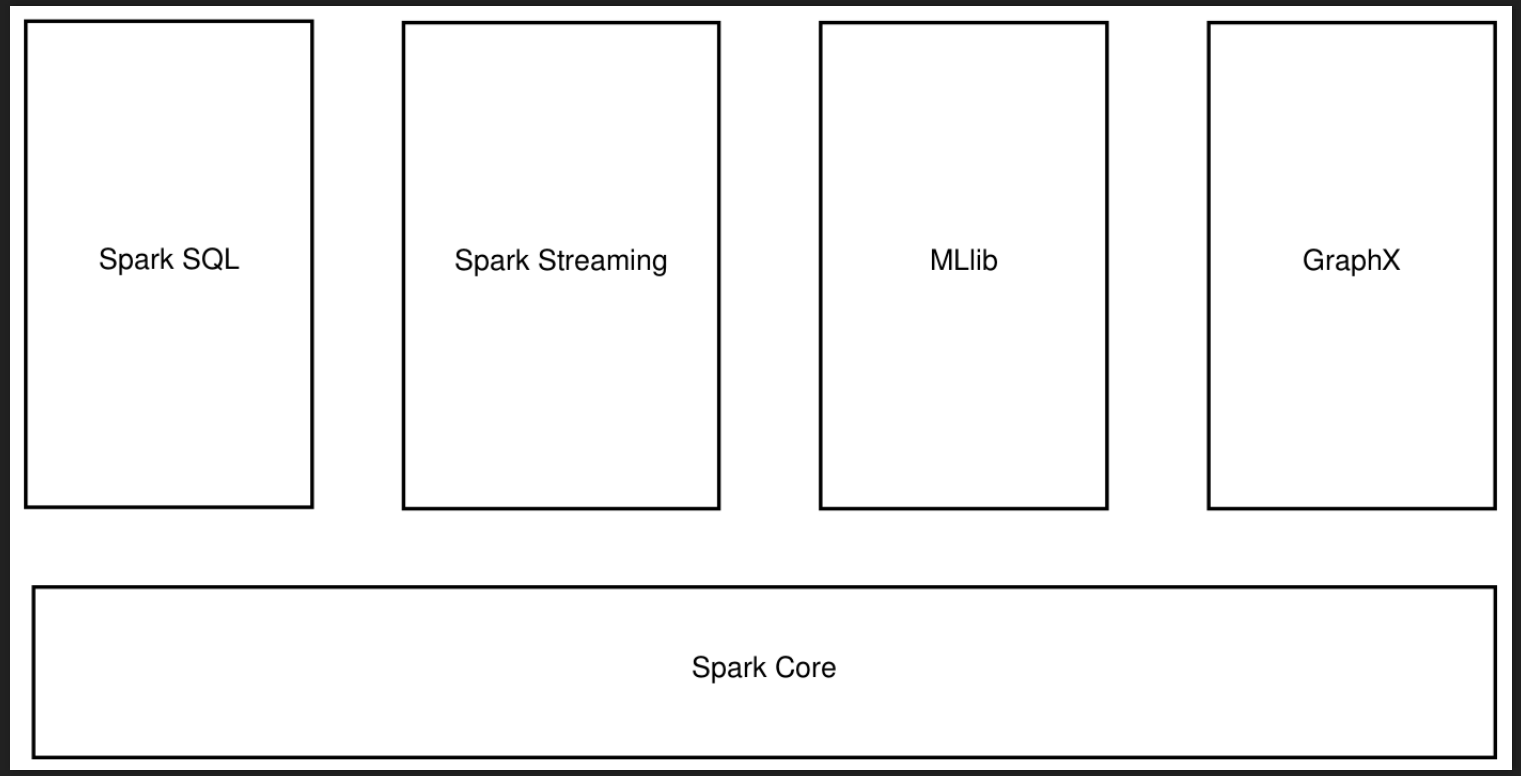


Δομή Συνιστωσών Apache Hadoop



Είδη Κόμβων στο Apache Hadoop

Apache Spark



Δομή Συνιστωσών Apache Spark

Ανάπτυξη Υλοποιήσεων Ανάλυσης Συναισθημάτων Κειμένου

- Η θεματική περιοχή στην οποία δραστηριοποιείται η εργασία αποτελεί κατά μία έννοια την τομή μεταξύ της κατηγοριοποίησης και της παράλληλης εκτέλεσης σε κατανεμημένο περιβάλλον
- Για την πλήρη αξιολόγηση της αποτελεσματικότητας των υλοποιήσεων, κάθε μία πρέπει να κριθεί βάσει των αναγκών και των δύο πεδίων που εργάζεται
- Σε μια προσπάθεια στοχευμένων βελτιστοποίησεων, αναπτύχθηκαν για κάθε υλοποίηση μια απλή και μια τροποποιημένη έκδοση της που στοχεύει στην ανάκτηση καλύτερων αποτελεσμάτων είτε στην ταξινόμηση εγγράφων είτε στην παράλληλη εκτέλεση βάσει του όγκου δεδομένων εισόδου/διαθέσιμου εξοπλισμού
- Η τροποποιημένη έκδοση κάθε υλοποίησης περιέχει μία φάση επιλογής χαρακτηριστικών (βασισμένη στο TFIDF) ούτως ώστε να ελαχιστοποιηθεί ο θόρυβος των δεδομένων, με ό,τι αυτό συνεπάγεται

Naïve Bayes στο Hadoop

- Οι λειτουργίες της μηχανικής μάθησης είναι διακριτές όσον αφορά τα δεδομένα που εκπαίδευσης και ελέγχου που δίνονται σε κάθε κύκλο εργασίας
- Διατηρούνται καθολικοί μετρητές σχετικοί με τα χαρακτηριστικά κειμένου που χρειάζονται στον υπολογισμό των πιθανοτήτων ταξινόμησης κάθε εγγράφου στην φάση του ελέγχου του μοντέλου
- Στον κύκλο της εκπαίδευσης φιλτράρονται και διασπόνται τα έγγραφα εισόδου σε χαρακτηριστικά, για να βρεθεί το πλήθος εμφανίσεων τους για κάθε κλάση για την ρύθμιση του μοντέλου
- Στον κύκλο του ελέγχου φιλτράρονται και διασπόνται αντίστοιχα τα έγγραφα εισόδου σε χαρακτηριστικά για τον υπολογισμό των πιθανοτήτων ταξινόμησης κλάσης και τον ορισμό της ετικέτας κλάσης με την μεγαλύτερη πιθανότητα στο εκάστοτε έγγραφο

Κύκλοι Εργασίας
1. Εκπαίδευση Μοντέλου
2. Έλεγχος Μοντέλου

Τροποποιημένη Έκδοση Naïve Bayes στο Hadoop

- Διατηρούνται και εδώ καθολικοί μετρητές σχετικοί με τα χαρακτηριστικά κειμένου που χρειάζονται στον υπολογισμό των πιθανοτήτων ταξινόμησης κάθε εγγράφου αλλά και για την επιλογή των πιο σχετικών χαρακτηριστικών
- Στους τρεις πρώτους κύκλους φιλτράρεται και διασπάται το κείμενο εισόδου σε χαρακτηριστικά για να βρεθεί η TF συχνότητα τους και στη συνέχεια το μέτρο TFIDF για κάθε χαρακτηριστικό ανά έγγραφο
- Στον κύκλο της επιλογής όρων επιλέγονται τα κορυφαία χαρακτηριστικά βάσει της TFIDF βαθμολογίας τους στο 75% επί του συνόλου και επανασυντάσσονται τα έγγραφα με τους όρους που απέμειναν
- Αντίστοιχη διαδικασία εκπαίδευσης και ελέγχου μοντέλου με την απλή έκδοση

Κύκλοι Εργασίας
1. Καταμέτρηση Όρων
2. Συχνότητα Όρων
3. TFIDF Όρων
4. Επιλογή Όρων
5. Εκπαίδευση Μοντέλου
6. Έλεγχος Μοντέλου

Απλές & Τροποποιημένες Εκδόσεις Naïve Bayes & SVM στο Spark

- Γίνεται χρήση των κατάλληλων εργαλείων από τις βιβλιοθήκες του Spark για τον υπολογισμό του TFIDF μέτρου και τη υλοποίησης των μοντέλων για Naïve Bayes και SVM
- Μοναδική διαφορά μεταξύ εκδόσεων είναι ο ορισμός των **5 εγγράφων ως τον ελάχιστο αριθμό εμφάνισης ενός χαρακτηριστικού για να ληφθεί υπόψην στην κατασκευή του μοντέλου**
- Το TFIDF εξυπηρετεί επίσης την μετατροπή του κειμένου στην χρήσιμη και φιλική κατά τους υπολογισμούς **διανυσματική αναπαράσταση των εγγράφων**

Βήματα Διαδικασίας
1. Φιλτράρισμα Κειμένου
2. Υπολογισμός TFIDF των Όρων
3. Διάσπαση Εγγράφων Εισόδου σε Σύνολα Εκπαίδευσης και Ελέγχου
4. Εκπαίδευση Μοντέλου
5. Έλεγχος Μοντέλου

Πειραματικά Αποτελέσματα και Αξιολόγηση

- Σαν είσοδος δόθηκε μια μεγάλη συλλογή αναρτήσεων από το *Twitter* με **2 κλάσεις πολικότητας συναισθημάτων**, όπου **75%** των tweets δίνονται **για εκπαίδευση του μοντέλου** ενώ τα υπόλοιπα **25%** των tweets **για έλεγχο του μοντέλου**
- Χρησιμοποιήθηκε **συστοιχία τριών (3) κόμβων** από τον διαθέσιμο εξοπλισμό του τμήματος Μηχανικών Πληροφορικής και Υπολογιστών στο Πανεπιστήμιο Δυτικής Αττικής
- Για να κριθεί πλήρης η αξιολόγηση κάθε υλοποίησης, πρέπει να εξεταστεί ξεχωριστά η επίδοση της στο πλαίσιο της κατηγοριοποίησης και της παράλληλης εκτέλεσης σε κατανεμημένο περιβάλλον

```
13,0,Sentiment140,      this weekend has sucked so far
14,0,Sentiment140,     jb isnt showing in australia any more!
15,0,Sentiment140,     ok thats it you win.
16,0,Sentiment140,    &amp;lt;-------- This is the way i feel right now...
17,0,Sentiment140,"    awhhe man.... I'm completely useless rt now. Funny, all I
18,1,Sentiment140,    Feeling strangely fine. Now I'm gonna go listen to some Se
19,0,Sentiment140,    HUGE roll of thunder just now...SO scary!!!!
20,0,Sentiment140,    I just cut my beard off. It's only been growing for well o
```

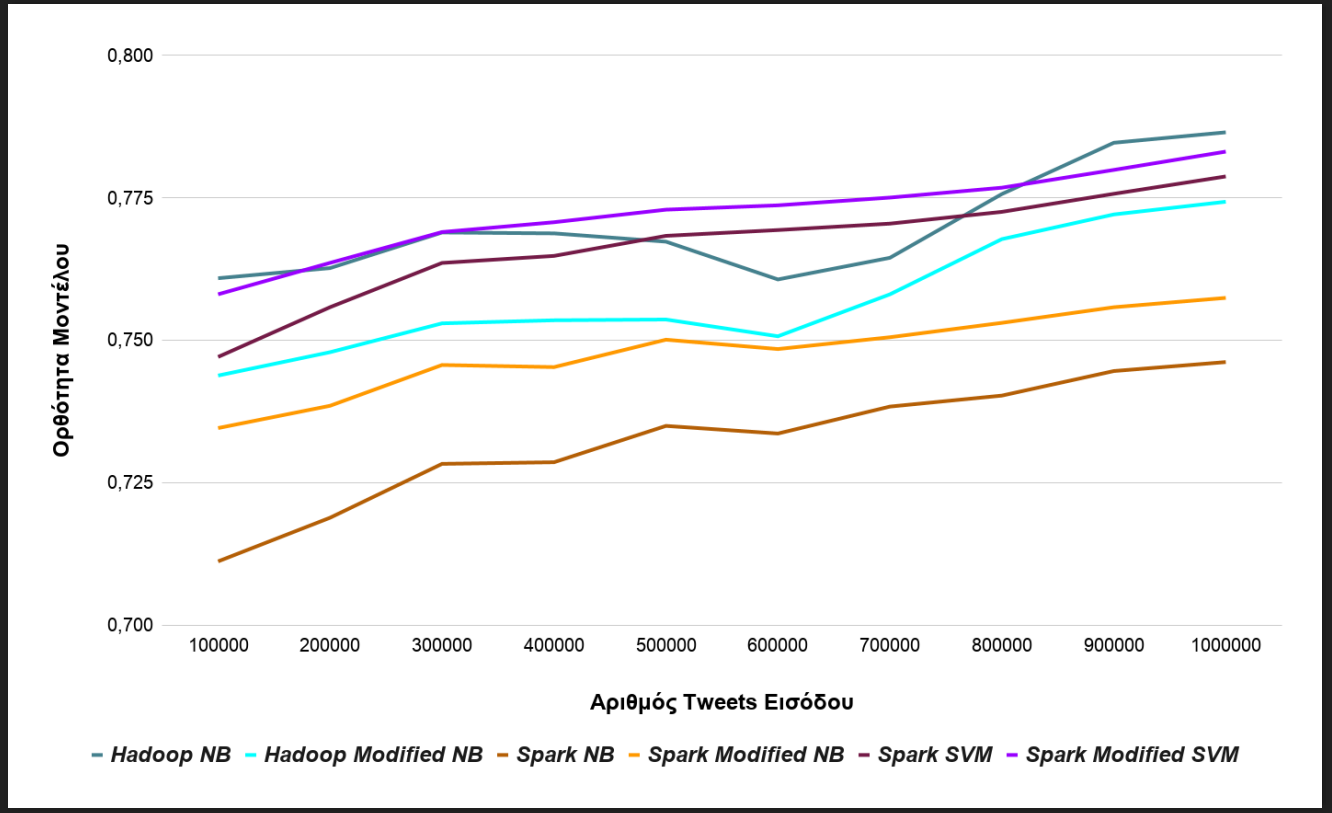
Μορφή Δεδομένων Εισόδου

Χαρακτηριστικό	Εξοπλισμός Κόμβου
Επεξεργαστής	Intel i3-4160, 3.6GHz
Πυρήνες	2
Νήματα	4
Κύρια Μνήμη	8GB
Αποθ. Χώρος	250GB

Χαρακτηριστικά Κόμβων Συστοιχίας Πειραματικής Αξιολόγηση

Ορθότητα (Accuracy)

- Από τα στοιχεία που προέκυψαν από την μήτρα σύγχυσης κάθε εκτέλεσης, μπορεί σε πρώτο χρόνο κάθε υλοποίηση να αξιολογηθεί κατά το ποσοστό των σωστών κατηγοριοποιήσεων κάθε μοντέλου στο σύνολο των προβλέψεων
- Παρόλα αυτά η ορθότητα ενδέχεται να μην είναι το καταλληλότερο μέτρο αξιολόγησης σε περιπτώσεις που δεν εγγυάται από κανέναν ότι όλες οι κλάσεις έχουν (περίπου) τον ίδιο αριθμό στιγμιοτύπων στην συλλογή δεδομένων που εξετάζεται



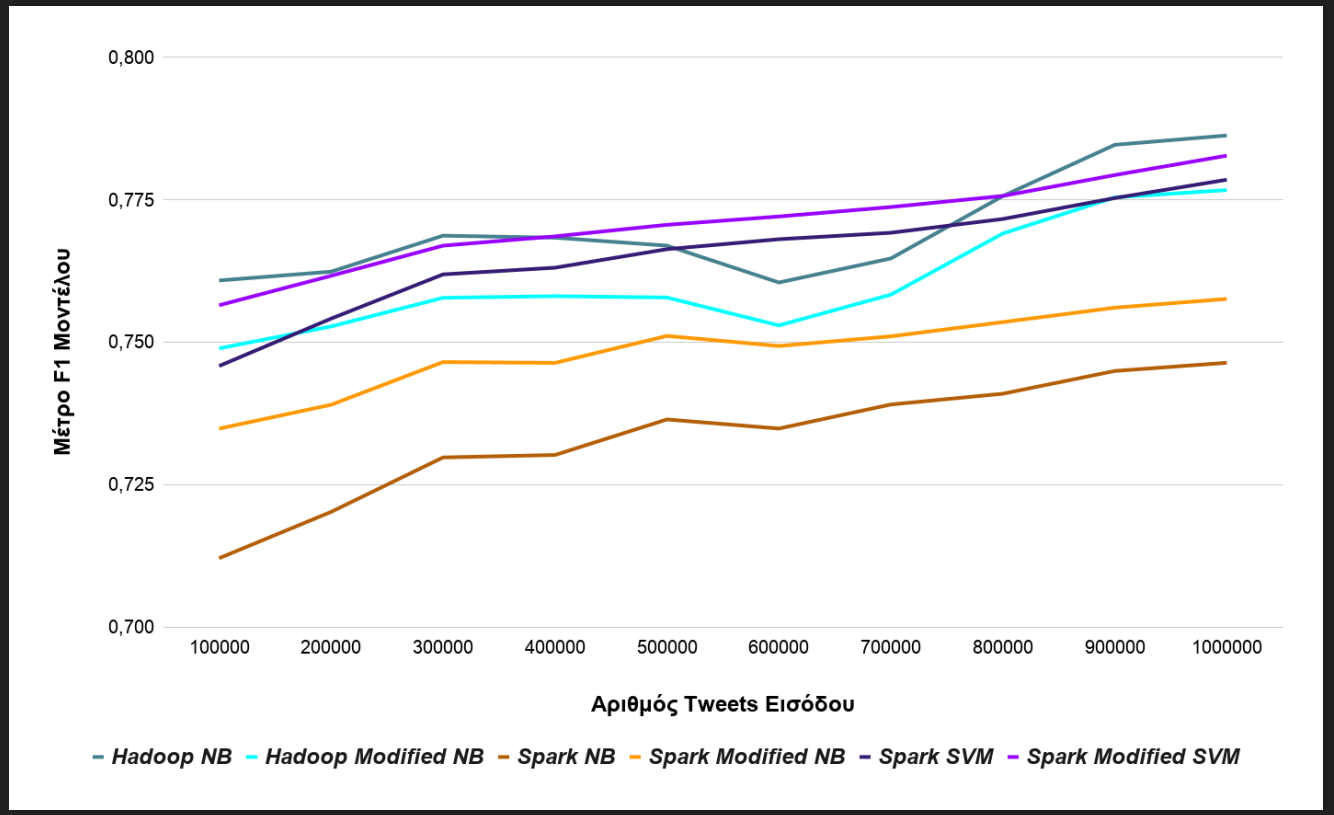
Γράφημα Γραμμών Ορθότητας Κατηγοριοποίησης Δειγμάτων Κάθε Υλοποίησης Ανά Αριθμό Δειγμάτων Εισόδου

$$A = \frac{TP + TN}{TP + FP + FN + TN}$$

Σχέση Υπολογισμού Ορθότητας A (Accuracy) Μοντέλου

Μέτρο F1 (F1 Measure)

- Για την μελέτη μοντέλων που δεν έχουν ισόποσα στιγμιότυπα μεταξύ κλάσεων χρησιμοποιούνται οι μετρικές της **ακρίβειας** (precision) και της **ανάκλησης** (recall) που ορίζονται για κάθε κλάση ξεχωριστά
- Υπολογίστηκαν αρχικά οι ποσότητες ακρίβειας και ανάκλησης για κάθε κλάση ξεχωριστά και έπειτα σταθμίστηκαν μεταξύ τους για την συνολική ακρίβεια και ανάκληση κάθε μοντέλου
- Εν τέλει με τις σταθμισμένες τιμές των μετρικών προσδιορίζεται η ποσότητα του Μέτρου F1 για κάθε υλοποίηση



Γράφημα Γραμμών Μέτρου F1 Κατηγοριοποίησης Δειγμάτων Κάθε Υλοποίησης Ανά Αριθμό Δειγμάτων Εισόδου

$$P = \frac{TP}{TP + FP}$$

Σχέση Υπολογισμού Ακρίβειας P (Precision) Μοντέλου για Θετική Κλάση

$$R = \frac{TP}{TP + FN}$$

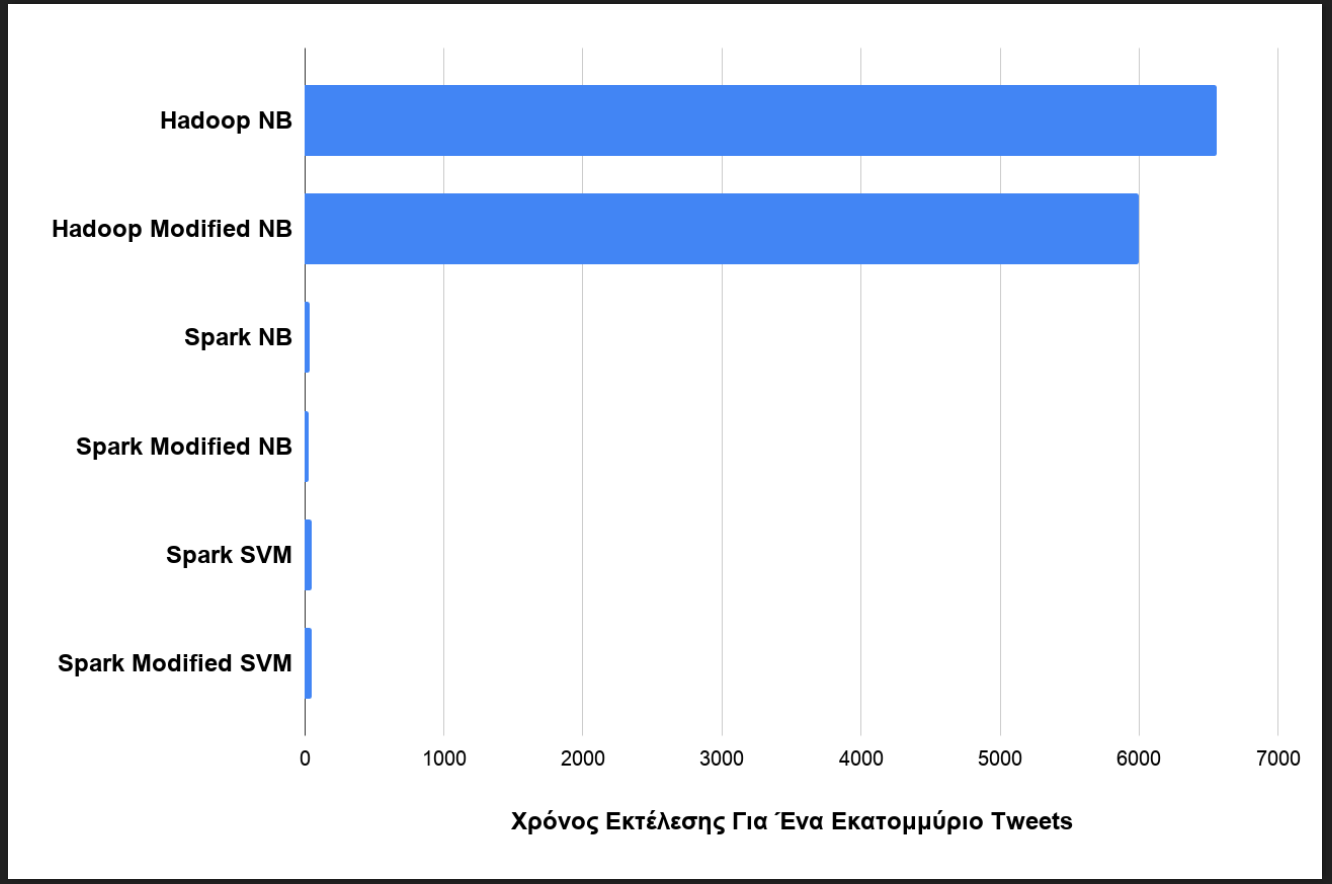
Σχέση Υπολογισμού Ανάκλησης R (Recall) Μοντέλου για Θετική Κλάση

$$F_1 = \frac{2PR}{P + R}$$

Σχέση Υπολογισμού μέτρου F1 (F1 measure) Μοντέλου

Χρόνος Εκτέλεσης

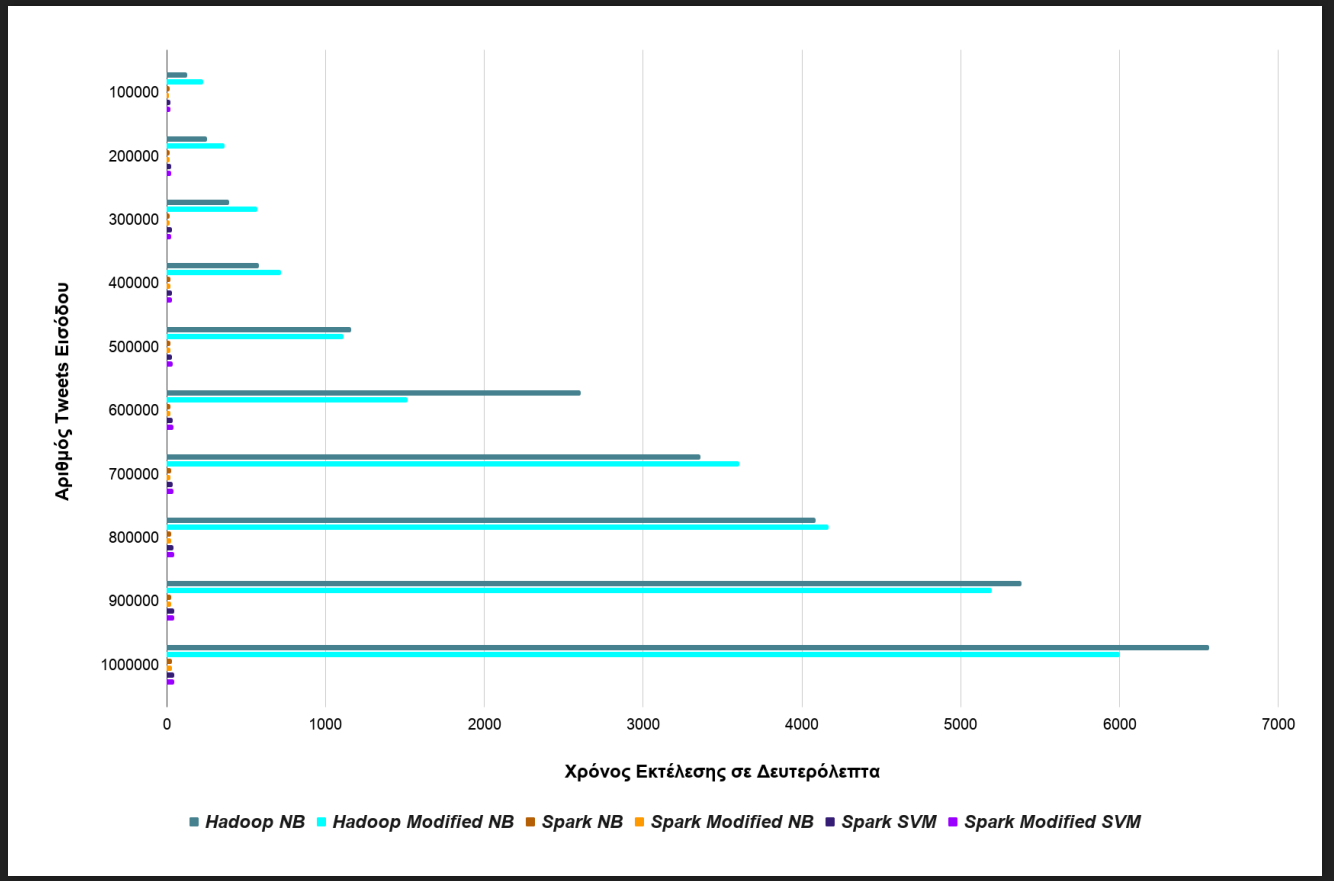
- Λήφθησαν τρεις (3) μετρήσεις για λογαριασμό κάθε υλοποίησης και υπολογίστηκε ο μέσος όρος του χρόνου εκτέλεσης που παρουσιάστηκε
- Χρησιμοποιήθηκαν και τα τρία (3) από τους διαθέσιμους υπολογιστές της συστοιχίας σε όλες τις μετρήσεις
- Μετρήθηκε ο χρόνος που απαιτείται κάθε φορά από την σάρωση των δεδομένων εισόδου ως την ολοκλήρωση της εκτέλεσης κάθε υλοποίησης



Γράφημα Ράβδων Χρόνου Εκτέλεσης (σε δευτερόλεπτα) Κάθε Υλοποίησης Για Ένα Εκατομμύριο Δείγματα Εισόδου

Κλιμακωσιμότητα (Scalability)

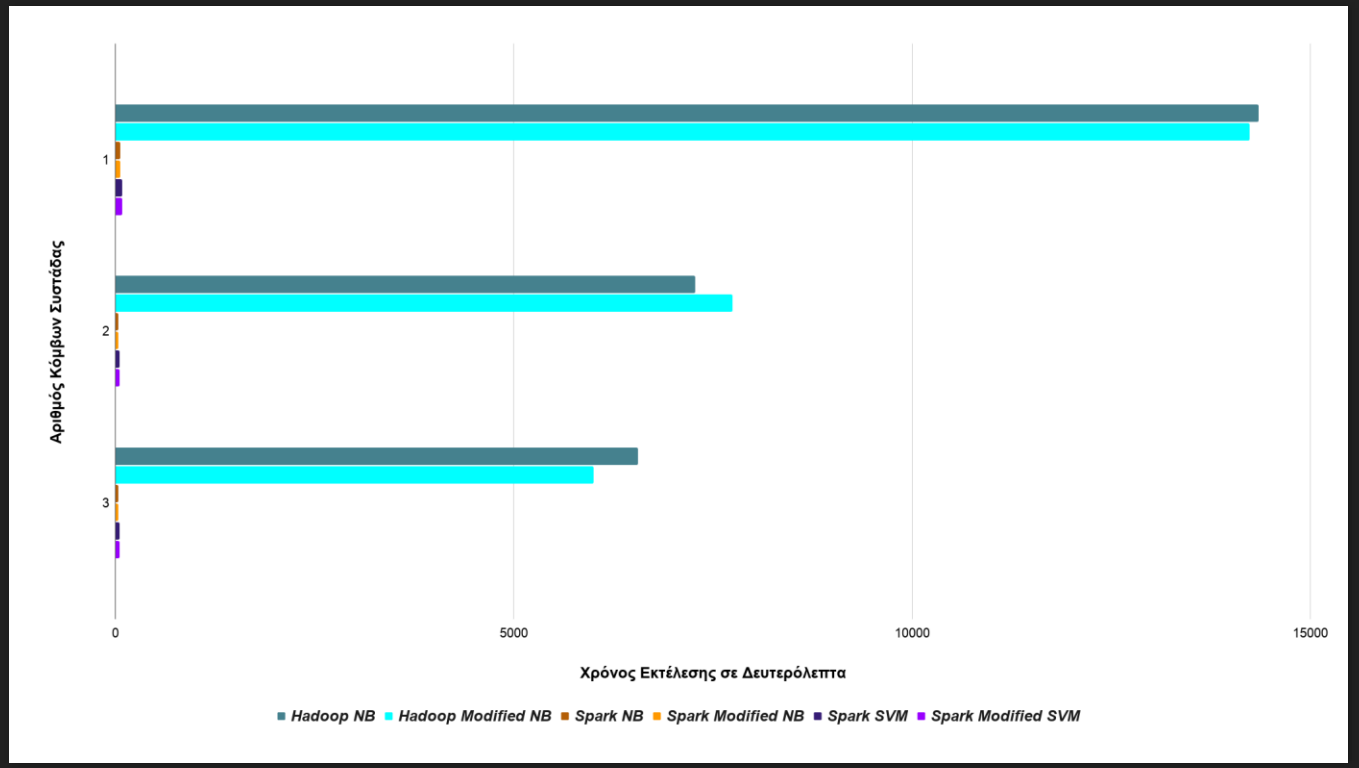
- Λήφθησαν τρεις (3) μετρήσεις για λογαριασμό κάθε υλοποίησης και συνόλου εισόδου και υπολογίστηκε ο μέσος όρος του χρόνου εκτέλεσης που παρουσιάστηκε
- Δοκιμάστηκαν εκτελέσεις όλων των υλοποιήσεων για όλο και μεγαλύτερα σύνολα δεδομένων εισόδου, από 100.000 έως 1.000.000 δείγματα
- Χρησιμοποιήθηκαν και τα τρία (3) από τους διαθέσιμους υπολογιστές της συστοιχίας σε όλες τις μετρήσεις
- Μετρήθηκε ο χρόνος που απαιτείται κάθε φορά από την σάρωση των δεδομένων εισόδου ως την ολοκλήρωση της εκτέλεσης κάθε υλοποίησης



Γράφημα Ράβδων Χρόνου Εκτέλεσης (σε δευτερόλεπτα) Κάθε Υλοποίησης Ανά Αριθμό Δειγμάτων Εισόδου

Επιτάχυνση (Speedup)

- Λήφθησαν και εδώ τρεις (3) μετρήσεις για λογαριασμό κάθε υλοποίησης και μεγέθους συστοιχίας και υπολογίστηκε ο μέσος όρος του χρόνου εκτέλεσης που παρουσιάστηκε
- Δοκιμάστηκαν εκτελέσεις για όλα τα δυνατά μεγέθη συστοιχίας, από έναν και μοναδικό κόμβο ως τους τρεις (3) που ήταν συνολικά διαθέσιμοι
- Μετρήθηκε ο χρόνος που απαιτείται κάθε φορά από την σάρωση των δεδομένων εισόδου ως την ολοκλήρωση της εκτέλεσης κάθε υλοποίησης



Γράφημα Ράβδων Χρόνου Εκτέλεσης (σε δευτερόλεπτα) Κάθε Υλοποίησης Ανά Αριθμό Δειγμάτων Εισόδου

Συμπεράσματα

- Στην περίπτωση που υπάρχει η ανάγκη της μέγιστης δυνατής αποτελεσματικότητας κατηγοριοποίησης, μπορεί να επιλεγεί μία από τις υλοποιήσεις του SVM στο Spark
- Στην περίπτωση που δίνεται βάρος στην βέλτιστη παράλληλη εκτέλεση ενός ικανοποιητικού μοντέλου κατηγοριοποίησης, προτείνεται η τροποποιημένη έκδοση του SVM στο Spark
- Μια καλή επίδοση στους χρόνους εκτέλεσης μιας υλοποίησης δεν συνεπάγεται πάντα από καλή επίδοση και στην ταξινόμηση των δειγμάτων, και το αντίστροφο
- Τα αποτελέσματα είναι όσο καλά είναι και τα δεδομένα εισόδου

Επεκτάσεις

- Επανεξέταση τροποποίησης υλοποιήσεων για την βελτιστοποίηση της επίδοσης ορθότητας στην ταξινόμηση των εγγράφων, χρησιμοποιώντας άλλο τρόπο αξιολόγησης περί των πιο σημαντικών χαρακτηριστικών
- Περαιτέρω μελέτη της συμπεριφοράς και των αποτελεσμάτων των υλοποιήσεων με χρήση περισσότερων κόμβων στο πειραματικό περιβάλλον της συστοιχίας
- Εφαρμογή δεδομένων εισόδου στην φάση του ελέγχου που συγκεντρώονται σε πραγματικό χρόνο από κάποια εξωτερική πηγή (με τη συνδρομή επεκτάσεων όπως το *Apache Flume* ή συνιστωσών όπως το *Spark Streaming*)

Τέλος

Ερωτήσεις;