Ψευδοκώδικας:

Βήμα #1:

Από το Twitter,παίρνω n tweets,n usernames,n λίστες από followers και n λίστες από friends.

Βήμα #2

Ι.Κατασκευάζω m λίστες με tweets και m usernames,όπου m≤n.

ΙΙ.Για καθεμία από τις m λιστες των tweets,υπολογίζω x μετρικές και y features

III.Βάζω σε μια λίστα τις x μετρικές και τα y features

IV. Στελνω προς τα έξω το username μαζί με μια λίστα με τις μετρικές+features.

Βήμα #3

Κατασκευάζω ένα map,στο οποίο βάζω τα k τελευταία username-list of metrics&features που υπολόγιστηκαν.

Μετατρέπω για καθένα από τα k entries,την λίστα σε διάνυσμα.

Βήμα #4: Εφαρμογή του K-Means

I.Από τα k datapoints,επιλέγω τυχαία κ datapoints,που θα είναι τα κ αρχικά centroids.

II.Εφαρμόζω τον αλγόριθμο k-means:

Για καθ'ένα από τα k data points:

a.Βρίσκω σε ποιο από τα κ centroid,το data point βρίσκεται πιο κοντά

b. Αποθηκεύω στο centroid εκείνο το διάνυσμα του data point

c.Υπολογίζω τις νέες τιμές του κάθε centroid.

ΙΙΙ.Ανάλογα με το index του κάθε cluster,δημιουργώ 2 hashmaps,το πρώτο με cluster index=0 θα έχει k1 στοιχεία και το δεύτερο με cluster index=1 θα έχει k2 στοιχεία.

Βήμα #5: Εφαρμογή του Gaussian Mixture Model -Gaussian Rank

Ι.Παίρνω τα k datapoints:

A. Εφαρμόζω τον αλγόριθμο k-means στα k data points και χωρίζω το σύνολο των k data points σε κ clusters. (Έστω το κάθε cluster αποτελείται από λ data points).

B.Σε καθ'ένα από τα κ clusters εφαρμόζω τον αλγόριθμο ΕΜ.

Γ.Βρίσκω το GMM gia καθ'ενα από τα κ clusters.

Δ.Για καθένα από τα κ clusters που προκύπτουν:

Έστω κάθε cluster αποτελείται από λ datapoints:

I.Βρίσκω την μέση τιμή και την standard deviation τιμή του κάθε στοιχείου των διανυσμάτων των λ datapoints.

II.Για κάθε ένα από τα λ datapoints,υπολογίζω την gaussian rank ως προς κάποιο στοιχείο του διανύσματος.

Βήμα #6: Fusion Lists

I.Στο data set( σύνολο των k data points) εφαρμόζω τον αλγόριθμο fuzzy means και χωρίζω τα k data points σε κ clusters.

II.Για καθ'ένα από τα κ clusters: Έστω ότι κάθε cluster αποτελείται από λ datapoints.Τότε:

Α.Αποθηκεύω καθ'ένα από τα λ datapoints σε μια λίστα

Β.Για κάθ'ένα από τα λ data points:

a.Βρίσκω την 'θέση' του κάθε data point μέσα στην λίστα.

b. Υπολογίζω την borda τιμή του κάθε datapoint στο σύνολο των λίστων.

Βήμα #5: Εφαρμογή του Gaussian Mixture Model-Gaussian Rank (NEW)

Ξεχωριστά για κάθε μια από τις 2 ομάδες που δημιουρήγηθηκαν από τον k-means:

1.Υπολογίζω το means,το standard deviation του κάθε feature .

2.Υπολογίζω τυχαία τα mixture weights του κάθε feature.

3. Υπολογίζω την αρχικό gaussian distribution του κάθε feature

4. Eκτελώ τον αλγόριθμο ΕΜ για κάθε feature:

(Την ιδέα για τον ΕΜ την πήρα από εδώ: http://perso.telecom- paristech.fr/~gfort/Preprints/lecorff:fort:moulines:LVAICA2012.pdf)

a.E-STEP: Υπολογίζω το gaussian posterior probability και κανονικοποιώ τις τιμές των πιθανοτήτων αυτών.

b.M-STEP: Υπολογίζω τις νέες τιμές των gaussian distribution,means,covariance,mixture weights για καθ'ένα feature.

5.Με βάση τις τιμές των features , means,covariance και mixture weights,υπολογίζω τα CDF και weighted CDF.

6.Για κάθε feature ξεχωριστά,υπολογίζω τα rank values,και αποθηκεύω το username σε λίστα ή σε hashmap(username+index).

Βήμα #6: Fuzzy C Means (NEW)

I. Από τα k datapoints,επιλέγω τυχαία C datapoints,που θα είναι τα C αρχικά centroids.

II.Για καθ'ένα από τα k data points:

a.Βρίσκω σε ποιο από τα C centroid,το data point βρίσκεται πιο κοντά. Για την απόσταση datapoint-centroid χρησιμοποιώ την σχέση:

b. Αποθηκεύω στο centroid εκείνο το διάνυσμα του data point

c.Υπολογίζω τις νέες τιμές του κάθε centroid.

ΙΙΙ.Ανάλογα με το index του κάθε cluster,δημιουργώ C(εδώ θεωρώ C=3) hashmaps,το πρώτο με cluster index=0 θα έχει C1 στοιχεία ,το δεύτερο με cluster index=1 θα έχει C2 στοιχεία και το τρίτο cluster με cluster index=2 με C3 στοιχεία.

Βήμα #7:Fusion Lists

Βήμα #8: Δημιουργώντας ένα database και υπολογίζοντας ένα relativity value (IDEA)

I.Δημιουργώ ένα database στο οποίο εκχωρώ τα tweets,τα αντίστοιχα usernames, timestamps ποσταρίσματος των tweets και λίστες των friends και followers.

II.Δημιουργώ δύο αρχεία όπου:

A.στο πρώτο,αποθηκεύω,τα usernames των twitterers

B.στο δεύτερό,αποθηκεύω τα tweets με τα αντίστοιχο username του twitterer, (σε μορφή hashmap username-tweet).Αν κάποιο username εμφανιστεί παραπάνω από μια φορά τότε αποθηκέυω τα usernames με τις αντίστοιχες λίστες των tweets(σε μορφή hashmap username-list of tweets).

III.

Αξιολόγηση Αλγορίθμων:

Τρόπος #1:

Αρχικά ταξινομώ τους χρήστες ανάλογα με τις απαντήσεις/tweets που έχουν δώσει και επιλέγονται έτσι οι τόπ 10 χρήστες.

Για την ακρίβεια,παίρνω τους αλγόριθμους και ανάλογα με το αποτέλεσμα που δίνει και την διάταξη των 10 τοπ χρηστών διαπιστώνω' πόσο ακριβείς είναι .

Στην συνέχεια τεστάρω πόσο καλοί είναι χρησιμοποιώντας τα άλλα 3-4 μοντέλα