

## Αποτελέσματα και συγκρίσεις

Παρακάτω παρουσιάζονται ενδεικτικά αποτελέσματα για διάφορους συνδιασμούς των αλγορίθμων συσταδοποίησης που υλοποιήθηκαν. Για την πλειοψηφία αυτών έχει χρησιμοποιηθεί η cosine μετρική καθώς είναι αρκετά πιο γρήγορη και έτσι διευκολύνονταν οι πειραματισμοί εξοικονομώντας χρόνο. Χρησιμοποιείται το twitter\_dataset\_small\_v2.csv

### Σύγκριση K-means Random και K-means++ initialization

| Euclidean I1A1U1 | Silhouette | Time    |
|------------------|------------|---------|
| k=4              | 0.0756523  | 5.20625 |
| k=10             | 0.0918686  | 10.7704 |
| k=15             | 0.0923296  | 18.9669 |
| k=20             | 0.174054   | 29.0892 |

| Cosine I1A1U1 | Silhouette | Time     |
|---------------|------------|----------|
| k=4           | 0.045536   | 0.960846 |
| k=10          | 0.0821437  | 2.58122  |
| k=15          | 0.128915   | 9.3551   |
| k=20          | 0.101497   | 9.44794  |

| Euclidean I2A1U1 | Silhouette | Time    |
|------------------|------------|---------|
| k=4              | 0.106346   | 3.06907 |
| k=10             | 0.0876862  | 16.9732 |
| k=15             | 0.0737217  | 25.1635 |
| k=20             | 0.0871114  | 45.7858 |

| Cosine I2A1U1 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0591045  | 1.77367 |
| k=10          | 0.0829732  | 4.03818 |

|      |           |         |
|------|-----------|---------|
| k=15 | 0.0858545 | 7.31536 |
| k=20 | 0.109943  | 13.2259 |

Όταν χρησιμοποιούμε τη random μέθοδο αρχικοποίησης τα κέντρα αρχικοποιούνται τυχαία οπότε τα αποτελέσματα παρουσιάζουν αρκετές αποκλίσεις και διαφοροποιήσεις ανάλογα με την εκτέλεση. Στην kmeans++ μέθοδο παρατηρούμε μια πιο ομοιόμορφη συμπεριφορά.

### **K-Means with Range Search LSH/CUBE**

| Cosine I2A2U1 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0595637  | 10.3628 |
| k=10          | 0.0879154  | 23.6646 |
| k=15          | 0.0909024  | 92.7514 |
| k=20          | 0.0938697  | 170.211 |

| Cosine I2A3U1 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0442006  | 6.78578 |
| k=8           | 0.0863154  | 13.8281 |
| k=15          | 0.0903255  | 38.5556 |
| k=20          | 0.100007   | 27.6177 |

Και οι δυο μέθοδοι απαιτούν περισσότερο χρόνο σε σχέση με το απλό k-means και θα είχαν μεγαλύτερο νόημα αν εφαρμόζονταν σε μεγάλα δεδομένα όπου και θα απέδιδαν καλύτερα. Σε γενικές γραμμές το cube φαίνεται πιο γρήγορο από το lsh.

### **K-medoids (PAM)**

| Cosine I2A1U2 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0664144  | 90.1732 |
| k=10          | 0.053678   | 35.5957 |
| k=15          | 0.0773082  | 16.8098 |

|      |           |         |
|------|-----------|---------|
| k=20 | 0.0662138 | 14.0084 |
|------|-----------|---------|

Ο pam συγκλίνει σε πιο λίγες επαναλήψεις σε σχέση με το k-means αλλά κάθε επανάληψη του αλγορίθμου απαιτεί αρκετό χρόνο.

### **PAM and Range Search**

| Cosine I2A2U2 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0399554  | 73.9766 |
| k=10          | 0.0774405  | 37.8583 |
| k=15          | 0.0815724  | 29.6546 |
| k=20          | 0.0748826  | 26.5943 |

| Cosine I2A3U2 | Silhouette | Time    |
|---------------|------------|---------|
| k=4           | 0.0617354  | 74.1029 |
| k=10          | 0.051018   | 34.3211 |
| k=15          | 0.0605227  | 26.9565 |
| k=20          | 0.078303   | 35.9199 |