

Nama : Usama
NPM : 1606862766

A. Clustering

Untuk mengelompokkan *tweet* berdasarkan *cluster*-nya akan menggunakan K-means, oleh karena menggunakan K-means harus ditentukan centroid awalnya, *centroid* awal sudah ditentukan di **centroid_awal.txt**. Oleh karena itu, untuk iterasi clustering pertama gunakan *centroid* yang sudah ditentukan.

Update *centroid* pada setiap iterasi *clustering*. Algoritma yang digunakan :

1. Setiap *tweet* di *cluster* k akan menjadi kandidat *centroid* baru untuk *cluster* k.
2. Hitung nilai *similarity* menggunakan *jaccard similarity* pada setiap *tweet* yang menjadi kandidat.
3. Jumlahkan setiap *similarity* yang didapatkan.
4. Hitung rata-rata dari *similarity*.
5. Gunakan *tweet* yang memiliki rata-rata *similarity* terbesar menjadi *centroid* baru.

Keluaran dapat dilihat di **1606862766_output_a.txt**

B. SentimentAnalysis

Untuk setiap *tweet* kita telah menyimpan *tweet_id* dan *cluster*-nya, oleh karena itu kita hanya perlu menghitung sentimen untuk setiap *tweet*. Kemudian masukkan sentimen dari *tweet* tersebut ke dalam daftar *cluster* yang sesuai.

Cara menghitung sentimen :

$$polarity\ sentiment = \frac{\#positif + \#negatif}{\#opinion\ words}$$

untuk menghitung apakah sebuah *tweet* memiliki sentimen negatif atau positif, kita hanya perlu menghitung penjumlahan dari $\#positif + \#negatif$, jika $\#positif + \#negatif$ menghasilkan nilai negatif, maka akan menjadi sentimen negatif dan begitupun sebaliknya. Karena nilai $\#opinion\ words$ bernilai positif, maka nilai negatif dan positif dari *polarity sentiment* hanya ditentukan oleh nilai pembilangnya.

Keluaran dapat dilihat di **1606862766_output_b.txt**