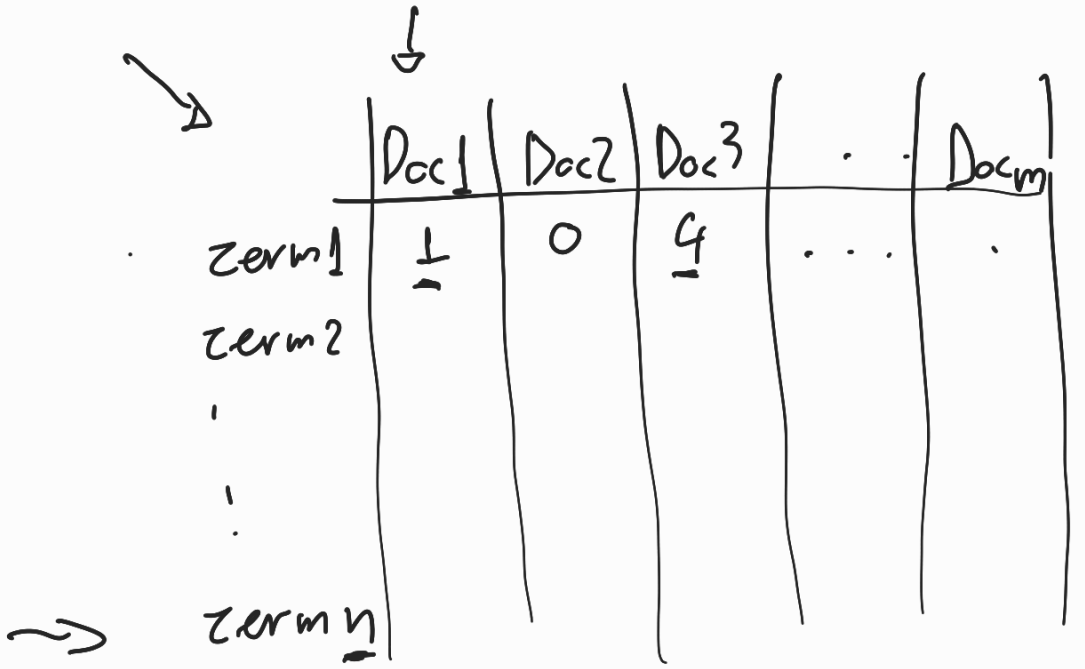


Προεργασία

- Αφαίρεση ανούσιων λέξεων (stop words) κοινών
- Εύρεση ριζών (stemming)

$q = term_{i_1} + term_{i_2} + \dots term_{i_m}$ → (0, 1, 0, 1, ...)

1: στη θέση αυτή

→ {Doc1, ... DocK}

$\cos(x, y)$ δείχνει την ομοιότητα x, y

$\cos(\theta(q, a_i)) > tol$

↓

i στήλη του A

$(A = U \Sigma V^T)$

↑

αραιός

$q^T a_i$

$\cos(\theta(q, a_i)) = \frac{a_i^T q}{\|a_i\|_2 \|q\|_2}$

$$\approx U_K \Sigma_K V_K^T = \underbrace{U_K H_K}_{H_K}$$

$$A \approx U_K H_K \qquad a_i \approx U_K h_i$$

$a_i \approx U_K \cdot h_i$ όπου h_i η i στήλη του H_K

$\vec{q}_i \approx U_K \cdot \vec{h}_i$

$$q^T A_K = \left(q^T (U_K H_K)\right)^T = \underbrace{(U_K^T q)}_{\vec{q}_K} \cdot H_K$$

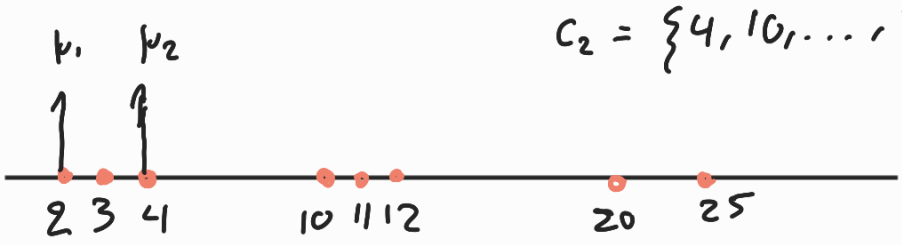
$$\cos\theta_i = \frac{(U_K^T q) \cdot h_i}{\|U_K^T q\|_2 \|h_i\|_2} = \frac{q_K^T \cdot h_i}{\|q_K\| \cdot \|h_i\|}$$

Συσταδοποίηση (Clustering) βασισμένη σε αντιπροσώπους

$K=2$

$C_1 = \{2,3\}$

$C_2 = \{4, 10, \dots, 25\}$



Έστω η σημεία $\{\vec{x}_1, \vec{x}_2, \dots \vec{x}_n\} \in \mathbb{R}^d$

$\downarrow \quad \downarrow$

$\begin{pmatrix} x_1^1 \\ x_2^1 \\ x_i^1 \end{pmatrix}$

Πλήθος συστάδων. Έστω K τότε $\{C_1, C_2, \dots C_K\}$

Κάθε συστάδα C_i υπάρχει ένας αντιπρόσωπος. Ένας συνηθισμένη δείκτης είναι ο μέσος όλων των σημείων της C_i που ονομάζεται κέντρο βάρους (centroid)

$$\underline{\mu_i} = \frac{1}{n_i} \sum_{j \in C_i} x_j = \frac{1}{|C_i|} \sum_{j \in C_i} x_j$$

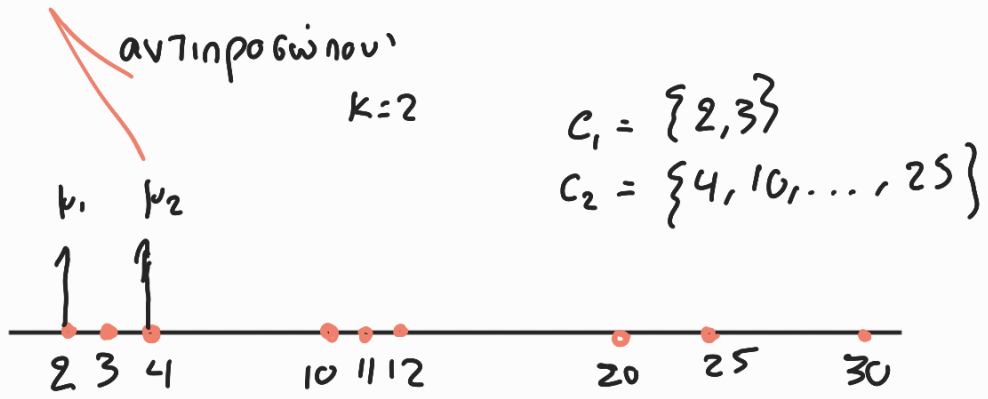
Αλγόριθμος K μέσων

$\mathcal{C} = \{C_1, C_2, \dots, C_K\}$

Η αξιολόγηση της συσταδοποίησης θα γίνει μέσα από την μετρική

$$\sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

Στόχος μας είναι να βρούμε $\mathcal{C}$ που ελαχιστοποιεί την παραπάνω μετρική



$C_1 = \{2, 3\}$

$C_2 = \{4, 10, \dots, 25\}$

2ο βήμα: Βρίσκω τους νέους μέσους όρους

$\mu_1^2 = \frac{2+3}{2} = 2.5$

$\mu_2^2 = \frac{4+10+11+12+20+25+30}{7} = \frac{112}{7} = 16$

$C_1^{\mu_1^2} = \{2, 3, 4\}$ || $C_2^{\mu_2^2} = \{10, 11 \dots 30\}$

$\mu_1^3 = 3$

$\mu_2^2 = \frac{10+11+12+20+25+30}{6} = \frac{108}{6} = 18$

Η διαδικασία συνεχίζεται μέχρι ???

$\{\mu_1^{(n)}\} = \{\mu_1^1, \mu_1^2, \mu_1^3 \dots\}$

$\{\mu_2^{(n)}\} = \{\mu_2^1, \mu_2^2, \dots\}$

$$\sum_{i=1}^{K} \| \mu_i^m - \mu_i^{m-1} \|^2 \leq \varepsilon$$