چیت‌شیت یک‌صفحه‌ای — Big Data

پاسخ = تعریف + فرمول + مثال. شماره صفحات = جزوه کامل همین نسخه.

مفهوم	یک خط
Shingle	دنباله k توکن متوالی
Jaccard	$\lvert A\cap B\rvert / \lvert A\cup B\rvert$
MinHash	Pr[هش برابر] = Jaccard
LSH	فقط candidateهای مشابه
Collab. Filtering	توصیه با Similarity کاربر/آیتم
Closed / Maximal	$M \subseteq C \subseteq F$

$$TF_{ij} = f_{ij} / \max_k(f_{kj})$$
$$IDF_i = \log_2(N / n_i)$$
$$TF\text{-}IDF = TF \times IDF$$

مثال استاد: $0.1 \times 3 = 0.3$

$$J = \lvert A\cap B\rvert / \lvert A\cup B\rvert$$
$$Pr[h_\pi(C1)=h_\pi(C2)] = J$$
$$P = 1 - (1 - s^r)^b$$

$$L_2 = \sqrt{\Sigma(x_i-y_i)^2}$$
$$L_1 = \Sigma\lvert x_i-y_i\rvert \quad L_\infty = \max\lvert\ldots\rvert$$
$$\text{Nominal: } (p - m) / p$$

$$\text{Map: } (i\ ,\ m_{ij} \times v_j)$$
$$\text{Reduce: } u_i = \Sigma\ (m_{ij} \times v_j)$$

مثال: U = [32, 26, 44]

$$sup(A\Rightarrow B) = sup(A\cup B)$$
$$conf = sup(A\cup B) / sup(A)$$

موضوع	صفحه جزوه
فهرست صفحات	1
چیت‌شیت داخل جزوه	1
Technical Glossary	2
TF-IDF / Collaborative	6
Cluster / Hadoop	8
MapReduce	11
Matrix × Vector	13
Shingle / MinHash / LSH	16
Distances	18
Apriori / Closed	20
Clustering	24
حل نمونه سوالات	**28**