Ex: for cluster validity.

Assume we have a test collection D of 900 documents from three topics (or 3 classes) Science, Sports and Politics. Each class has 300 documents. Each document in D is labelled with one of the topic (class). These documents are grouped into 3 clusters. Measure the effectiveness of the clustering.

Cluster	Science	Sports	Politics	→	Total
1	250	20	10		280
2	20	180	80		280
3	30	100	210		340
Total	300	300	300		

Solution:

Step 1: Calculate the total in each cluster.

cluster 1 → 250 + 20 + 10 = 280

Step 2: Find out the probability of each cluster.

Cluster	Science	Sports	Politics	Purity
1	0.893	0.0714	0.035	0.893
2	0.0714	0.643	0.286	0.643
3	0.0882	0.2941	0.6176	0.617
Total	300	300	300	0.711

C1: Prob (Science) = 250/280 = 0.893, Prob (Politics) = 0.0357

Prob (Sports) = 20/280 = 0.0714

Step 3: Calculation of Purity, by considering the Maximum probability. = Max (Prob).

C1 : Max (0.893, 0.0714, 0.035) = 0.893

Similarly for all other clusters.

Step 4:
Purity of the clustering = $\sum_{i=1}^{3} \frac{m_i}{m}$ Purity.

$$= \left[\frac{280}{900} \times 0.893 + \frac{280}{900} \times 0.643 + \frac{340}{900} \times 0.617\right]$$

$$= 0.277 + 0.2000 + 0.2330$$

$$= 0.711$$

Step 5: Calculate the Entropy of each cluster.

$$C1 : -\left(\frac{250}{280}\log_2\frac{250}{280} + \frac{20}{280}\log_2\frac{20}{280} + \frac{10}{280}\log_2\frac{10}{280}\right)$$

$$= -(0.893\log_2 0.893 + 0.0714\log_2 0.0714 + 0.035 + \log_2 0.035)$$

$$= -(-0.14579 + 0.271886 - 0.1692775)$$

$$= 0.587$$

Similarly for all the clusters calculate the Entropy.

Prob	Science	Sports	Politics	Entropy	Purity.
1	0.893	0.0714	0.035	0.587	0.893
2	0.0714	0.643	0.286	1.198	0.643
3	0.088	0.294	0.617	1.257	0.617
Total	300	300	300	1.0301	0.711

$$-\left(\frac{20}{280}\log_2\frac{20}{28} + \frac{180}{280}\log_2\frac{180}{280} + \frac{80}{280}\log_2\frac{80}{280}\right)$$

$$-(0.071\log_2 0.071 + 0.642\log_2 0.642 + 0.285\log_2 0.285)$$

$$-(-0.27093 - 0.41046 - 0.51612) = 1.198$$

$$-\left(\frac{30}{340}\log_2\frac{30}{340} + \frac{100}{340}\log_2\frac{100}{340} + \frac{210}{340}\log_2\frac{210}{340}\right)$$

$$-(0.0882\log_2 0.0882 + 0.294\log_2 0.294 + 0.617\log_2 0.617)$$

$$= -(-0.30897 - 0.5192 - 0.4298)$$

$$= 1.257$$

Step 6: Calculate the entropy of clustering.

$$= \sum_{i=1}^{k} \frac{m_i}{m} e_i$$

$$= \frac{280}{900} \times 0.587 + \frac{280}{900} \times 1.198 + \frac{340}{900} \times 1.257$$

$$= 0.1826 + 0.3727 + 0.4748$$

$$= 1.0301$$