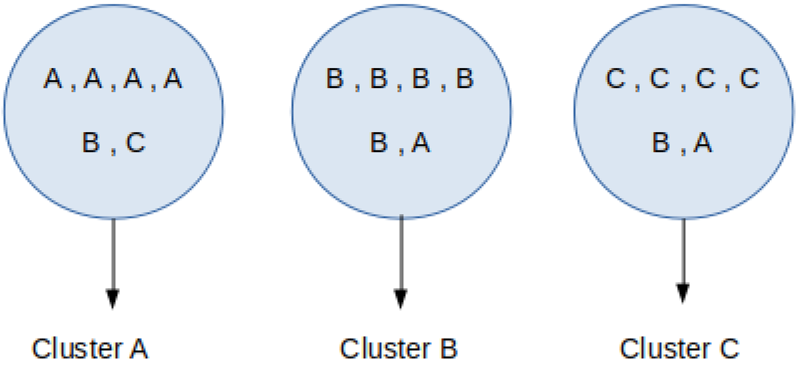
CS 620–Introduction to Data Science and Analytics, HW5

1) **(40 points)** Consider the following 3 clusters.
   a. (10 pts) Calculate the Purity
   b. Create the contingency table (confusion matrix) and using the contingency table,
      i. (20 pts) Calculate the Rand index
      ii. (5 pts) Calculate the Precision and Recall
      iii. (5 pts) Calculate the Balanced F measure



2) **(60 points)**
   a. (20 pts) The following list of R's and N's represents relevant (R) and non-relevant (N) documents in a ranked list of 50 documents. The top of the ranked list is on the left of the list, so that represents the most highly weighted document, the one that the system believes is most likely to be relevant. The list runs across the page to the right. This list shows 10 relevant documents. Assume that there are only 10 relevant documents for this query.

      R,R,N,R,N,R,N,N,N,N,R,N,N,N,R,N,N,N,N,R,R,N,N,N,N,N,N,N,R,N,N,N,N,N,N,N,N,N,N,N,R,N,N,N,N,N,N,N,N,N

      Based on that list, calculate the following measures:
      i. (15 pts) Precision, Recall and F Measure at each position.
      ii. (5 pts) Average Precision of the ranked list

   b. (20 pts) Now, imagine another system retrieves the following ranked list for the same query.

      R,N,N,R,N,N,N,R,N,N,N,N,N,R,N,N,N,N,N,N,N,R,N,N,N,N,N,N,R,N,N,N,R,N,N,N,N,R,N,N,N,N,N,N,R,N,N,N,N,R

      Based on that list, calculate the following measures:
      i. (15 pts) Precision, Recall and F Measure at each position.
      ii. (5 pts) Average Precision of the ranked list

   c. (15 pts) Interpolation defines precision at any recall level as the maximum precision observed in any recall-precision point at a higher recall level. Calculate the Interpolated Precision for each standard recall values (0.0,0.1….1.0) and generate the Recall-Precision graph for the ranking lists in part (a) and (b).

   d. (5 pts) What do the graphs tell you about the system in (a) and the one in (b), i.e., if you were given only these evaluation metrics (Average Precision, Precision-Recall curve) what can you determine about the relative performance of the two systems in general.

What to turn in: PDF should contain the following information at the top. **Lastname-hw5.pdf**
CS620
HW5
@author:

Submit your pdf file to Blackboard. Due: Monday, April 25, 2022, 11.59pm