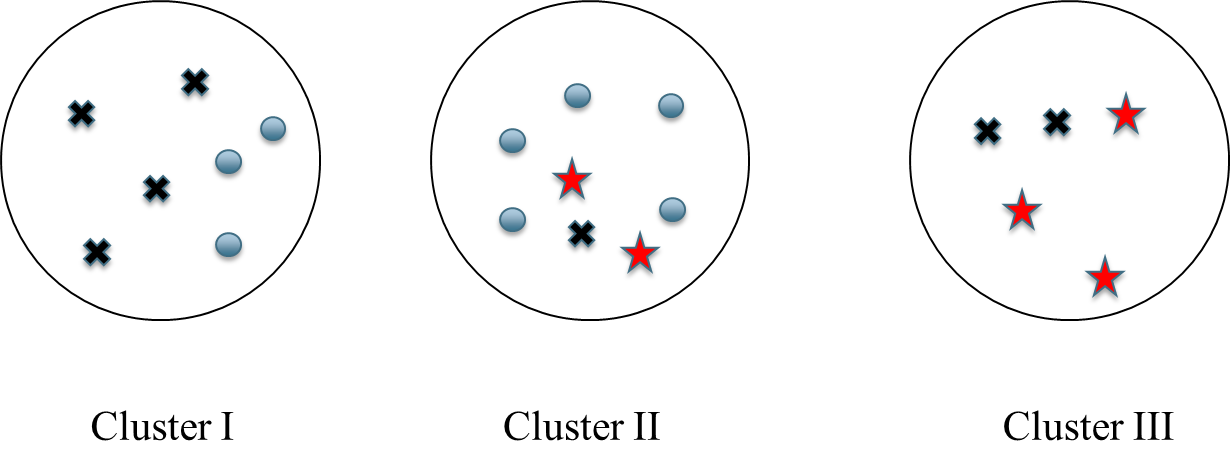
1. (20%) Naïve Bayes text classification

Based on the data in the following table, use (i)Bernoulli model (ii) multinomial model to predict class label of test set. (Use Laplacian correction to avoid overfitting)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Doc id | Words in document | Class label |
| Training set | 1 | Hualien Taiwan | Y |
| 2 | Hualien Hualien Taiwan Taiwan | Y |
| 3 | Sapporo Shanghai | Y |
| 4 | Hualien Sapporo Japan Japan | N |
| 5 | Sapporo Japan Taiwan | N |
| Test set | 6 | Hualien Taiwan Taiwan | ? |

1. (35%) Consider a web graph with four nodes 1, 2, 3 and 4. The links are as follows: 1->2, 1->3, 2->1, 2->4, 4->1, 4->2, 4->3. (i) write the transition probability matrices for the surfer’s walk with teleport probability 0.2. (ii) assume the initial probability vector the four nodes is <1, 0, 0, 0>. What is the probability vector at step 2. (iii) devise an algorithm for computing the static state probabilities of the four nodes.
2. (20%) Assume we cluster the data points into three groups (cluster I, cluster II, and cluster III). Compute the rand index.



1. (25%) Seed selection is an important job in K-means algorithm. Bad seeds can result in poor convergence rate or convergence to sub-optimal clusters. Please propose 2 methods to improve the seeds quality. Furthermore, use example to illustrate the idea you proposed.