

Indexation: (3pt)

Following is portion of positional index in the form (Word : docId<position1,position2,.....>):

- Computer: 2<1,15,26,38>, 3<15,152,169>, 5<1,18,26,35,59,156,256>
 - Science: 2<2,16,27>, 5<19,36,157,260>
 - Department: 2<3,28,150,269>, 5<2,20,22,158,261>, 6<15,26,59,68>
 - Faculty: 2<17,39,200>, 5<8,11,37>, 7<5,19,256,350,456>
1. Describe the set of documents (with positions) that satisfy the query “Computer science department”.
 2. Describe the set of documents (with positions) that satisfy the query “Computer science faculty”.

Vector Space Model: (3pt)

Consider the following collection of documents:

- document1 = ‘one two ‘
 - document2 = ‘three two four ‘
 - document3 = ‘one two three ‘
 - document4 = ‘one two ‘
1. Calculate the IDF of each term
 2. Calculate similarity between document 3 and the query “one three three” using NTC.LNN smart notation.

Decision Tree: (8pt)

As of September 2012, 800 planets have been identified in our galaxy. Many spaceships sent to all these planets have established whether they are habitable for humans or not, but sending a spaceship to each planet is expensive. In this problem, you will come up with decision trees to predict if a planet is habitable based only on features observable using telescopes.

In nearby table you are given the data from all 800 planets surveyed so far. The features observed by telescope are Size (“Big” or “Small”), and Orbit (“Near” or “Far”). Each row indicates the values of the features and habitability, and how many times that set of values was observed. So, for example, there were 20 “Big” planets “Near” their star that were habitable.

Size	Orbit	Habitable	Count
Big	Near	Yes	20
Big	Far	Yes	170
Small	Near	Yes	139
Small	Far	Yes	45
Big	Near	No	130
Big	Far	No	30
Small	Near	No	11
Small	Far	No	255

Draw the decision tree?

Clustering: (6pt)

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters:

$A1=(2,10)$, $A2=(2,5)$, $A3=(8,4)$, $A4=(5,8)$, $A5=(7,5)$, $A6=(6,4)$, $A7=(1,2)$, $A8=(4,9)$.

The distance matrix based on the Euclidean distance is $d(a,b)=\sqrt{(x_b-x_a)^2+(y_b-y_a)^2}$.

Suppose that the initial seeds (centers of each cluster) are A3, A5 and A8. Run the k-means algorithm for only one iteration. At the end of this iteration show:

1. The new clusters (with the examples belonging to each cluster)
2. The centers of the new clusters
3. Draw a 10 by 10 space with all the 8 points and show the clusters after the first iteration and the new centroids.