Question 1

If for the χ^2 statistics for a binary feature, we obtain $P(\chi^2 | DF = 1) < 0.05$, this means:

a. **That the class labels depends on the feature**
b. That the class label is independent of the feature
c. That the class label correlates with the feature
d. No conclusion can be drawn

Question 2

Given a document collection, if we change the ordering of the words in the documents, which of the following will not change?

a. **Singular values in LSI**
b. The entities extracted using HMM
c. The embedding vectors produced by Word2vec
d. All the previous will change

Question 3

We want to return, from the two posting lists below, the top-2 documents matching a query using Fagin's algorithm with the aggregation function taken as the sum of the tf-idf weights. How many entries (total of both lists) are accessed in the first phase of the algorithm performing round robin starting at List 1 (i.e., before performing the random access)?

List 1			List 2	
document	tf-idf		document	tf-idf
d3	0.8		d1	0.8
d2	0.6		d3	0.6
d1	0.5		d4	0.5
d4	0.4		d2	0.4

a. 3
b. 4
c. 5
d. 6

Question 4

In the physical representation of an inverted file, the size of the index file is typically in the order of (where n is the number of documents):

a. O(log(n))
b. O(sqrt(n))
c. O(n)
d. O(n^2)

Question 5

Which is **true** about the use of entropy in decision tree induction?

a. The entropy of the set of class labels of the samples from the training set at the leaf

level is always 0

b. We split on the attribute that has the highest entropy
c. **The entropy of the set of class labels of the samples from the training set at the leaf level can be 1**
d. We split on the attribute that has the lowest entropy

Question 6

Given the *2-itemsets* {1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 5}, when generating the *3-itemset* we will:

a. Have **4** *3-itemsets* after the join and **4** *3-itemsets* after the prune
b. **Have 4 *3-itemsets* after the join and 2 *3-itemsets* after the prune**
c. Have **3** *3-itemsets* after the join and **3** *3-itemsets* after the prune
d. Have **2** *3-itemsets* after the join and **2** *3-itemsets* after the prune