

[illegible]

Part I : String/Text Processing**[43 marks in total]**

1. Describe two (or more) steps that we would typically perform in the Tokenisation process for an Information Retrieval collection, according to the discussion in this subject. [4 marks]
 - Folding case — making everything lowercase
 - Stripping punctuation — removing some non-alphabetic characters like “,”
 - Stemming — removing suffixes (in English) to change a word to a base form
 - Splitting into tokens — splitting the document based on whitespace (in English) and maybe some punctuation
 - And other possible answers

2. It has been claimed that there are three primary types of “information need” in a web search context: “informational”, “navigational”, and “transactional”. Briefly describe each, optionally with the aid of an example. [3 marks]
 - Informational: tell me more about this topic, e.g. **history of Australia**
 - Navigational: take me to the URL corresponding to this topic, e.g. **Unimelb homepage**
 - Transactional: interface with a database, so that I can perform some service (like buying a product), e.g. **iphone ebay**

3. In the context of Information Retrieval:

(a) Explain how Data retrieval is different to Information Retrieval. [2 marks]

- Data retrieval: getting some variable value out of memory, or a record out of a database, etc.
- Information retrieval: trying to find some document(s) which meet the user's information need expressed by the query
- Information retrieval doesn't have an exact answer; whether the results are useful depends on the user issuing the query

(b) Give an example of a method or source of information that we might incorporate in our engine, that is specific to Web-scale Information Retrieval. [1 marks]

- Link analysis
- Click-through data
- And other possible answers

4. ...And more questions to add up to the marks stated above. (-:

Part II: Data Mining/Machine Learning**[42 marks in total]**

For these questions, we have a training dataset comprised of the following 6 instances, 3 attributes, and two classes F and T, and a single test instance labelled with ?:

ele	fed	aust	CLASS
1	1	1	F
1	0	0	F
1	1	0	T
1	1	0	T
1	1	1	T
1	1	1	T
0	0	0	?

5. Classify the given test instance using the method of Naive Bayes, as described in this subject. [4 marks]

- We need to pre-calculate a bunch of probabilities: $P(f) = \frac{2}{6}$, $P(t) = \frac{4}{6}$; $P(l = 0|f) = 0$, $P(l = 0|t) = 0$, $P(e = 0|f) = \frac{1}{2}$, $P(e = 0|t) = 0$, $P(u = 0|f) = \frac{1}{2}$, $P(u = 0|t) = \frac{1}{2}$,
- When we substitute in, we need to replace 0 values with ϵ , a small positive constant value.
- We calculate the scores for the two classes F and T:

$$\begin{aligned}
 \text{F} &: P(f)P(l = 0|f)P(e = 0|f)P(u = 0|f) \\
 &= \frac{1}{3}(\epsilon)(\frac{1}{2})(\frac{1}{2}) = \frac{\epsilon}{12} \\
 \text{T} &: P(t)P(l = 0|t)P(e = 0|t)P(u = 0|t) \\
 &= \frac{2}{3}(\epsilon)(\epsilon)(\frac{1}{2}) = \frac{\epsilon^2}{3}
 \end{aligned}$$

- ϵ is less than $\frac{1}{4}$, so F has the larger value — so that is the class we choose.

6. Explain why 1-Nearest Neighbour will give a different prediction to 3-Nearest Neighbour on this test instance. (Note that it is not required to show all of your workings for this question.) [2 marks]

- Regardless of the distance metric we're using, clearly the second instance (1,0,0:F) has the smallest distance; so, 1-NN will say F.
- The next best instance(s) are (1,1,0:T), of which there are two.
- So, for 3-NN, we will observe 2 T instances and 1 F instance among the 3 nearest neighbours; there are more T than F, so we classify it as T.
- (Since the question doesn't ask for working, it is possible to explain this more compactly.)

7. Consider the method of Random Forests:

(a) Briefly explain how a Random Forest would be constructed on the training data above. [4 marks]

- For a Random Forest, we will construct a bunch of “Random Trees”, in this case, let’s say 10 of them.
- For each tree, we will use Bagging to come up with a different training dataset: we will re-sample the instances with replacement, until we have 6 (possibly repeated) training instances.
- When building our tree, at each node, we only consider a proportion of the attributes. Because we have so few attributes, let’s say: we randomly choose 2 of the 3 attributes for consideration at the root node; we consider both of the remaining attributes at the second layer; we consider the final attribute at the third layer.

(b) Is there any evidence that a Random Forest would label the given test distance differently to a regular Decision Tree? [3 marks]

- (Aside: there will be some more difficult questions like this one. If you need to think about the problem, the harder questions might take longer to answer than the marks suggest!)
- Probably yes:
- The regular decision tree will have **e** at the root — it is clearly the most useful attribute — and therefore classify the test instance as **F**.
- When bagging, the chance of any individual instance being present in the training data is about 63%. (The lectures say $\frac{2}{3}$.) If the second instance isn’t present, we are going to say **T** pretty much no matter what.
- Even if the second instance is present, $\frac{1}{3}$ of the time, **e** won’t be in the options for the root — therefore **u** will be placed at the root (**1** is useless). If we’ve bagged more of the 1,0,0:**F** instances than 1,1,0:**T** instances, we’ll say **F**, but this will be very uncommon, given that there are twice as many **T** instances with **u**=0.
- To a rough approximation: $37\% + 63\% \frac{1}{3} = 58\%$ of the trees will choose **T**; this is more than half, so probably the Random Forest will choose **T**.

8. Exclude the CLASS labels from the dataset, and cluster all 7 instances using the method of k -means. Apply the Manhattan Distance as a similarity measure; use the second (1,0,0) and third (1,1,0) instances as seeds. [4 marks]

- Let's say Cluster 1 C_1 begins at 1,0,0 and Cluster 2 C_2 begins at 1,1,0.
- For each instance, we calculate the Manhattan distance to the two clusters. I will show the workings for one instance; it is obviously crazy to try to write the whole formula 14 times in 5.6 minutes.
 - First instance to C_1 : $|1 - 1| + |1 - 0| + |1 - 0| = 2$; to C_2 : $|1 - 1| + |1 - 1| + |1 - 0| = 1$.
 - Second instance to C_1 : 0; to C_2 : 1.
 - Third instance to C_1 : 1; to C_2 : 0.
 - Fourth instance is the same as third instance; fifth and sixth instances are the same as first instance.
 - Seventh instance to C_1 : 1; to C_2 : 2.
- So, the first, third, fourth, fifth, and sixth instances are closer to C_2 ; the second and seventh are closer to C_1 . We now update our centroids:

$$C_1 : \frac{1}{2}[(1, 0, 0) + (0, 0, 0)] = (0.5, 0, 0)$$

$$C_2 : \frac{1}{5}[(1, 1, 1) + (1, 1, 0) + (1, 1, 0) + (1, 1, 1) + (1, 1, 1)] = (1, 1, 0.6)$$

- Now, we re-calculate the Manhattan distances:
 - First instance to C_1 : $|1 - 0.5| + |1 - 0| + |1 - 0| = 2.5$; to C_2 : $|1 - 1| + |1 - 1| + |1 - 0.6| = 0.4$.
 - Second instance to C_1 : 0.5; to C_2 : 1.6.
 - Third instance to C_1 : 1.5; to C_2 : 0.6.
 - Fourth instance is the same as third instance; fifth and sixth instances are the same as first instance.
 - Seventh instance to C_1 : 0.5; to C_2 : 2.6.
- So, the first, third, fourth, fifth, and sixth instances are closer to C_2 ; the second and seventh are closer to C_1 . This is the same as the previous iteration, so this is the clustering.

9. ...And more questions to add up to the marks stated above. :-)