

IR Coursework (2)

Dell Zhang
DCSIS, Birkbeck, University of London

Spring 2012

This is part 2 of the coursework (worth 8 marks) covering the topics of Dell's lectures. The total coursework, together with Sven's part, will be worth 20 marks.

1. (4 marks)

Suppose that the following labelled document collection is used to train a Naive Bayes classifier for two classes — UK and US.

<i>Class</i>	<i>Documents</i>
UK	d_1 : "London, England."
UK	d_2 : "York, England."
US	d_3 : "New England."
US	d_4 : "Newark, New Jersey."

- (a) Compute the probability $P(\text{UK})$.
- (b) Compute the probabilities $P(\text{New}|\text{UK})$ and $P(\text{York}|\text{UK})$ with Add-One (Laplace) smoothing.
- (c) Compute the probability $P(d_5|\text{UK})$ for the test document d_5 : "York. New York."
- (d) Which class should document d_5 be classified into? Why?

2. (4 marks)

Suppose that the pair-wise similarity information of a document collection $\{d_1, d_2, d_3, d_4, d_5, d_6\}$ is given by the following table.

d_1						
d_2	0.7					
d_3	0.4	0.8				
d_4	0.2	0.1	0.5			
d_5	0.0	0.1	0.0	0.6		
d_6	0.1	0.4	0.2	0.3	0.4	
	d_1	d_2	d_3	d_4	d_5	d_6

- (a) If it is known that $\{d_1, d_2, d_3\}$ are *Facts* and $\{d_4, d_5\}$ are *Myths*, how will document d_6 be classified by the k NN algorithm with $k = 3$, $k = 4$, and $k = 5$ respectively? Please use the *standard* k NN algorithm but not its weighted variant.
- (b) Use the Single-Link Hierarchical Agglomerative Clustering (HAC) algorithm to cluster these documents, and draw the generated dendrogram.