



## Question 2

Assume we have the query “garbage recycle” and are considering two documents which mention these words. The word counts are shown in table below – N is the total number of words in a document.

	Document 1	Document 2
N	398	491
Count of “garbage”	35	15
Count of “recycle”	13	14
Count of “garbage recycle”	3	4

- a) Calculate  $P(Q|D_i, r)$  for the given query for each document using unigram and bigram models built on each document and applied to the query. Q means the query,  $D_i$  means document with index i, r means relevance = true.

- b) Assume that in 200,000 previous queries, document 2 was determined to be relevant by users 8 times and document 1 was relevant 4 times. Calculate the un-normalised odds ratio of the relevance to irrelevance of the given query for each document using the formula below. Which document seems more relevant to the query under each of the unigram and bigram models?

$$\frac{P(r | D, Q)}{P(\neg r | D, Q)} = \frac{P(Q | D, r)P(r | D)}{\alpha(1 - P(r | D))}$$

### Question 3

The following two paragraphs are taken from Assignment 2 and each will be considered a 'document'.

Paragraph 1:

Create an Ensemble of Classifiers: Make multiple copies of your classifier, and train these copies on random subsets of your training data. Produce an overall prediction in response to a test pattern by combining the results from multiple classifiers via a voting system. See Russell and Norvig, section 18.10 on p. 748-752 for some background.

Paragraph 2:

Overtraining Prevention: There are a number of techniques that may be used to ensure the classifier is not overtraining. One option is to partition your training set into two new sets (a training set, and a validation or verification set) and to use this validation set to estimate when to stop training the classifier. Again, the choice is up to you.

A query "train classifier" is received. Use a unigram model of the document and query to determine which document would be ranked as the most relevant. Show working.

Ignore punctuation marks. Decide and describe how you treat numbers. Where relevant to determining which document best matches the query, use stemming (reduction to base words) via simple, programmable rules and list these rules. Assume that both paragraphs are on average equally relevant to a general range of queries.