COMP4650/COMP6490 Document Analysis
2025 Semester 2

Computing Lab 2

Q1: Information Retrieval

Part I

(This is a theory question that does not require coding.)

Suppose that we are evaluating our IR system. For a given query, our system retrieves 10 documents, which are marked as being relevant (`R`) or irrelevant (`I`) in the following order:
`R, I, R, R, I, R, R, R, R, R`
The list is ordered left to right, so the leftmost `R` is the relevance of the first retrieved document. There are 12 relevant documents in the entire collection.

(a) Calculate the recall at 5 documents retrieved.

(b) Calculate the interpolated precision at 20% recall.

(c) Calculate the F1-score at 5 documents retrieved.

(d) Consider the task of building an IR system for a collection of legal documents (patents, court transcripts, etc) to be used by legal firms. How would you evaluate this IR system? Compare the differences in user needs between this system and a typical web search application. How would these differences influence what metrics you use to measure your system performance?

Part II

(This is a practice question that requires coding.)

In the notebook `lab2-inverted_index.ipynb` you will implement a simple indexer to construct inverted index from raw text. Work through the notebook and answer the questions in it.

Q2: Text Classification

Part I

(This is a theory question that does not require coding.)

The binary logistic regression classifier that we have learned in class is a type of *linear* classification models, where we must represent our input objects as a vectors and the classification decision is based on a score that is a linear function of the input vector.

Show how a binary logistic regression classifier can be viewed as a linear classifier.

Part II

(This is a practice question that requires coding.)

In this part, you will build a text classification model on a provided movie review dataset. The dataset consists of 50,000 review articles written on movies in IMDb, each review is labelled with the sentiment – either positive or negative. Your text classification model will be able to infer the sentiment of a review from its text. The main goal of this question is for you to gain familiarity with the `scikit-learn`[1] machine learning package and its application to text data.

One simple approach to classifying the sentiment of documents from their text is to train a logistic regression classifier using bag of words features. This approach is relatively straightforward to implement and can be very hard to beat in practice.

You have been provided with a notebook `lab2-sparse_linear_classifier.ipynb` which loads the movie reviews and splits the data into training, validation, and testing sets. Your task is to apply logistic regression to the movie review dataset, to predict the sentiment label from the review text. To do this you will need to write a function `fit_model` that takes a set of training sentences as input, tokenizes the sentences, calculates TF vectors and then trains a logistic regression model.

(*HINT:* `CountVectorizer`*, and* `LogisticRegression` *in the* `scikit-learn` *package will be helpful for this. You should use them after reading the documentation*).

You should also implement `test_model` as it will be useful in the next part. Using `fit_model`, `test_model`, and your training and validation sets you should then search over possible values for the regularisation parameter C. It is suggested that you try $C = 3^k$ where k is an element of $\{-5, -4, ..., 0, 1, ..., 5\}$, and select based on accuracy. Though other choices are possible, you might want to test them out. Next, re-train your classifier using the training set concatenated with the validation set and your best C value. Evaluate the performance of your model on the test set.

Answer the following questions:

(a) What was the best performing C value?

(b) What was your final accuracy? (The accuracy of the reference solution is around 0.87)

(c) Look at the co-efficients of the logistic regression classifier (i.e. the `coef_` attribute of the logistic regression object), what words do the 5 largest (most positive) and 5 smallest (most negative) co-efficients correspond to? (List as words not their token ids, `CountVectorizer` has a `vocabluary_` attribute that may help.)

Briefly explain why these words do or do not make sense.

[1] `https://scikit-learn.org/`