COMP9313 Project2 Report

By z5244467(Chen Wu)

Task2

1. F1 score: 0.7483312619309965
2. In order to reduce runtime, I use cache() to store the read data sets of train and test in memory, which is convenient for retrieval. In each filter, cache() is also performed, and the filtered data does not need to be filtered again, which reduces the running time from 190s to 110s. Then more data processing can be done. According to data in the development data, there are considerable punctuation marks. When using tokenizer, these punctuation marks will also be output, which may affect performance. Therefore, it is possible to improve performance by removing punctuation. After searching source, RegexpTokenizer can be utilized to remove punctuation. After using RegexpTokenizer, the F1 score has a little change, which is 0.7483314623513457.