Answers to Q2

A2.

1. After tokenizing the text in the corpus, it was found that the most commonly occurring tokens are commas, full stops, other punctuation symbols and stop words like ‘a’, ‘and’, ‘the’, etc.
2. Since these most commonly occurring words add no contextual value in analyzing similarity of patent documents, they must be cleaned from our corpus to obtain a set of relevant keywords for further analysis.
3. The pairwise cosine similarity has been calculated and dumped in a csv file named - **cosine\_similarity\_scores.csv**.

Additional note: Please refer to the code and output file attached.