一、

網路上雖然有提供一般的 stopwords 提供下載，但是很多時候不同領域的文章 stopwords 會不同，例如：金融相關文章某些詞彙或許就不那麼重要。請設計一個方式，可以把這樣的領域知識考慮進入 stopwords 去除的流程之中。

什麼是 Heap Law 請寫出公式並說明。

為什麼在 Heap Law 的前提下，我們可以知道透過不斷加大 Corpus 無法解決 sparseness of data 的問題？

二、

搜尋 pattree 的過程中，當結果停在 subtree 時，為什麼還需要經過比對其中的 external nodes 才能判定搜尋解果的正確性。

搜尋 pattree 的過程中，若最後結果停在 external nodes，是否可以證明結果一定不存在於 Collection 中？

使用者透過「好吃的午餐」這樣的 query 在 pattree 中成功搜尋到結果，請問若使用者輸入「吃的午」這樣的 query 是否可以搜尋到結果，請說明原因。

三、

Euclidean distance 是常見的一種直觀用來計算兩篇文章差異的方式，請說明為何這樣的計算方式會對長文較有利。

請說明為何使用 Okapi BM25 的公式，其結果容易對短文較有利。(有給公式)

四、

TP、TN、FP、FN 分別代表什麼意思？

ROC Curve 的 x 軸、 y 軸是什麼？

請說明什麼是 ROC Curve？怎麼繪製？

為何 points 越靠近左上角代表 model 表現越好？

五、

現在有一個情況，使用者 query 只有一個字「iphone」，但剛好 collection 中每篇文章都有這個 term，請說明若使用 tf-idf 做為計算 weight 的方式，這會帶來的問題。

呈上，若使用者堅持使用 tf-idf 來做 vector space model 請設計一個解決的辦法、流程。

六、

Good Turing 計算並排名

請算出 Good Turing 中的 Psuedo Count 並說明計算過程