Topic Model

1. 考慮以下的以下的短文和投影片第三頁的自訂字典：
   台灣近年來房地產市場持續火熱，房價不斷攀升，尤其是在北部都會區，已經成為許多年輕人難以負擔的沉重壓力。高昂的房價造成一種普遍現象：年輕族群即使擁有穩定的工作機會，仍然難以存到足夠的錢支付頭期款，更遑論日後的房貸負面對房市如此不友善的環境，許多年輕人只能選擇租屋。此外，商品的持續上漲，也讓年輕人吃不消，然而面對原物料的上漲，業者也很無奈。

   以上文章有涵蓋到哪些主題（複選）

A. 產業
B. 國安
C. 經濟
D. 社會

ANS：A 和 C，因為裡頭包括了如房地產、房價、房市、房租、租金等經濟主題的字詞。也包括了如商品、業者等產業字詞

2. 承上題，若只能選擇一個主題，請問哪一個主題是重點討論的
   A. 產業
   B. 國安
   C. 經濟
   D. 社會

ANS：C，因為經濟主題的字詞出現比較多

3. 考慮一份文件只有一個主題，以下敘述何者正確?
   A. 通常適用於短文章
   B. 可用如 K-means 的分群演算法來達成
   C. 分群的結果可將每一群裡的文件當成屬於同一個主題
   D. 以上皆是

ANS：D，這些都是。

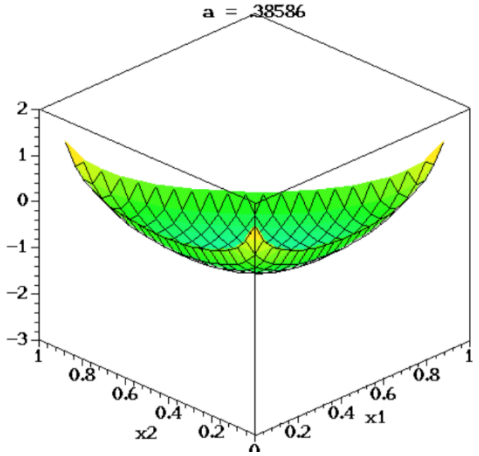
4. 考慮一份文件可有多個主題，以下敘述何者為非?
   A. 如事先訂主題字，可用類似情緒字典法來找出文件的主題
   B. 事先訂主題字的解釋性高但不容易訂
   C. 如事先不訂主題，LDA 是最普遍的方式
   D. LDA 可以找出主題敘述

ANS：D，LDA 可以找出每一個主題的重要字，但不是主題的敘述

5. 下列何者是 LDA 可以輸出的結果（複選）
   A. 每一個主題的字詞分佈
   B. 每一個文件的主題分佈
   C. 最佳主題數
   D. 主題敘述

ANS：A 和 B，主題數是要預先設定，主題敘述或標籤必須看輸出的主題字詞分佈自行推敲而得。

6. 使用 LDA 時必須設定先驗分佈 Dirichlett Distribution 的參數，比如文件-主題的先驗分佈參數α設成 0.38, 假設只有三個主題，則其分佈可以視覺化表示如下



請問以下敘述何者正確?
   A. 代表每一文件通常有多個主題
   B. 主題分佈為(0.3, 0.3, 0.4)的文件出現機率會高於主題分佈為(0.9, 0.05, 0.05)的文件
   C. 代表每一文件通常只有一個主題
   D. 以上皆非

ANS：C，參數α設的小，代表每一文件的主題數很少，所以出現像(0.3, 0.3, 0.4)的主題分佈概率很低。

7. 以下有關 LDA 的敘述何者為非?
   A. LDA 常被用來了解一堆文件是討論哪些重要議題
   B. LDA 可以自動給予每一個主題一個標籤(label)代表主題名稱
   C. 不同於事先訂主題字的方法，LDA 裡每一主題的字詞是有權重的
   D. LDA 是一個有紮實統計理論基礎的主題模型

ANS：B，主題的標籤 LDA 無法給定，只有主題分佈。

8. 以下有關 GuidedLDA 的敘述何者錯誤
   A. 想要的主題可以用給定主題種子字的方式來設定
   B. 會產生文件的主題分佈
   C. 會產生主題的字詞分佈
   D. 僅能產生給定種子字的主題。

ANS：D，如果給定主題數大於有種子字的主題數，則會產生其他主題。