第2題

在 $\sum_{i=1}^{M} p_i = 1, p_i \geq 0$ 的限制之下，$H(X) = -\sum_{i=1}^{M} p_i[\log(p_i)]$ 的最大值出現在機率分佈爲: $p_j = 1.0, 1 \leq j \leq M$ and $p_k = 0, 1 \leq k \leq M, k \neq j$ 的情況之下，此時 $H(X) = 0$。(3 points) 最小值則是在出現在 uniform distribution 時， 即 $p_j = \frac{1}{M}, 1 \leq j \leq M$ 的時候，此時 $H(X)$ 爲最小。(3 points)
因此 $H(X)$ 可以作爲測量一個機率分佈分散的情形，越高則表示越集中， 越小則越分散。(4 points)

第4題

Word 是中文裡的"詞"，像是"考卷"、"學校"。Character 則是中文裡的"字"，像是上例裡的"考"、"卷"、"學"、"校"。(2 points)
用詞做爲 language model 單位的好處有:

- 詞是構成句子的主要單位。
- 詞帶有更多的語意在其中。(1 point for 1 item)

但是壞處有:

- 沒有公認的 lexicon ，所以會影響 language model 的表現。
- 在訓練語料上需要有相當精確的斷字，不易取得。
- 有嚴重的 "OOV" (Out-of-vocabulary) 問題，會有之前沒有見過的新詞。 (1 point for 1 item)

而用字做爲 language model 單位的好處是:

- 不必考慮斷字的問題。
- 不用擔心 OOV 的問題 (1 point for 1 item)

而壞處則是:

- 較不具有語意於其上。
- 和 word 相比，當使用 N-gram language model 需要更多的 N 值才會有好的表現。 (1 point for 1 item)

第5題

Back-off 的做法是在當訓練資料量不夠而產生 data sparseness ，使得我們需要的資訊成爲不可信賴時，我們可以轉而使用次一級的資訊。在 N-gram Language Model 裡若N-gram 爲 unseen event 或語料不足時我們可以轉而向 (N-1)-gram 查詢，並且將所得到的結果給予適當的 weighting 。(5 point)
Interpolation 的目的則是直接視資料量較小的資訊爲需要修正的。在 N-gram Language Model 裡 N-gram 的訓練語料往往不如 (N-1)-gram 來得充份而可靠，因此若同時使用兩者則可能達到更好的效果。因此在計算時便同時考慮兩者，並且使用兩者linear interpolation 來取代 N-gram (5 point)

第8題

和純文字文件不同，多媒體文件在瀏覽上是相當不方便的。人們無法很快得知其 中的資訊，因此需要額外的組織以及整理。
Spoken Document Understanding and Organization 是爲了幫助這樣的文件在 瀏覽上能夠更有效率。
Understanding 包括了以下技術: (1 point for each item listed, another 2 points for description)

- 關鍵字以及類專有名詞的抽取(Key Term/Name Entity extraction for spoken documents).
  找出文章之中的人名，地名，組織名，事件名稱，以及其他關鍵詞例如"投票"、"遊行"等等。
- 文件的分段 (spoken document segmentation).
  把一連串的語音依照其內容分割成段落。
- 資訊萃取(Information Extraction for spoken documents).
  找出文章中所提及的事件的人地時事物等等，以及其中的關係。

Organization 包括了以下技術: (1 point for each item listed, another 2 points for description)

- 自動摘要產生(Automatic Generation of Summaries for spoken documents)
  自動爲文件產生出一份簡短扼要的摘要。
- 自動標題產生(Automatic Generation of Titles for spoken documents)
  自動依文件內容爲文件產生出一個標題。
- 主題分析以及組織(Topic Analysis and Organization for spoken documents)
  分析文件之中的主題，並將主題之間的關係找出並呈現出來。