自然語言理解的解釋性資訊標記競賽：競賽任務與資料說明

任務述敘

本競賽的**每一筆輸入資料為一個三元組**(q, r, s)，q 是一則英文論述，r 是一則對 q 進行回應的英文短文，s 則是 r 對 q 的議論關係，可能是同意（agree）或不同意（disagree）。

輸出資料則是一個雙元組(q', r')，q' 與 r' 分別是 q 與 r 的子序列（subsequence），且 q' 與 r' 提供了關鍵性的資訊，足以判斷 q 與 r 呈現 s 的關係。

下表呈現了一筆範例，q 為一則與槍支管制有關的論述，r 是 q 的一則回應，r 與 q 的關係為不同意。預期的輸出 q' 與 r' 則如 q 與 r 中黃底的片段，提供關鍵性的資訊呈現 r 不同意 q。注意 q' 與 r' 可以是不連續的片段（但不同片段間的先後順序必須與原文之順序相同）。

輸入資料	
q	Originally posted by voiceofreason Well if you're going to quote that article that mentions Canada to support your opinion, then does that mean you're in favor of gun registration and licensing? Canada has them. And if the NRA is really interested in enforcing laws why do they ask congress to reduce the budget of law enforcement? The NRA uses its political muscle to make it easier for criminals to obtain guns. Robert Ricker, former top lawyer of the NRA, talks about it. So if you're really interested in enforcing gun laws, don't support the NRA.
r	What is true, Ricker says, is that gun manufacturers have long known that distributors and retailers supply thousands of guns each year to criminals, and yet gun makers deliberately look the other way. The quote says it all. That attitude breaks many cornerstone laws regarding personal responsibility. Manufactures are NOT responsible for customers illegal activity. GM and Ford know that 100% of all the product they sell will be used illegally. Everyone of their cars will be used to break speed limits. No one obeys speed limts, the only time speed limits are obeyed is when a cop is watching. Yet GM and Ford are not responsible for all the illegal activity of their products. Shifting the responsbility for enforcement breaks many traditions and laws regarding individual personal responsibility. When a customer breaks a law using a product, manufactures are not

	responsible, that goes for both guns and cars, or any product that is sold and used for criminal purposes. There certainly a lack of cooperation between the ATF and the NRA. They really should be working together. It takes two sides to make a conflict. The ATF is trying to single out gun manufactures to be responsible for policing customers. NO OTHER INDUSTRY is being held accountable like that. That is why there is a conflict between the NRA and ATF. The NRA asks that gun manufactures to be treated with the same consideration of existing laws and standards applied to other industries. The NRA supports any efforts to take firearms away from criminals, but they are asking that gun manufactures not be held accountable for customer illegal activity. Which is the position of every single product manufacturer not matter what you make and sell. It is not that gun manufactures are looking away, they do. So does every other manufacture who makes any other product.
s	Disagree
輸出資料	
q´	then does that mean you're in favor of gun registration The NRA uses its political muscle to make it easier for criminals to obtain guns. Robert Ricker, former top lawyer of the NRA, talks about it. if you're really interested in enforcing gun laws, don't support the NRA.
r´	What is true, Ricker says, is, gun manufacturers have long known that distributors and retailers supply thousands of guns each year to criminals, and yet gun makers deliberately look the other way. Manufactures are NOT responsible When a customer breaks a law using a product, manufactures are not responsible trying to single out gun manufactures to be responsible for policing customers. NRA supports any efforts to take firearms away from criminals It is not that gun manufactures are looking away, they do.

競賽資料下載格式

測試資料為一 csv 檔，以無 BOM 檔頭之 UTF-8 編碼，並以半形逗號分隔欄位。除了 id 與 s 兩欄位以外，其餘欄位資料之首尾皆有半形雙引號包夾。如欄位內容中亦有雙引號字元，則以反斜線字元作為脫逸符號。檔案中每一行資料包含四欄，依序為測試資料 **id、*q*、*r*、*s***。其中，測試資料 id 為整數，*q* 與 *r* 為長度不等之字串，*s* 則僅可能為 DISAGREE 或 AGREE。

主辦單位提供的**訓練資料集**，格式類似測試資料，惟每一行中將包含六欄，依序為訓練資料 id、*q*、*r*、*s*、*q*´、*r*´，供建模訓練。訓練資料集內 id 相同者，代表同一組題目可接受的*q*´、*r*´答案可能組合，以利建模運用。

請注意，*q*、*r* 欄位原始內容若存在因編碼差異而出現的亂碼，則競賽資料亦如實呈現。

競賽答案上傳格式

上傳資料應為一 csv 檔，以無 BOM 檔頭之 UTF-8 編碼，並以半形逗號分隔欄位。除了 id 之外，其餘欄位資料之首尾皆有半形雙引號包夾。如欄位內容中亦有雙引號字元，則以反斜線字元作為脫逸符號。**檔案中每一行資料需包含三欄，依序為測試資料 id、*q*´與*r*´**，其中，測試資料 id 為整數，*q*´與*r*´為長度不等之字串，不同片段的字串之間請以 space (空白鍵) 隔開。

上傳資料之內容必須包含欄位名稱 (三個欄位名稱依序為 id, q, r)，且總行數必須與測試資料的總行數一致。上傳檔案格式請至下載區，參考上傳檔案範本「Submission template.csv」。
針對一個測試資料 id，請勿上傳多個答案。請勿上傳未知 id 之內容。

評分預處理

評分**以詞組(token)為計算單位**。評分時會先使用 nltk 套件(3.7 版本)裡的 tokenize.word_tokenize() 函式，分別對*q*´與*r*´進行分詞（word tokenization），並且排除長度為 1 且只有標點符號的詞組。

標點符號的內容為以下字元之一：

```
!"#$%&'()*+, -./:;<=>?@[\]^_`{|}~
```

範例 1：

對字串 `today is my day.` 進行分詞，會得到`['today', 'is', 'my', 'day', '.']`，其中僅有`'.'`為標點符號的詞組，因此最終的分詞結果為`['today', 'is', 'my', 'day']`。

範例 2：

對字串 `It's a question.` 進行分詞，會得到`['It', "'s", 'a', 'question', '.']`，其中僅有`'.'`為標點符號的詞組，`"'s"`不為標點符號的詞組，因此最終的分詞結果為`['It', "'s", 'a', 'question']`。

評分方式

評分以提交上傳之q'與r'與答案q'與r'之間的最長共同子序列（Longest Common Subsequence, LCS）重疊率為依據，詳如下式。

$$Score = \frac{1}{2N}\sum_{i}^{N}\max_{j}\left(\frac{|q_i' \cap \widehat{q_{i,j}}'|}{|q_i' \cup \widehat{q_{i,j}}'|} + \frac{|r_i' \cap \widehat{r_{i,j}}'|}{|r_i' \cup \widehat{r_{i,j}}'|}\right)$$

其中 N 為評分資料筆數，q_i'與r_i'為提交資料裡第 i 筆資料的輸出結果，$\widehat{q_{i,j}}'$與$\widehat{r_{i,j}}'$為第 i 筆資料的答案，由於每筆資料的答案可能有若干組，因此以 j 表示其中一組。

上傳之結果將與每一組答案分別計分，取最高分者作為第 i 筆資料所得之成績。q 與 r 長度之單位為詞組（token），以 q 的計分為例，分子q_i' 與 $\widehat{q_{i,j}}'$之間的交集長度即兩者的 LCS 長度，分母 q_i' 與 $\widehat{q_{i,j}}'$之間的聯集長度則為q_i' 與 $\widehat{q_{i,j}}'$個別長度之和再減掉兩者 LCS 的長度。同理，r 的計分方式相同，最後取 q 與 r 的平均。

範例：

假設q_i'為字串 `not to be? No, to be` 且$\widehat{q_{i,j}}'$為字串 `to be or not to be`。先對q_i'和$\widehat{q_{i,j}}'$進行預處理(分詞)分別得到`['not', 'to', 'be', 'No', 'to', 'be']`與`['to', 'be', 'or', 'not', 'to', 'be']`，這兩個詞組的 LCS 為`['to', 'be', 'to', 'be']`，因此$|q_i' \cap \widehat{q_{i,j}}'|$為 4 且$|q_i' \cup \widehat{q_{i,j}}'|$為 6+6-4=8，該項得分為 4/8=0.5。

特別注意，為了降低人工標註答案對比賽公平性可能造成的影響，測試資料中混有大量不列入最終評分的資料。