110-1 資訊檢索與擷取
期末專案報告

朱紹瑜 B06705028　周敦翔 B07902050　廖政華 B07207063

1 Introduction

本篇報告為 110-1 學期資訊檢索與擷取 (CSIE 5460) 課程之期末計畫結報，主要任務為建構檢索系統，使得能根據給定的醫療紀錄，從生物醫學文獻資料集中檢索出主題相近的文件，並依據相關性進行排名。
程式碼連結：https://github.com/Ethan07902050/1101-IR-final

1.1 資料集簡介

- **Document**：共 100,000 篇生物醫學領域之英文文獻，以 XML 格式儲存，其中每篇包含摘要（abstract）和本文（body）兩部分。
- **Query**：共 30 篇醫療紀錄，每篇包含病歷資料 (note)、病歷描述 (description) 以及病例大綱 (summary) 三個部分，其中病歷資料所記錄的資訊最為詳細，如各種生理指標、症狀和用藥劑量等，形式上較偏向條列式筆記，且使用較多縮寫。病歷描述可視為病歷資料的統整敘述，較有一致的行文結構。而病例大綱則是病歷描述之再精簡化，只由一至兩個句子組成。

1.2 使用套件

我們主要使用 Python 進行開發，並大量利用資訊檢索任務相關的既有套件，簡介如下：

Gensim. [1] 主要專注於語料庫之主題探勘，支援多種傳統相似度檢索模型，並提供多元的資料前處理函式，我們利用之進行斷詞、詞幹提取、篩選停止詞等前處理程序。

PyTerrier. [2] 為資訊檢索函式庫 Terrier 之 Python 介面，提供直覺化的資料操作運算子，有助於快速建立模型流水線。支援豐富的相似度檢索模型和擴展查詢（query expansion）方法。我們以 PyTerrer 作為建構檢索模型與運行實驗的平台。

SciSpacy. [3] 為自然語言處理套件 Spacy 專注在生物醫學領域的子分支，提供多種預訓練模型，能針對文章中的醫學專有詞彙進行命名實體標記，並支援如 umls 或 MeSH 等資料庫的串接，此相關模組為我們所利用，一方面藉由標記後的詞彙特性篩選較重要的查詢關鍵字，另一方面則使用命名實體於資料庫中之豐富資訊進行擴展查詢。

2 Methods

2.1 資料前處理

2.1.1 Document

為了去除格式的資訊，僅保留有語意資訊的內容，對於每一份 document，我們擷取所有 <p> 標籤及其子標籤的段落內容，將所有英文字元轉為小寫，並移除停用詞 (stopwords)、標點符號、數字。

2.1.2 Query

我們採用與 Document 相似的方法，將 query 中所有英文字元轉為小寫，並移除標點符號。除此之外，在內容方面，原先 query 的資料中包含病歷資料 (note)、簡化至約十句話以內的病歷描述 (description)，以及最為精簡的病例大綱 (summary)，長度約為一至兩句敘述。我們嘗試擷取不同的段落組合組成 query，搭配 BM25 及 DPH 兩個模型進行比較，結果如表 1。我們發現僅包含 summary 可取得較好的檢索表現。

	note + description + summary	note	description	summary
BM25	0.1642	0.1510	0.1423	0.1771
DPH	0.1673	0.1508	0.1374	0.1795

Table 1: 不同 query 的內容對應訓練資料 MAP@50 表現

爾後，我們嘗試進一步精簡 query 的內容。考量到有些詞彙並非停用詞，但對於醫學領域而言意義較為廣泛，較無助於病歷資料的檢索，因此我們透過 scispacy 實作生物醫學相關名詞的命名實體辨識 (Named Entity Recognition, NER)，將 query 的內容精簡至僅保留 NER 辨識出的命名實體。此外，我們亦嘗試若將縮寫的命名實體展開為完整的名稱。我們將這些方法搭配 BM25 及 DPH 兩個模型進行比較，結果如 2，篩選 summary 中的命名實體對於兩個模型有不同的影響，另外，展開 summary 中的縮寫並沒有影響整體檢索的表現。本次期末專案中，後續模型比較使用的 query 均僅保留 summary 中的命名實體，並展開縮寫。

	summary	summary + NER	summary + NER + 展開縮寫
BM25	0.1771	0.1642	0.1642
DPH	0.1795	0.1813	0.1813

Table 2: query 有無經過 NER 及展開縮寫對應訓練資料 MAP@50 表現

2.2 Retrieval Models

接續上述的前處理流程，我們挑選了 pyterrier 提供的數個 retrieval 模型進行實驗，包含 TF-IDF, BM25, InL2, PL2, DLH, DPH 和 Dirichlet LM，並比較不同模型在訓練資料上的表現。在這些模型當中，PL2 表現最好，MAP@50 達到 0.2196。以下簡單介紹我們使用的模型。

2.2.1 TF-IDF

TF-IDF 計算每一個字在文章中的重要性，這個權重與詞語在該文章中出現的次數成正比，與包含該語詞的文章數量成反比。第 j 篇文章中第 i 個語詞可以表示如下

$$w_{i,j} = \frac{f_{i,j}}{\sum_j f_{i,j}} \cdot \log \frac{N}{n_i}$$

其中，$f_{i,j}$ 為詞語 k_i 在文章 d_j 共出現了幾次，N 為資料集中文章的個數，n_i 為有多少文章包含 k_i。雖然 TF-IDF 概念簡單，但在許多時候的表現卻勝過更加複雜的模型，因此在我們的實驗當中，TF-IDF 主要作為與其他模型比較的基礎。

2.2.2 BM25

BM25 是在 70 與 80 年代由 Stephen E. Robertson, Karen Spärck Jones 等人提出的機率模型，給定一個 query 之後，可以用來衡量每一篇文章的相關性。自從這個模型提出之後，發展出許多變型，我們使用的是 Okapi BM25。給定一個 query Q，當中包含字詞 $q_1, \ldots, q_n$，一個文章的相關性分數可以如下計算：

$$\text{score}(d_j, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f_{i,j} \cdot (k_1 + 1)}{f_{i,j} + k_1 \cdot (1 - b + b \cdot \frac{|d_j|}{\text{avgdl}})}$$

其中，$f_{i,j}$ 是 q_i 相對於文章 d_j 的 term frequency，$|d_j|$ 是文章的長度，avgdl 是文章平均的長度，k_1 和 b 是可以調整的參數。在接下來的實驗當中，我們將 k_1 設為 1.2，b 設為 0.75。IDF 的計算如下

$$\text{IDF}(q_i) = \ln\left(\frac{N - n_i + 0.5}{n_i + 0.5} + 1\right)$$

N 與 n_i 的定義如上一小節所述。

2.2.3 InL2

InL2 主要考慮正規化後的 term frequency，計算方法如下

$$w_{i,j} = \left(\frac{1}{tf_n + 1}\right)\left(tf_n \cdot \log_2(\frac{N+1}{n_i + 0.5})\right)$$

tf_n 表示經過正規化的 term frequency

$$tf_n = f_{i,j} \cdot \log_2\left(1 + c \cdot \frac{\text{avgdl}}{|d_j|}\right)$$

c 為正規化的參數。

Table I. Models are Made Up of Three Components[a]

	Basic Models	Formula
P	Poisson approximation of the binomial model	(6)
D	Approximation of the binomial model with the divergence	(8)
G	Geometric as limiting form of Bose–Einstein	(17)
B_E	Limiting form of Bose–Einstein	(19)
$I(n_e)$	Mixture of Poisson and inverse document frequency	(20)
$I(n)$	Inverse document frequency	(19)
$I(F)$	Approximation of $I(n_e)$	(21)
	First Normalization	
L	Laplace's law of succession	(23)
B	Ratio of two Bernoulli processes	(26)
	Second (Length) Normalization	
$H1$	Uniform distribution of the term frequency	(41)
$H2$	The term frequency density is inversely related to the length	(42)

[a]For example, $B_E B2$ uses the limiting form of Bose–Einstein formula (19), normalized by the incremental rate of the Bernoulli process of formula (26) and whose within document–term frequency is normalized by formula (42).

Figure 1: Divergence from Randomness 架構下的三個模組

2.2.4 PL2

PL2 是 Divergence from Randomness [4] 架構下的模型，這個架構下的模型主要由三個模組組成，如圖 1 所示。PL2 中詞語的權重計算方法如下

$$\begin{cases} w_{i,j} & = \left(\frac{1}{tf_n+1}\right)\left(A+B+C\right) \\ A & = tf_n \cdot \log_2(\frac{tf_n}{\lambda}) \\ B & = \left(\lambda + (\frac{1}{12 \cdot tf_n} - tf_n\right) \cdot \log_2 e \\ C & = 0.5 \cdot \log_2(2\pi tf_n) \end{cases}$$

tf_n 與上一小節相同，代表正規化後的 term frequency，λ 則代表 Poisson distribution 中的平均和變異數。

2.2.5 DLH

DLH 主要是以 term frequency 的 hyper geometric distribution 為基礎，這個模型假設 query term 在文章中的出現次數是從所有文章抽樣，而非單篇文章 [5]。文章與 query 的相關性計算如下

$$\text{Score}(d_j, Q) = \sum_{q_i \in Q} \text{qtw} \cdot \left(\frac{1}{f_{i,j}+0.5}\right)\left(\log_2\left(\frac{f_{i,j} \cdot \text{avgdl}}{|d_j| \cdot (N/F)}\right) + 0.5 \cdot \log_2\left(2\pi f_{i,j}(1-\frac{f_{i,j}}{|d_j|})\right)\right)$$

- $F = f_{i,j}/|d_j|$
- qtw = qtf / qtf_{max}
- qtf 為 query term frequency 而 qtf_{max} 為所有 query term 當中最大的 qtf

2.2.6 DPH

與 DLH 同樣以 hyper geometric distribution 為基礎，但是使用 Popper's normalization。

2.2.7 Dirichlet Language Model

language model 將每一篇文章都視為一個機率模型。把一個字餵進這個模型，就會產生對應的機率。定義清楚計算 language model 的流程之後，我們就可以依據每一篇文章的內容，計算其對應的 language model，再來計算產生 query 當中每一個詞的機率，全部相乘之後作為文章與 query 相關性的指標。簡而言之，如果一篇文章越有可能產生一個 query，那兩者的相關性應該愈高，反之亦然。

一個簡易建立 language model 的方法，是計算文章當中每一個詞出現的頻率。這樣會產生一個問題：依據上述的計算方式，如果 query 包含一個文章中沒有出現過的詞語，兩著的相關性會降低為 0，但這顯然無法反映現實的狀況，因此，我們可以將一部分有出現在文章中詞語的機率，分給沒有出現在文章中的詞語，這個技巧稱作 smoothing，Dirichlet prior smoothing 是其中一種作法。給定一個 language model M_j，我們可以計算產生 index term k_i 的機率為

$$P(k_i|M_j) = \begin{cases} P^S_{\in}(k_i|M_j) & \text{if } k_i \in d_j \\ \alpha_j P(k_i|C) & \text{otherwise} \end{cases}$$

算式中的 C 代表所有文章的集合。在 Dirichlet prior smoothing 當中，產生存在於文章的詞語的機率為

$$P^S_{\in}(k_i|M_j) = \frac{f_{i,j} + \lambda \frac{F_i}{\sum_i F_i}}{\sum_i f_{i,j} + \lambda}$$

其中，$f_{i,j}$ 代表詞語 k_i 在文章 d_j 共出現了幾次，F_i 則代表 k_i 在所有文章中出現的次數，我們可以進一步推導出

$$\alpha_j = \frac{\lambda}{\sum_i f_{i,j} + \lambda}$$

並且依此計算不存在於文章中的語詞所對應的機率。

2.2.8 實驗結果

從表 3的數據中我們可以發現，相較於 weak baseline 的 0.1141，基礎的 TF-IDF 模型已經有不錯的表現，MAP@50 達到 0.1856，雖然這只是訓練資料的數據。其他模型如 BM25, InL2, DLH 大概是在同一個水準，PL2 與 DPH 表現較為突出，分別達到 0.2196 與 0.2023，顯示正規化對於模型的表現有正面的影響；另外，Dirichlet LM 的分數相當接近 PL2，可以推測在這個資料集上，language model 也是一個可以考慮的切入點。我們挑選表現最好的三個模型—PL2, DPH, Dirichlet LM—進行後續的實驗。

	TF-IDF	BM25	InL2	PL2	DLH	DPH	Dirichlet LM
MAP@50	0.1856	0.1853	0.1820	0.2196	0.1820	0.2023	0.2117

Table 3: 不同模型對應訓練資料 MAP@50 表現

2.3 Query Expansion

根據第一次檢索回來的文件，query expansion 模型會找出文件中適當的關鍵字，加入原本的 query，並再次進行搜尋，以找出更多相關的文件。我們嘗試了 Bo1 Divergence, Kullback Leibler divergence [4], RM3 relevance model [6] 三種不同的模型，接續在上一節結果中表現較佳的三個檢索模型之後，找出合適的組合；再進一步實驗兩次 query expansion 是否會比一次 query expansion 擁有更好的表現，最後，我們的得到最好的結果，是以 DPH 作為檢索模型，第一次 query expansion 使用 Bo1，第二次 query expansion 使用 KL，這樣的流程在訓練資料上 MAP@50 的分數達到 0.2612，public 的測試資料分數為 0.2120，private 測試資料分數為 0.2340。

表 4為不同的檢索模型搭配不同 query expansion 模型之後得到的 MAP@50 分數，在以下的實驗中，我們統一讓 query expansion 模型考慮檢索結果中相關性最高的十份文件，並且在每一次的 expansion 中加入 5 個新的詞語。整體來看，模型表現有所提升。PL2 原本的分數為 0.2196，搭配 Bo1 之後提升到 0.2517；DPH 原本的分數是 0.2023，搭配 Bo1 之後提升到 0.2578；Dirichlet 原本的分數為 0.2117，搭配 KL 之後分數提升到 0.2264，與其他兩個模型相較，進步的幅度較小。我們選擇表現提升幅度最大的模型 DPH，再進行第二次的 query expansion，檢驗加入更多的詞彙是否還能進一步提升模型的表現。

表 5為 DPH 與兩次 query expansion 模型組合後的檢索結果，表中左方為第一次 query expansion 使用的模型，上方則是第二次 query expansion 使用的模型，舉例來說，第一次套用 Bo1、第二次套用 KL 的 MAP@50 為 0.2612，相較於只使用 Bo1 的檢索結果 0.2578，表現略為提升；不過，DPH 搭配兩次 RM3 的 expansion 結果明顯較差。因此，我們最後決定選用 DPH 加上 Bo1 與 KL 作為最後檢索模型的組成，在 public 的測試資料集上分數達到 0.2120，private 的測試資料集分數則是 0.2340。

	Bo1	**KL**	**RM3**
PL2	0.2517	0.2408	0.2311
DPH	0.2578	0.2497	0.2389
Dirichlet LM	0.2199	0.2264	0.2150

Table 4: 經過一次 query expansion 之後訓練資料的 MAP@50

	Bo1	**KL**	**RM3**
Bo1	0.2601	0.2612	0.2372
KL	0.2625	0.2545	0.2335
RM3	0.2516	0.2487	0.2147

Table 5: DPH 經過兩次 query expansion 之後訓練資料的 MAP@50，左方欄位為第一次 query expansion 使用的模型，上方欄位則為第二次 query expansion 使用的模型

2.4 命名實體連結（Named Entity Linking）

被標記的命名實體會進一步跟生醫資訊相關之資料庫中概念相同的實體進行連結，此處我們使用 UMLS 和 MeSH 這兩個資料庫。SciSpacy 提供了存取資料庫資訊的統一介面，其中每個實體對應到的資訊中，我們提取了語意類型（Semantic type）、標準名稱（Canonical name）和別名（Aliases）這三種資訊並進行下列嘗試：

2.4.1 以語意類型篩選關鍵字

UMLS 額外提供了語意類型的資訊，我們以此篩去與疾病或生理狀況較無相關的命名實體，應有助於使 query 更加精確。經過對資料的仔細審視（Appendix A），我們選擇篩去語意類型為"temporal concept" 的實體，不過跟基準表現並沒有太大差異。

	discard "temporal concept" entities
UMLS	0.2652

Table 6: query 使用語意類型篩選關鍵字 MAP@50 表現 (baseline=0.2625)

2.4.2 加入別名擴展查詢

由於醫學領域用語經常出現同義異形字的情形（如通俗用語與專業術語之別），又我們使用較為精簡的 summary 進行檢索，其能捕捉到的脈絡資訊較為有限而可能無法彌補用字差異的問題，故推測在傳統模型的情況下，將同義字納入應能在與文件計算相似度時提供更完整的訊息。我們分別嘗試將正規名稱、第一個別名、前五個別名以字串形式連接於原始 query 中相應字彙。結果如表 7，可以發現表現反而下降了。

	canonical_name	aliases-1	aliases-5
UMLS	0.2422	0.2368	0.2092
MeSH	0.2247	0.2054	0.1918

Table 7: query 使用不同資訊進行擴展查詢 MAP@50 表現 (baseline=0.2625)

3 Discussions and Conclusion

本報告旨在建構以電子化醫療紀錄對生醫領域文獻資料庫進行檢索之檢索模型。總體而言，僅保留醫療紀錄之 summary 部分中的命名實體作為 query，並搭配以 Bo1 進行擴展查詢的 DPH 檢索模型達到最佳表現。由此，我們發現針對檢索資訊去蕪存菁能有效提升表現，然而這進一步產生資訊量過少的問題，為突破此瓶頸，我們引介進擴展查詢，而基於 Bo1 模型的局部上下文分析（Local Context Analysis）做法獲得了顯著的表現提升。此外，我們也嘗試提取命名實體於 UMLS 和 MeSH 等資料庫中的資訊，一方面進一步篩選核心相關的關鍵字，另一方面用以擴展查詢，然而表現不佳。

而未來仍有許多能繼續進展的方向。首先由於時間與計算效能的限制，我們對 documents 的處理僅止於簡單的字串清理與斷詞，而未來可對 document 的主題組成或文本特性等諸多面向進行深度分析，應能為前處理程序和模型選擇等議題提供許多啟發，並可進一步探討 query 和 document 處理方式的特性與差異。此外，我們認為命名實體連結仍有其潛力，而本篇報告中直接將字串串接回去之作法確實較為粗糙，未來如篩選候選命名實體和加權函數等機制皆應被審慎地考慮與設計。最後，也可以嘗試將神經網路生成之特徵納入檢索模型，應能捕捉到許多與傳統模型面向相異的資訊，有望能增進表現。

References

[1] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[2] Craig Macdonald and Nicola Tonellotto. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*, 2020.

[3] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[4] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, oct 2002.

[5] Ben He and Iadh Ounis. Combining fields for query expansion and adaptive query expansion. *Information Processing Management*, 43(5):1294–1307, 2007. Patent Processing.

[6] Nasreen Jaleel, James Allan, W. Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark Smucker, and Courtney Wade. Umass at trec 2004: Novelty and hard. 01 2004.

Appendix A　Query 中命名實體的語意類型

Semantic Type	Count	Semantic Type	Count
Disease or Syndrome	48	Animal	1
Temporal Concept	34	Indicator, Reagent, or Diagnostic Aid	1
Sign or Symptom	34	Phenomenon or Process	1
Finding	28	Molecular Function	1
Population Group	21	Immunologic Factor	1
Pharmacologic Substance	17	Health Care Related Organization	1
Mental or Behavioral Dysfunction	13	Inorganic Chemical	1
Qualitative Concept	12	Hormone	1
Body Part, Organ, or Organ Component	11	Fungus	1
Organic Chemical	11	Cell	1
Therapeutic or Preventive Procedure	10	Mental Process	1
Patient or Disabled Group	9	Injury or Poisoning	1
Pathologic Function	9	Age Group	1
Organism Attribute	7	Enzyme	3
Functional Concept	7	Intellectual Product	2
Occupation or Discipline	7	Element, Ion, or Isotope	2
Spatial Concept	6	Idea or Concept	2
Health Care Activity	6	Neoplastic Process	2
Organism Function	6	Laboratory or Test Result	2
Body Substance	5	Manufactured Object	2
Body Location or Region	4	Bacterium	2
Geographic Area	4	Biomedical or Dental Material	1
Amino Acid, Peptide, or Protein	4	Medical Device	1
Quantitative Concept	3		
Clinical Attribute	3		
Diagnostic Procedure	3		
Biologically Active Substance	3		
Laboratory Procedure	3		