

AI Cup 2018 生醫論文自動分析熱身賽簡介

近年來隨著人工智慧技術的蓬勃發展，人工智慧的應用逐漸落實在日常生活周遭。全球性的市場調查諮詢公司 Markets & Markets 預估人工智慧應用的市場將從 2017 年的 160.6 億美元成長到 2025 年的 1,906.1 億美元，年均複合增長率達到 36%。獲益於人工智慧技術的產業包括製造業、金融科技、醫療領域等。這一波人工智慧的發展預期將會在 2022 年為醫療領域帶來 79 億 880 萬美金的產值。主要應用的人工智慧技術包括機器學習、自然語言處理、情境感知以及電腦視覺。其中自然語言處理技術更是在促成智慧醫療領域中扮演最重要的推手，依此技術所開發的各種醫療決策分析系統，如 IBM Watson 等，已經開始蓬勃發展與應用。

臺灣醫療技術在世界上頗負盛名，這波的人工智慧技術更是掀起了產官學界一股智慧醫療的風潮。本計畫期望能夠藉由此計畫的推動，在這股即將到來的熱潮中扮演領頭羊的角色，引領學子在學習人工智慧技術的同時，將其應用到跨領域的生物醫學智慧計算領域。本計畫將以自然語言處理技術為核心，開放競賽語料讓學子能夠將人工智慧技術應用於基礎的臨床醫療病歷資料分析到進階的生物資訊文獻語料，據此培養具備人工智慧、機器學習、自然語言處理與倫理道德素養的專業人才，並期待在本競賽中發展的技術成為臺灣健康資料加值應用中最重要的一環。參與本競賽的學子將可實務的了解分析生物醫學資料的流程與技術，並體現轉譯醫學思維，讓學子能夠了解利用自然語言處理技術將基礎醫學、生物資訊研究，和臨床治療連結的大思維。

生物命名實體辨識做為文獻探勘研究的基礎，本熱身賽會提供人工標註的生醫文件（來源包含：PubMed、PubMed Central、以及專利文獻摘要），並將資料分為訓練集、建構集與測試集。熱身賽內容包含三項任務，細節如下：

任務一：參賽隊伍須辨識文獻中所提及的基因、疾病以及化學物質共三種命名實體類型。

任務二：回答上述三類之基因、疾病與化合物分別對應至生醫資料庫中之 ID 編號，其中基因 ID 會以人類基因為主，但包含部分非人類基因。

任務三：進一步回答文章中疾病源發自何種器官或組織（例如：肺臟、皮膚、血液或骨髓，限縮在給定之 57 項器官組織中，相關規定亦會於標註準則中說明並公佈），以及該篇文章中的化學物質命名實體是否會引發或造成該文章中何種疾病。

- 初階參賽者須完成任務一實體辨識
- 進階參賽者須完成任務一及任務二所指定回答項目
- 高階參賽者則須完成任務一至三所有項目。