專題報告

賴冠綸

指導教授: 謝孫源 教授、張升懋 教授
技術支援: 研發處 邱鈺期 小姐
網頁設計: 陳星文 同學

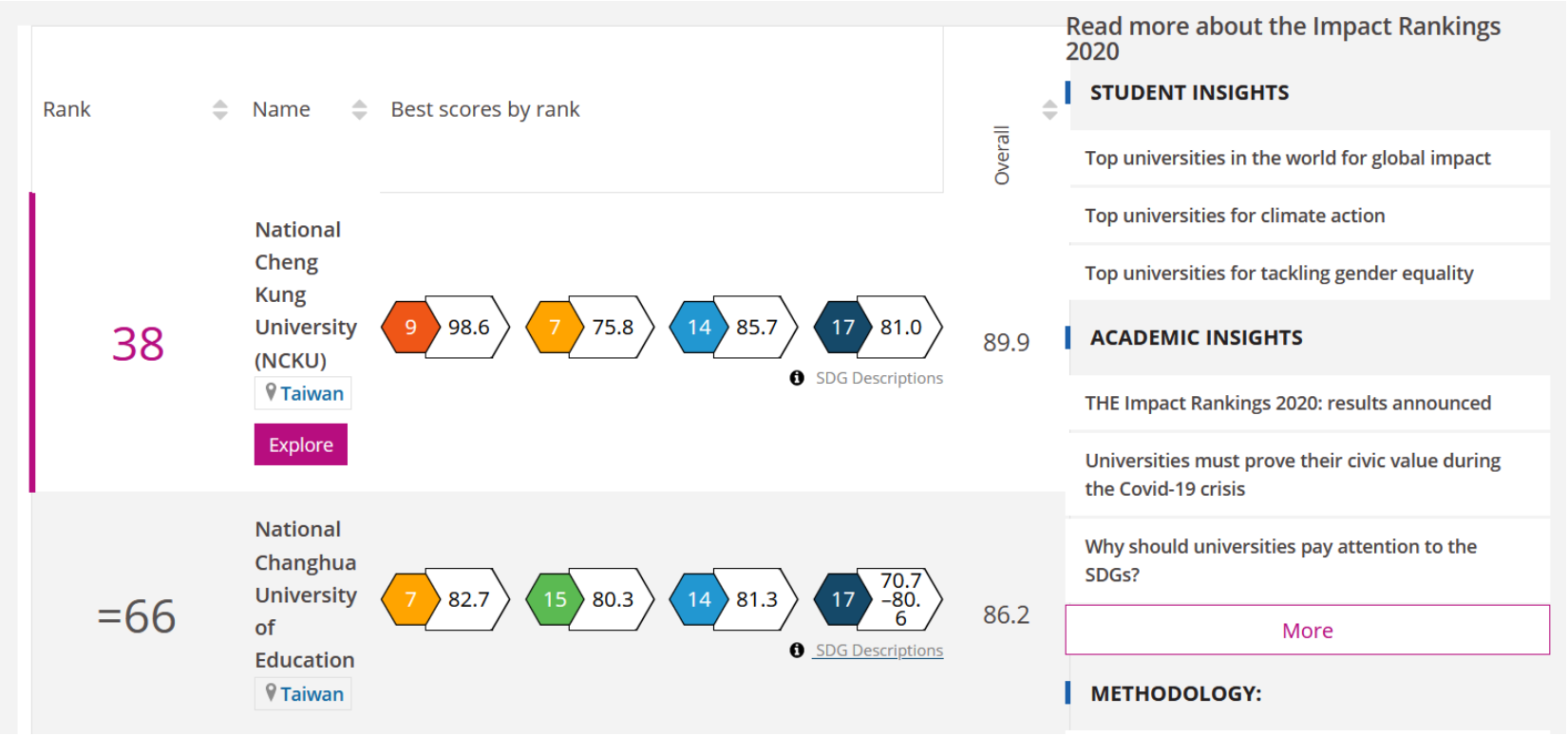
大綱

1. 前言
2. 動機
3. 流程
4. 蒐集資料
    - Google 搜尋
    - 爬蟲抓取摘要
5. 資料分析
    - 單字萃取
    - Tf-Idf 分析
    - 成大關鍵字與國際關鍵字的機率分析
6. 成果展示
7. 結論

前言

SDGs(Sustainable Development Goals): 是聯合國所制定一系列的發展目標，具體有 17 個目標以及其隨附的 169 項指標。

而我們成大在 2020 年的 impact rankings 為台灣第一。



動機

雖然成大在 2020 的排名上有著優異的表現，但我相信仍有一些論文沒能被算入 SDG 的範疇。

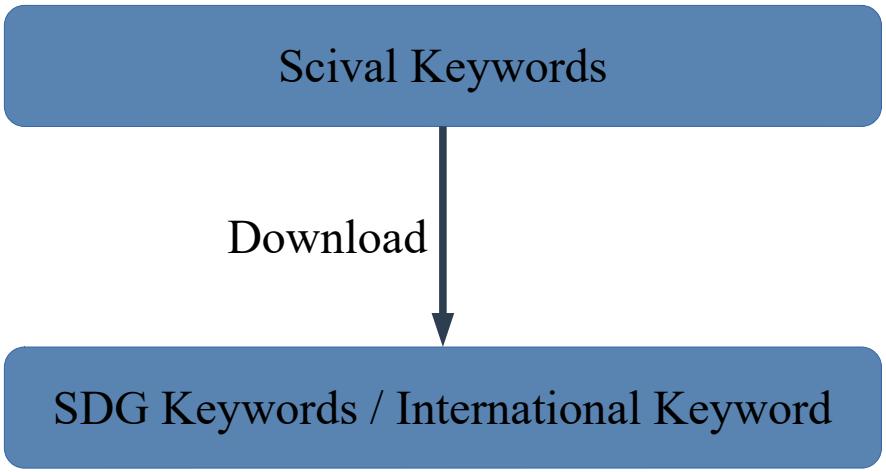
這可能是因為我們在字詞的使用沒有符合分類的標準，而導致沒有被分類到正確的 SDG 當中，甚至不被計算。

因此我想透過這個專題來提供一個針對成大論文的推薦系統，對於相對應的 SDG 做出分類以及推薦。

流程 - Data Collection

Step 1:

從 Scival 下載 SDG 的關鍵字來作為國際關鍵字。

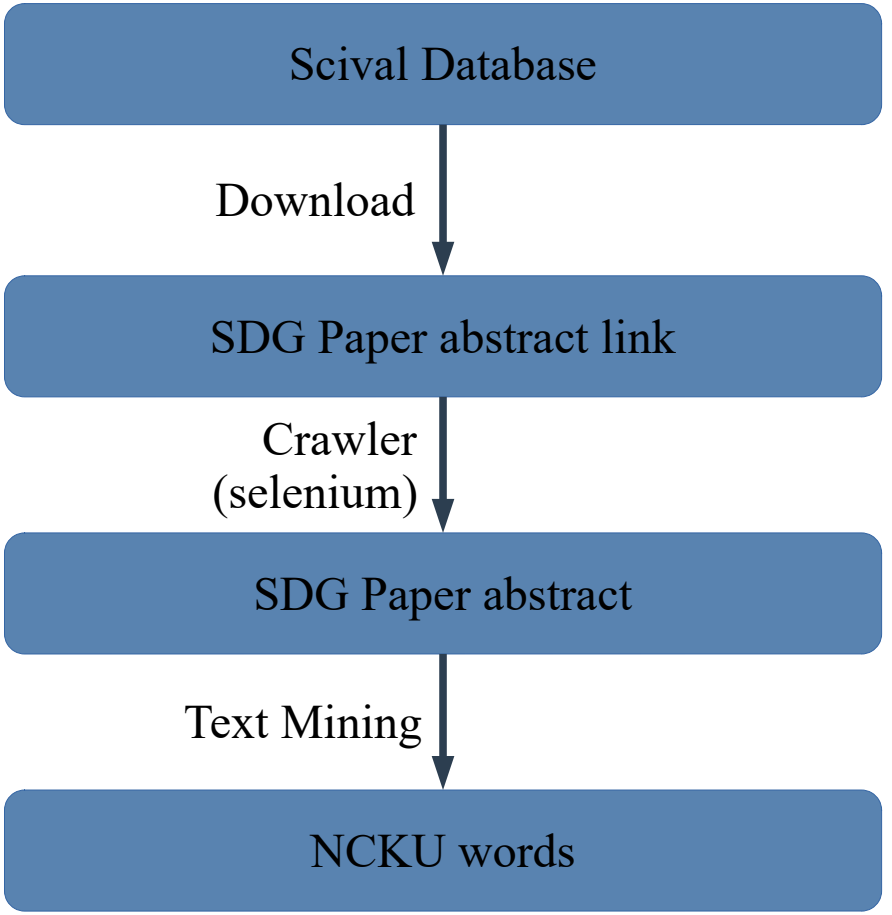


關鍵字網址連結

流程 - Data Collection

Step 2:

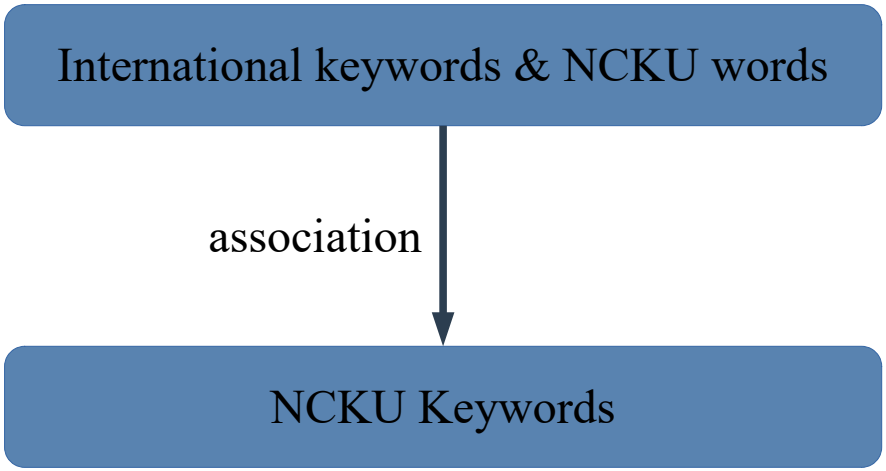
從 Scival 上獲取成大論文摘要的網址，
(因為 Scival 不提供大量的摘要下載)
再使用爬蟲獲取摘要，
最後透過文字處理來分離出屬於成大的單詞。



流程 - Data Collection

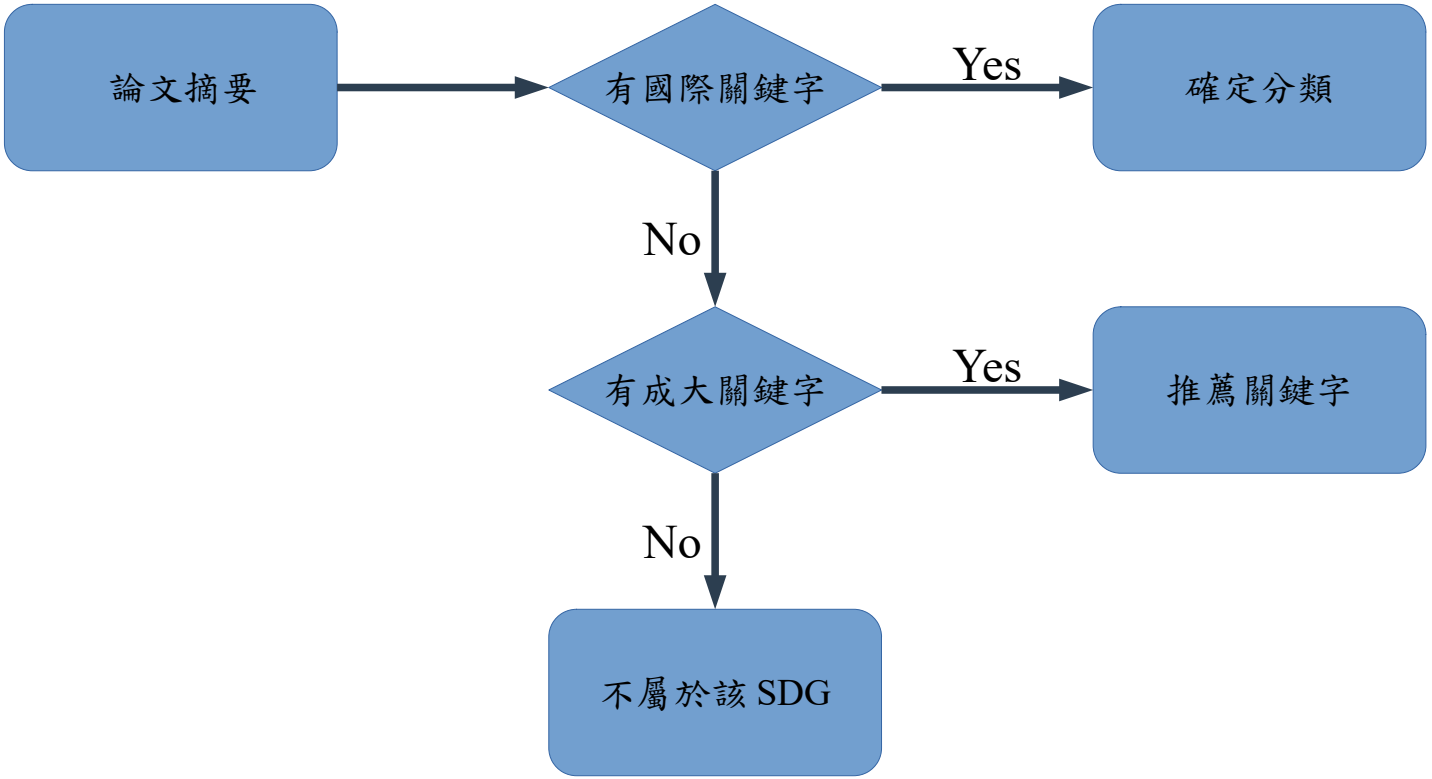
Step 3:

將國際關鍵字與成大的單詞進行連結，
找出其中有相關的組合，
並整理出成大關鍵字。



流程 - Keyword Recommendation

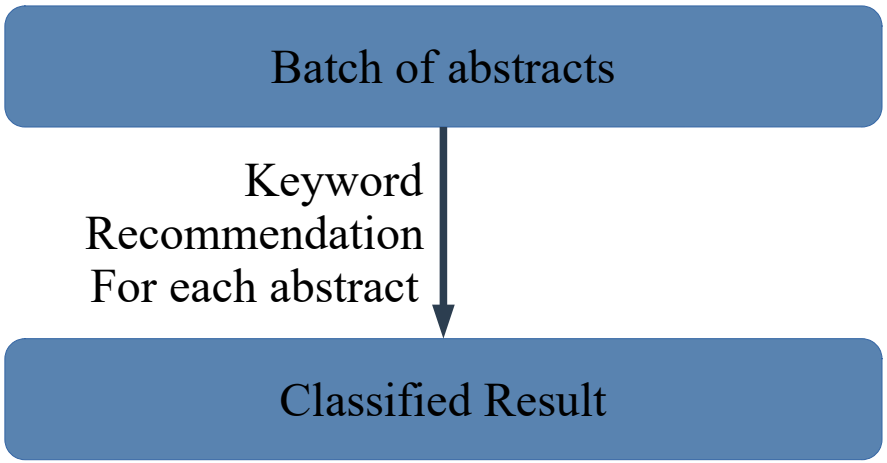
Step 4： Check SDG 1~16



流程 - Batch

Step 5 :

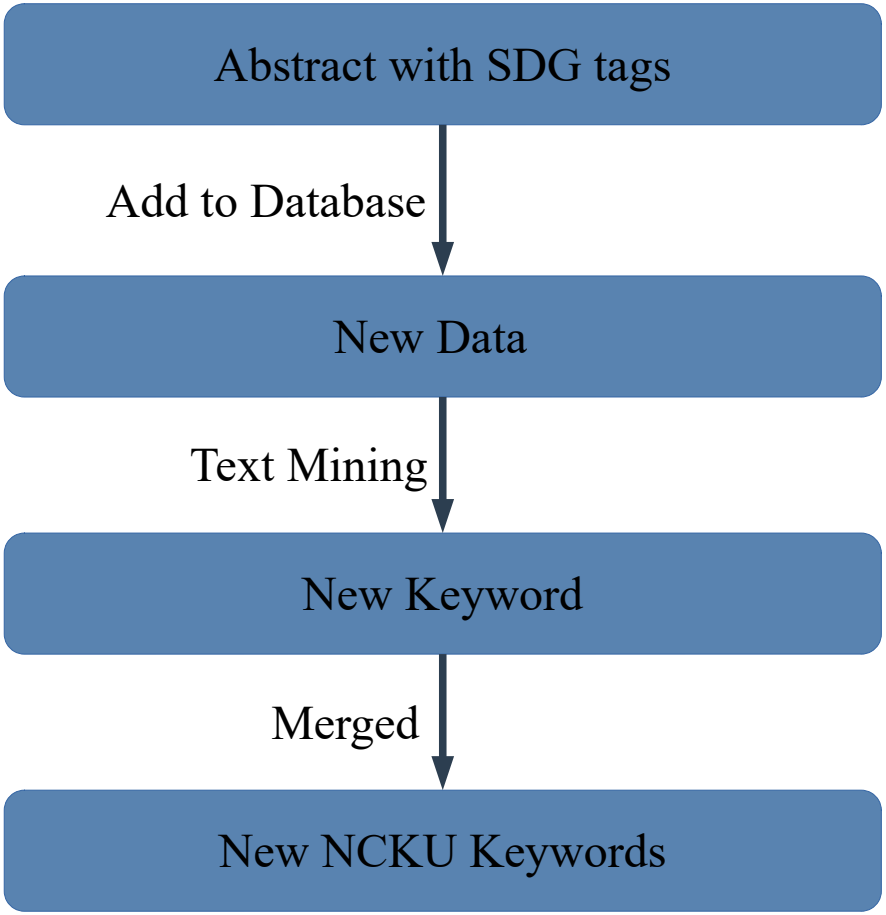
在使用者上傳檔案後，
對整個檔案進行分類以及推薦。



流程 - Further Data Collection

Step 6 :

使用者也可以在分類時提供該論文的正確分類，
我們會將其進行標記以及收錄，
在其滿足一定數量或是在固定周期後，
我們會對其進行分析，
並擴張原有的關鍵字庫。



實作細節

透過爬蟲取得摘要

原本我使用的資料是教授所提供的，其中並沒有包含論文摘要的部分，所以我使用爬蟲來抓取連結中的資料。

在爬蟲的部分我使用了 selenium 的 webdriver 並透過 BeautifulSoup 來進行分析與資料提取。

使用 selenium 而非 scrapy 的原因是這個論文網站有較強的反爬蟲能力，而 selenium 能夠更好的應對反爬蟲。

該論文網站有提供摘要下載的功能，但是有資料量的限制，所以我在這裡使用爬蟲來完成抓取摘要的工作。

Scrapy/BS4/Selenium 優缺比較

	Scrapy	Bs4	Selenium
優點	基本的爬蟲架構已經搭好了，只需要填充自己的規則就可以了，結構清晰。	簡單，容易上手。	動態頁面的爬蟲，頁面交互能力，避免反爬，因為真的是模擬用戶打開瀏覽器的操作。
缺點	在有反爬蟲的網站上效果不太好。	由於 BS4 是通過層級關係一層一層的達到目的標籤，速度比較慢，解析數據慢。	爬取數據比較慢。

原文網址：https://kknews.cc/tech/l6bo5r9.html

資料分析 - 單字萃取

原因 : 為了分析單詞之間的關係，所以需要將摘要拆分成單詞，又因為單詞太多，所以最後決定只挑選其中的名詞部分。

做法 : 使用 nltk 套件來進行拆分，以及挑選名詞的工作。

資料分析 - Tf-Idf 分析

原因 : 單字雖然只保留了名詞的部分，但我希望可以在過濾掉一些不重要的單字，所以透過 Tf-Idf 演算法來了解各個單詞的重要性，進一步作為挑選單字的依據。

做法 : 實作 Tf-Idf 算法

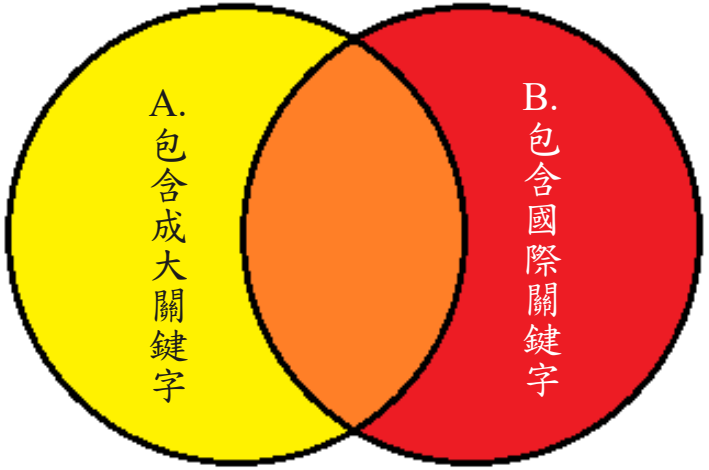
Tf = 單詞出現次數 / 文章所有單詞數

Idf = $\log_{10}$(所有文章數 / 出現該單詞的文章數)

Tf-Idf = Tf * Idf

成大關鍵字與國際關鍵字的機率分析

想法：透過在成大的論文中查找國際關鍵字的方式，來將同一篇論文中的其他字與之連接，希望透過這樣的方式來找出與國際關鍵字同義且獨屬於成大的關鍵字。



作法：

透過計算：$\frac{(P(A\cap B))}{(P(B))}$

可以判斷出現國際關鍵字時，

其他字出現的可能性，

並進一步地把有較高機率的字定為成大關鍵字。

成果展示

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
Volume 11937 LNCS, 2019, Pages 207-215
2nd International Conference on Innovative Technologies and Learning, ICITL 2019; Tromsø; Norway; 2 December 2019 到 5 December 2019; 代碼 234499

Improving Programming Education Quality with Automatic Grading System (Conference Paper)

Cai, Y.-Z. ✉, Tsai, M.-H.

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

這一篇論文為 SDG4
有被成功分類出來

摘要

As the rapid growth of information technology, the demand for proficiency traditionally, hands-on programming training is more beneficial and pract all assignments in person. Besides, students may not get feedback immed system is required to grade and send feedback to students. Based on an ex behave as expected, we develop a set of course management tools and dep server to run and test the programs. However, the server is susceptible to l and prevent malicious network traffic are demonstrated in this paper as we with 140 students enrolled. Around 72% of the students indicate the autor Switzerland AG 2019.

SDG	chance	matched	suggest
SDG 4 - Quality Education	2	environmental education	
SDG 6 - Clean Water and Sanitation	1	communities; communities	wastewater, treatment, freshwater; water purification, freshwater
SDG 7 - Affordable and Clean Energy	1	sectors	energy transition
SDG 11 - Sustainable Cities and Communities	1	program; education; students; building; school; campus	city, circular economy; city, circular economy; city, circular economy; human settlement, disaster; urban, pollutant; urban, pollutant
SDG 15 - Life On Land	1	value	land, ecosystem
SDG 3- Good Health and Well-being	0.5	process; research	child abuse; child abuse
SDG 8 - Decent Work and Economic Growth	0.5	intention	medium enterprise
SDG 9 - Industry, Innovation and Infrastructure	0.5	intention; value	medium enterprise; product innovation
SDG 10 - Reducing Inequality	0.5	energy	migration polic
SDG 13 - Climate Action	0.5	energy; building	climate, energy conservation; climate, energy conservation
SDG 14 - Life Below Water	0.5	development; development	marine, fishery; ocean, fisheries
SDG 16 - Peace, Justice, and Strong Institutions	0.5	process; research	child abuse; child abuse
SDG 1 - No Poverty			

成果展示

Cities
Volume 84, January 2019, Pages 56-65

Does bus accessibility affect property prices? (Article)

Yang, L.[a], Zhou, J.[b], Shyr, O.F.[c], Huo, D.D.[a]

[a]Department of Real Estate and Construction, Faculty of Architecture, The University of Hong Kong, Hong Kong
[b]Department of Urban Planning and Design, Faculty of Architecture, The University of Hong Kong, Hong Kong
[c]Department of Urban Planning, National Cheng Kung University, Taiwan

這一篇為 SDG11
有成功被分類且進行推薦

摘要

Existing studies have yet reached consistent conclusions on accessibility b
of the West, where bus patronage is generally low. In this study, we used a
estates in Xiamen, China to develop four non-spatial hedonic pricing moc
econometric models to quantify the effects of bus accessibility on property
would influence estimates of such benefits. Our findings are as follows. (1
outcome is in contrast with findings of mainstream research (or conventic
0.5% higher, all else being equal. (2) Bus travel times to essential destinati
that account for spatial autocorrelation outperform traditional hedonic pri
plausibility of this study. However, the price premiums offered by bus acc
declining attractiveness for bus travel and continuous transit service enha

SDG	chance	matched	suggest
SDG 5 - Gender Equality	1	effects	female employment
SDG 8 - Decent Work and Economic Growth	1	travel	sustainable tourism
SDG 11 - Sustainable Cities and Communities	1	bus; transit; bus; transit; bus; transit; property	cities, congestion; cities, congestion; cities, transportation; cities, transportation; cities, public transport; cities, public transport; human settlement, disaster
SDG 12 - Responsible Consumption and Production	1	travel	sustainable tourism
SDG 13 - Climate Action	1	property; travel	climate, awareness; climate, decision-making
SDG 3- Good Health and Well-being	0.5	research	child abuse
SDG 9 - Industry, Innovation and Infrastructure	0.5	benefits	product innovation
SDG 16 - Peace, Justice, and Strong Institutions	0.5	research	child abuse
SDG 1 - No Poverty			

結論

推薦器的功能達成預期，能夠進行分類以及推薦。

然而在效果上仍有改進的空間，因為依然有論文無法被正確分類以及推薦。

問題可能出在於以下幾點：

1. 提取 International Keywords：所參考的研究成果可能不完全是現在分類的依據，在之後需要進一步更新。

2. 論文樣本太少：導致在建立 NCKU Keywords 時，出現偏差。

3. NCKU Keywords 的建立：目前是根據單字同時出現的頻率來進行建立，無法確保 NCKU Keywords 與 International Keywords 的關聯。