

國立高雄大學電機工程學系(研究所)

碩(博)士論文

基於自注意力機制自動化產生多文本摘要

The Study of Self-Attention Automation generate Multi-document Summarization

研究生：葉柏漢 撰

指導教授：施明昌 教授

吳東穎 博士

中華民國113年07月

**謝辭**

時光匆匆，兩年的時間如同白駒過隙，碩士生涯即將落幕，在完成這篇論文的過程中獲益良多，我要感謝所有支持和幫助過我的人。首先，我要感謝我的指導教授-吳東穎博士。總能在我苦惱不已時提供最即時最有用的知識與資料，當論文方向走偏時也能即時的修正軌道，讓論文的內容擁有最佳的品質。

再來也要感謝曾經參與產碩專班的在職同仁們，給予我最寶貴的意見，不僅僅是論文上的意見，甚至在課堂上的一些事務根據過往經驗來提供我幫助，是我努力支撐下去的動力。

接著，我要感謝評閱委員們，感謝您們抽出寶貴的時間閱讀我的論文，並提出寶貴的意見和建議。您們的專業評論將使我的研究更加完善。

指導教授：吳東穎 博士

國立高雄大學電機工程所

學生：葉柏漢

國立高雄大學電機工程所

**摘要**

在日新月異的現代，眾人能得到的資訊量越來越多，不管是報章雜誌或是網路媒體的盛行；然而獲得的資訊越多也就代表人們沒幫助、混淆視聽的髒資料也越來越多，例如：重複公布的資料、冗言贅詞、錯誤的資料等…

摘要的出現解決了文章冗長的問題，但並不是每一則文章都有作者提供的摘要，況且摘要通常針對單篇文章做撰寫，若人們閱讀多篇文章，也依然無法解決資料重複的問題，因此有了自動產生多文本摘要的概念。本研究以自注意力機制為主軸，依序由斷詞、提取關鍵詞、輸入自注意力機制、並將過濾完的詞投入GPT生成出容易閱讀的摘要。

本研究選用TFIDF來初步篩選關鍵詞，由於本研究的主軸是「多文本」，TFIDF不僅能選出單篇文章內高頻率的詞，也能去掃描各個文本內的詞頻，並使用自注意力機制將初步篩選出的關鍵詞在進一步篩選，使用自注意力機制選出注意力分數最高的幾個詞作為後續輸入GPT生成摘要的關鍵詞。最後使用ROUGE評估由GPT生成出的摘要，並且參考其他網頁系統所生成的自動摘要，本研究得到的ROUGE分數高於其他系統所生成之摘要。

總而言之，本研究的重點研究方向有以下幾點：

1. 如何精準的抓到關鍵詞？
2. 如何生成人們看得懂的內文？
3. 摘要的質量是否與預想相同？

**關鍵字：**關鍵詞、自動摘要、資訊量、自注意力機制、多文本

Advisor(s): Dr.(Professor) 吳東穎

Institute of National University of Kaohsiung

Student: 葉柏漢

Institute of National University of Kaohsiung

**ABSTRACT**

In this rapidly changing modern world, people have access to more and more information, whether it is the popularity of newspapers, magazines or online media; however, the more information they obtain, it also means that there are more and more dirty information that is not helpful and confusing to people. Many, such as: repeated published information, redundant words, wrong information, etc...

The emergence of abstracts solves the problem of lengthy articles, but not every article has an abstract provided by the author. Moreover, abstracts are usually written for a single article. If people read multiple articles, it still cannot solve the problem of duplication of information. Hence the concept of automatically generating multi-text summaries. This study takes the self-attention mechanism as the main axis, sequentially segmenting words, extracting keywords, inputting the self-attention mechanism, and putting the filtered words into GPT to generate an easy-to-read summary.

This study uses TFIDF to initially screen keywords. Since the main focus of this study is "multi-text", TFIDF can not only select high-frequency words in a single article, but also scan the frequency of words in each text and use self-attention The mechanism will further screen the keywords initially screened out, and use the self-attention mechanism to select the words with the highest attention scores as keywords for subsequent input into GPT to generate a summary. Finally, ROUGE is used to evaluate the summary generated by GPT, and with reference to the automatic summaries generated by other webpage systems, the ROUGE score obtained in this study is higher than the summaries generated by other systems.

In summary, the key research directions of this study are as follows:

1. How to accurately capture keywords?
2. How to generate content that people can understand?
3. Is the quality of the abstract as expected?

**Keywords:** Keywords, automatic summary, amount of information, self-attention mechanism, multi-text

**目錄**

# 謝辭 02

# 摘要 03

# ABSTRACT 04

# 緒論 09

* 1. 研究背景 09
  2. 研究動機 10
  3. 研究目的 11
  4. 研究方法 15
  5. 論文架構 17

# 相關研究 21

* 1. 摘要種類 21
  2. 文字探勘 21
     1. 分詞 23
     2. 關鍵字提取 23
     3. Word Emdeddings 24
  3. 網路爬蟲 23
  4. 中文斷詞 23
     1. 中研院CKIP斷詞系統 23
     2. Jieba斷詞 23
  5. 關鍵詞提取 23
     1. TFIDF 23
     2. Textrank 23
  6. 大型語言模型（Large Language Model) 23
     1. Transformer 23
     2. 自注意力機制(Self-Attention) 23
     3. 多頭注意力(Multi-Head-Attention) 23
     4. 位置編碼(Positional Encoding) 23
     5. 預測與微調(Pre-training and Fine-tuning) 23
     6. 生成和解碼 23
  7. GPT 23

# 系統架構 29

* 1. 系統架構說明 29
  2. 資料集 30
  3. 資料前處理 31
     1. CKIP斷詞 40
     2. 過濾停用詞 40
  4. 關鍵詞提取 32
  5. 詞向量轉換 33
  6. 注意力機制 33

# 實驗分析 34

* 1. 實驗方法 34
     1. 關鍵詞數量 40
     2. ChatGPT下達命令方式 40
  2. 實驗結果評估方式 35
     1. ROUGE-N 40
     2. 摘要生成評估結果 40

# 結論 43

* 1. 研究總結 43
  2. 未來期望 44

# 參考文獻 46

**圖目錄**

圖1.1，美國1940年-2022年的報紙發行量 09

圖1.2，2015-2021年全球的社群媒體使用人數 10

圖2.1，Transformer架構圖 22

圖2.2，Attention架構圖 23

圖3.1，系統流程圖 29

圖3.2，CKIP斷詞結果 30

圖3.3，篩選關鍵詞流程圖 30

圖3.4，TFIDF輸出結果示意圖 31

圖3.5，斷詞流程圖 31

圖3.6，詞向量轉換結果示意圖 32

圖3.7，注意力機制流程圖 33

圖4.1，十個關鍵詞 34

圖4.2，二十個關鍵詞 35

圖4.3，三十個關鍵詞 36

圖4.4，四十個關鍵詞 36

圖4.5，五十個關鍵詞 36

圖4.6，Reference(人工標注)示例 37

圖4.7，自動摘要(程式生成)示例 37

圖4.8，2-grams示例 38

**表目錄(沒動到 先保留)**

表1.1，機械手臂種類表 11

表1.2，自動導引車(AGV)種類表 12

表1.3，工業烤箱(OVEN)作業分類表 14

表1.4，工業烤箱(OVEN)機種分類表 14

1. **緒論**
   1. 研究背景

在21世紀的今天，我們生活在一個資訊爆炸的時代，科技的迅速發展與數位化轉型使得我們能夠即時獲得龐大且多元的資訊。資訊的爆炸性增長不僅改變了我們的生活方式，也對社會、經濟、文化等多個層面產生深遠影響。因此，資訊爆炸成為當代社會、科技與文化研究中一個極具挑戰性且具有深刻影響的主題。

近年來，美國報紙的發行量（請見圖1.1）逐年下降，這種趨勢主要受到數位化媒體和有線新聞平台的競爭影響（請見圖1.2），許多讀者更傾向於以線上形式獲取資訊。隨著智慧手機和社交媒體的普及，閱讀習慣發生了重大改變。許多人更傾向於透過即時、便捷的方式獲取新聞。這種數位轉型使得傳統報紙面臨著挑戰，需要不斷調整以滿足現代讀者的需求。然而，這也提醒我們媒體產業正處於不斷變遷的時代，需要創新和適應。

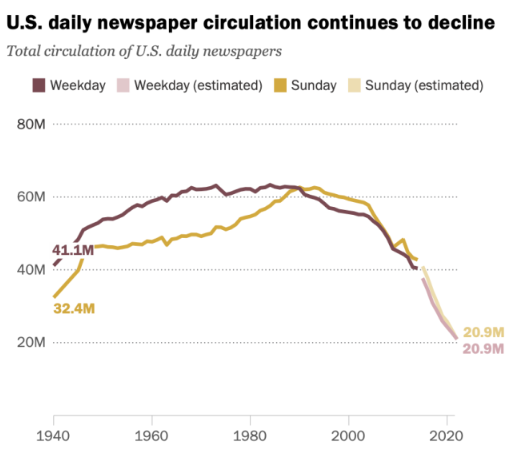


圖1.1，美國1940年-2022年的報紙發行量

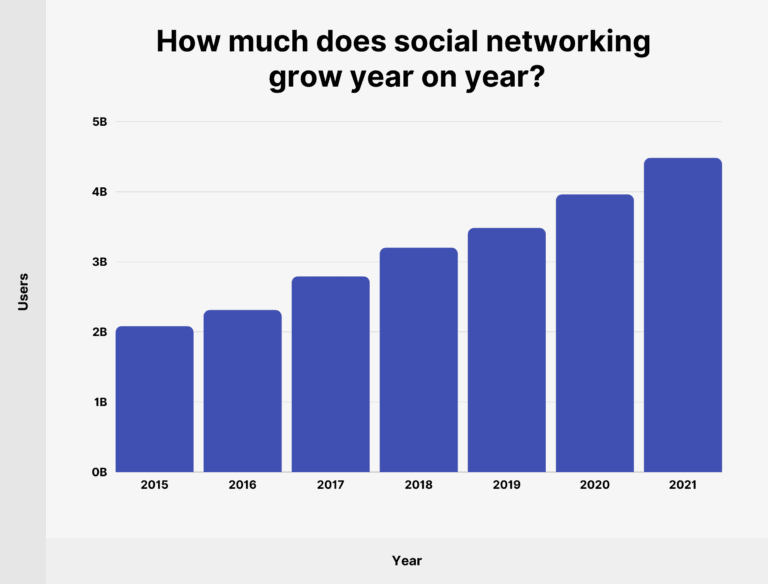


圖1.2，2015-2021年全球的社群媒體使用人數

論及2019年紐約時報的情境，特別是在顛峰時段，例如選舉、奧運等大型活動期間，每天可達1000至1500篇文章。這樣龐大的新聞數據資料庫背後，人們面臨著龐雜的資訊整理工作。在此情境下，發展迅速且有效提取關鍵字的技術成為研究中不可或缺的一環。這種技術不僅對新聞業務有所貢獻，同時也在研究領域中提供了深入挖掘龐雜資料的有效途徑，成為解決資訊過載問題的重要工具。

* 1. 研究動機

作為一位碩士生，我深刻地意識到這個問題對於學術研究的影響。在進行文獻回顧時，我發現自己需要處理大量的文獻資料，而這些資料往往包含了各種各樣的信息。這不僅使得篩選出重要資訊變得極為困難，而且也影響了對於整體研究方向的把握。

這個問題的根本原因在於我們缺乏有效的工具和方法來處理大量資料。傳統的閱讀和分析方式已經無法應對這一挑戰，我們需要新的技術和方法來解決這個問題。因此，我對於如何利用先進的資訊技術，例如自然語言處理和文字探勘，來提高資料處理的效率和準確性產生了濃厚的興趣。

另外在網路搜尋中，使用者經常陷入不確定的情境，不明確自己需要或欲探索的方向。可能查詢結果並非符合預期，導致資訊迷失或不完全滿足需求。若能開發一套系統以文章當作輸入，輸出一篇滿是重點的短篇，就能解決花費大量時間找資料的這一部份。

本次研究我們選擇以文章當作目標，文章的改變是最直接的，從以前的報章雜誌到現在的各大論壇，加上近年年資訊量尋訴的膨脹，也間接改變了人們閱讀的習慣；如何簡潔有力的挑出重點是現在人必學的課題，因此選擇文章當作本論文研究的目標。

本論文的另一個研究重點為ChatGPT，當前的ChatGPT功能非常強大，最具代表性的特性就是具有上下文關聯。但目前市面上的AI工具大多是將資訊餵進ChatGPT並生成出一套「看似」通順的語句，然而只是將一個資訊講成一串一串的句子，前後文並無關聯。會產生這樣的問題我認為給予資料的方式跟下達關鍵字的方式有相當大的關連，因此本篇論文也致力於透過關鍵詞來優化這部分的缺點，讓語句能夠在表達重點摘要的同時，產生出上下文通順的文章。

* 1. 研究目的

本論文的研究目的希望讓使用者將多篇文章做為系統的輸入，經過系統的統整與摘要後，讓使用者看到多篇文章內的所有重點。本研究中，自動摘要的作法有以下特徵:

1. 多文本的摘要

2. 抽象式的摘要

3. 信息檢索摘要

* 1. 研究方法

這份研究旨在探索如何透過先進的資訊處理技術，提高對於大量財經股票相關新聞的處理效率。首先，我們運用網路爬蟲技術擷取聯合新聞網上的財經股票相關新聞，以建立龐大的資料庫。接著，利用中研院開發的ckip斷詞系統，將這些新聞進行初步的語言處理，以備進一步的分析。

在資料預處理的階段，我們選擇使用TFIDF演算法進行初步的關鍵詞選取，以篩選出在文本中具有重要性的詞語。這有助於我們快速捕捉新聞內容中的關鍵信息，但同時也挑戰著如何更精準地捕捉主題。

為了進一步提升我們研究的深度和準確性，本論文的重心轉向自注意力機制的探討。這種機制能夠更精密地挑選經過TFIDF選取的關鍵詞，使我們更好地理解每個詞彙在文本中的重要性。這樣的過程不僅提高了對於新聞內容的理解力，同時也為後續的摘要生成奠定了基礎。

我們的目標是透過自注意力機制彙整出具有代表性的關鍵詞，並透過ChatGPT產生出具有上下文關聯的摘要。

這不僅有助於研究者迅速瞭解大量新聞中的重要信息，同時也為財經股票領域的深入研究提供了一個強大的分析工具。這項研究有望突破在資訊爆炸時代面臨的資訊過載困擾，為資訊處理領域帶來新的思維與方法。

* 1. 論文架構

第一章:前言，講述本論文之動機目的

第二章:相關研究，講述本論文所相關之技術

第三章:系統設計，本論文之系統架構與概念

第四章:實驗分析，探討實驗結果與市面上的摘要生成做比較

第五章:結論，套論本研究的未來展望並做一個總結

1. **相關研究**
   1. 摘要種類

自動摘要[10]因各自需求區分出不同種類，以下簡單介紹幾個:

* + 1. 單文本與多文本摘要(single-document summarization and multi-document summarization)

單文本摘要專注於個別文檔，透過保留主要信息和要點，形成簡潔而具代表性的總結，讓讀者能快速掌握內容。這有助於在信息龐雜的環境中提高閱讀效率，使人們能夠迅速獲取所需資訊。

多文本摘要從多個文檔中提取信息，建構一個融匯多源、突顯最關鍵信息的總結。這樣的摘要不僅使我們能夠全面瞭解多方面的內容，還能發現文檔間的共同主題和關鍵信息，促進跨領域知識整合。

* + 1. 資訊性摘要與指示性摘要(Informative Summarization and Indicative Summarization)

資訊性摘要致力於深入且詳盡地呈現原文中最重要、最具信息價值的內容，以完整展現文檔的主題和主旨。這包括了主要事實、觀點、論據以及其他關鍵細節，旨在提供讀者對文檔內容的深入理解。資訊性摘要常見於需要傳達詳細信息的場景，如新聞報導或學術論文。

指示性摘要則專注於提供簡要概述，指示文檔的主題和要點，而不追求涵蓋所有具體細節。它強調的是文檔的主題、核心觀點和重要信息，以提供一個更輕量、高層次的摘要。指示性摘要適用於一般性概述或摘要回顧等場景，強調簡潔而具有高層次的表達。在實際應用中，選擇使用資訊性摘要或指示性摘要通常取決於讀者的需求和特定應用場景的要求。

* + 1. 單一語言與多語言摘要(monolingial summarization and multilingual summarization)

單一語言摘要專注於處理和生成來自單一語言的文本摘要。透過保留文本中的主要信息和要點，生成簡潔而具有代表性的摘要，使讀者能夠迅速了解單一語言文檔的內容。 常見於處理單一語言的新聞文章、博客、學術論文等。多語言摘要處理包含多種語言的文本，旨在生成橫跨多語言的總結。將不同語言的文檔轉換為共通的摘要，使讀者能夠在跨語言環境中快速瞭解信息。適用於國際性新聞、多語言社交媒體內容，以及全球性企業或研究領域的文本。

* + 1. 選取式與抽象式摘要(single-document summarization and multi-document summarization)

選取式摘要直接從原文中挑選最重要的句子或段落，形成摘要，而不進行內容的改寫。保留了原文的精確性，選取的內容直接來自原文，不會引入不確定性或語法轉換的問題。抽象式摘要不僅選取原文的句子，還進行內容的重新組織和改寫，生成新的表達形式，以更簡潔和具有概括性的方式呈現摘要。具有更高的自由度，能夠以更自然的語言風格呈現，不僅是原文的重組，還可能包含未在原文中出現的表達方式。

* 1. 文字探勘

文字探勘（Text mining）[9]在當今資訊爆炸的時代，扮演著不可或缺的技術角色。它作為資訊探勘的一個分支，雖然與資訊探勘有所聯繫，但同時具有著明顯的區別。資訊探勘主要處理的是「結構性資料」，這類資料已經具有預先定義的欄位名稱和對應的數值，可以直接套入資訊探勘演算法進行分析。相對而言，文字探勘所處理的資料則是最原始的文字資料，這些文字是人們透過自然語言所表達的，並且尚未經過結構化處理，因此無法直接利用資訊探勘演算法進行有意義的數據分析。文字探勘的主要目標在於將這些多樣的文本資料轉換為具有高品質資訊的形式。

* + 1. 分詞

分詞[6]在文字探勘裡屬最基本的步驟，目的在於將連續的文本轉換成許多有意義的單詞，方便後續的分析與處理，例如:情感分析、主題建模等..另外在處理分詞時會考慮廷用詞過濾，將多餘的連接詞與受詞給剃除掉，不同的語言所使用的分詞方法大不相同，如英文可直接使用空格進行分詞，中文則必須經過一連串的步驟才能過濾出完整的詞彙。

* + 1. 關鍵詞提取

文字探勘中的關鍵詞提取技術[11]是一種從文本中自動識別並提取出最具代表性和重要性的詞彙的方法。這項技術有助於總結文本內容、提供快速檢視和理解文本主題的方式。

通常，關鍵詞提取技術可以分為基於統計方法和基於機器學習方法兩大類。基於統計方法的技術通常根據詞彙在文本中的頻率、關聯性和位置等因素來評估詞彙的重要性。而基於機器學習的方法則使用機器學習算法，根據文本的特徵來訓練模型，進而預測關鍵詞。

關鍵詞提取技術在信息檢索、文本摘要、知識庫構建等領域中發揮著重要作用。通過自動識別文本的重要詞彙，可以有效地縮小文本的範圍，幫助用戶更迅速地理解文本內容，同時也提供了文本分析和挖掘的基礎。

* + 1. Word embeddings

詞嵌入（Word embeddings）[8]是一種將詞彙轉換成連續向量空間中的向量表示的自然語言處理技術。這種表示法使得相似意義的詞彙在向量空間中靠近，並且可以捕捉詞彙之間的語義關係。傳統的詞袋模型無法表達詞彙之間的相似度和關聯性，而詞嵌入通過學習詞彙的分佈式表示，使得具有相似語義的詞在向量空間中更接近。

* 1. 網路爬蟲

網路爬蟲（Web Crawler）[7]是一種在互聯網上獲取網頁內容的技術，透過剖析DOM（Document Object Model）一層一層地抓取網頁上的資訊。這項技術的出現，讓我們能夠更有效地自動化獲取互聯網上所需的資訊，使得資料的擷取過程更加迅速、精確。

早期的網路爬蟲主要用於資料的收集與統計。這對於當時互聯網上的資源整合和應用來說是一個巨大的進步。隨著技術的不斷發展和改進，網路爬蟲的應用範疇也逐漸擴大。現代的網路爬蟲不僅能夠進行資料的收集，還能夠對網頁進行漏洞檢查，以提高網站的安全性。

網路爬蟲在搜尋引擎領域的應用尤其廣泛。知名的搜尋引擎如Google、Bing、Yahoo等都依賴於強大的網路爬蟲，透過不斷的爬取和更新，提供用戶最新、最相關的搜尋結果。當使用者進行瀏覽器搜索時，網路爬蟲會自動訪問全球資訊網，將訪問的頁面儲存下來以供使用者瀏覽。

網路爬蟲的使用也受到一些限制。有些網站不歡迎網路爬蟲訪問他們的網站，因為網路爬蟲每次抓取資料都相當於對網頁發出一次請求，過於頻繁的訪問可能導致網頁的負擔，進而發生網頁癱瘓的現象。此外，為了規範網路爬蟲的行為，有些網站透過robots.txt文件來制定規則。網頁管理者可以自訂規範在robots.txt，並要求網路爬蟲遵循這些規則進行抓取，這也是網路爬蟲在操作時需要注意的一點。

* 1. 中文斷詞

在中文寫作中，句子的分界是以標點符號來區別，而句子內部則充滿了各種詞彙。詞彙是由一個或多個字所組成的單位，而句子和詞彙之間存在著明顯的分界。然而，令人困擾的是，詞彙之間卻沒有像句子那樣的明確分隔標誌。這種特性對於詞性標記和資料檢索等後續處理步驟產生了挑戰。

在中文語境下，斷詞的過程變得至關重要。由於詞彙之間沒有像英文單詞一樣的明確邊界，斷詞的正確性直接影響到後續分析的準確性。一個合理、準確的斷詞結果可以提供有意義的詞彙單位，有助於進行詞性標記、語義分析和文本挖掘等工作。因此，中文斷詞[6]的過程變得格外重要。

相對於中文，英文句子的結構相對簡單。英文以空格作為單詞的邊界，單詞之間有明確的分隔符號，這使得詞的擷取相對較簡單。然而，中文的複雜結構和詞彙之間的模糊邊界使得斷詞變得更加具有挑戰性。

在現代資訊處理和自然語言處理的應用中，中文斷詞的技術發展至關重要。為了提高斷詞的正確性，研究者們開發了各種基於統計和機器學習的斷詞算法，並不斷優化演算法以應對中文語境的多樣性。這包括基於詞典的方法、基於統計模型的方法和基於深度學習的方法等。

* + 1. 中研院CKIP中文斷詞系統

CKIP中文斷詞系統（Chinese Knowledge and Information Processing Group，中文簡稱CKIP）[12]是一個由中央研究院資訊科 學研究所所開發的中文斷詞系統。該系統以中文語言處理為核心，致 力於解決中文斷詞的挑戰，並在自然語言處理領域中發揮了重要作用。

CKIP系統採用了統計和規則結合的方法，以提高斷詞的精準度和 效能。它基於大量的中文文本資料進行訓練，不僅考慮了詞彙之間的 統計特徵，還充分利用了語言學上的規則。這種綜合的方法使得CKIP 系統能夠更好地應對中文語言的多樣性和複雜性。

* + 1. Jieba斷詞

Jieba（結巴）是一個基於Python語言的中文斷詞系統，廣泛應用於 自然語言處理領域。其獨特之處在於結合了基於統計的詞頻方法和基 於規則的字典匹配方法，提供了高效且準確的中文斷詞解決方案。

Jieba的斷詞算法主要基於統計模型，通過分析大量的中文語料庫，學習詞彙的出現頻率和詞彙之間的概率關係。同時，Jieba還引入了HMM（Hidden Markov Model）模型，有效處理了一些新詞或罕見詞的斷詞難題。此外，Jieba支援使用者自定義詞典，可以根據特定應用領域的需求，添加專有名詞或新詞，提升斷詞的個性化效果。

* 1. 關鍵詞提取
     1. TDIDF

TF-IDF（Term Frequency-Inverse Document Frequency）[2]是一種 廣泛 應用於自然語言處理和文本探勘的詞權重算法，其全名即代表了 該算法的 核心概念：詞頻（TF）和逆文檔頻率（IDF）。這種算法通過 綜合考慮詞在 文檔中的出現頻率以及在整個文檔集合中的重要性，為 每個詞賦予一個權 重，揭示了詞在文本中的相對重要性。

詞頻（TF）衡量了一個詞在特定文檔中的出現頻率，其計算方式是該 詞在文檔中出現的次數除以文檔的總詞數。這一步驟反映了在單一文 檔中詞的相對重要性，但並未考慮到該詞在整個文檔集合中的普遍程 度。

逆文檔頻率（IDF）則評估了詞在整個文檔集合中的重要性。IDF的計 算方式是總文檔數除以包含該詞的文檔數，再對這個比值取對數。這 一步驟的目的在於找出在整個文檔集合中相對罕見但有助於區分文檔 的詞，賦予其較高的權重。透過這樣的計算方式，IDF強調了那些在較 少文檔中出現的詞彙，提供了一種區分性的度量。

TF-IDF的計算方式是將詞頻和逆文檔頻率相乘，這樣的操作將突顯在 特定文檔中高頻出現且在整個文檔集合中相對罕見的詞。這些詞的權 重較高，因此在文本分析和檢索中具有更大的重要性。

TF-IDF算法的應用不僅僅限於文本檢索，還廣泛運用於文本分類、關 鍵詞提取、信息檢索等領域。在處理大量文本數據時，TF-IDF不僅提 供了一種有效的特徵提取方式，還有助於捕捉文本之間的關聯性和重 要性。總體而言，TF-IDF作為一種權威的詞權重算法，為文本處理和 理解提供了強大的工具，同時也促進了相關領域的不斷發展和深化。

* + 1. Textrank

TextRank[1]是一種用於自動關鍵詞提取的演算法，它基於圖的排序算 法，最初由Mihalcea和Tarau於2004年提出，並在NLP（自然語言 處理）領域中得到廣泛應用。這種演算法的靈感來自於PageRank，一 種用於分析網頁的排名算法。

TextRank的核心思想是將文本表示為一個圖，其中每個節點表示文本 中的詞，而邊表示詞之間的關係。這種圖的構建過程通常使用詞彙窗 口和共現矩陣，以確定詞之間的關聯性。然後，基於這個圖結構，使 用圖的排名算法來確定每個詞的權重，這些權重用於衡量詞的重要性。

* 1. 大型語言模型(Large Language Model)

大型語言模型（Large Language Model)[14]是一種擁有龐大參數量和複雜結構的語言模型。這些模型被設計用來理解和生成自然語言，並在各種自然語言處理任務中表現出色。它們基於深度學習技術，特別是Transformer架構。

大型語言模型的參數量非常巨大，通常在數十億到數千億個參數之間，如近期最紅的GPT-3.5擁有2000億參數。這使得這些模型能夠學習更複雜的語言模式和語境。

這些模型通常使用預訓練的方式進行訓練。在預訓練階段，它們在龐大的文本數據集上學習通用的語言表示。然後，在微調階段，模型使用特定任務的標記數據進行微調，以提高在特定任務上的性能。

大型語言模型能夠理解語言的上下文，包括句子和段落的結構。這種上下文感知的能力有助於提高模型在理解語境相關性方面的性能。

* + 1. Transformer

大型語言模型的核心是Transformer[3]（請見圖2.1），這是一種由 Google於2017年提出的神經網絡架構。Transformer的技術是基於注 意力 （attention）與自我注意力（self-attention）[3]，基本以取 代CNN及RNN，在訓練Transformer時大幅減少，這使得機器翻譯的 發展快速躍進。

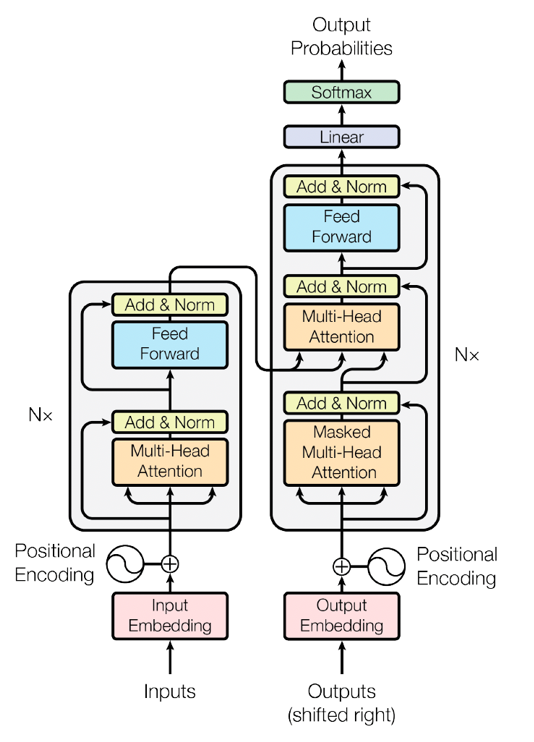


圖2.1，Transformer架構圖

* + 1. 自注意機制(Self-Attention Mechanism)

自注意力機制（Self-Attention Mechanism）[3] （請見圖2.2）是一 種深度學習模型中的關鍵技術，它在處理序列型數據時取得了顯著的 成功。最初由Vaswani等人於2017年提出，自注意力機制旨在解決傳 統序列模型中面臨的長距離依賴問題，同時提高模型對於不同位置信 息的感知。

Transformer使用自注意機制來處理輸入序列中不同位置的信息。這 使得模型能夠在處理每個單詞或標記時考慮其他所有位置的上下文。 自注意機制的引入增強了模型對長距離相依性的建模能力。

自注意力機制的核心思想是通過賦予不同位置的詞不同的權重，使模型能夠在處理輸入序列時更加靈活地分配注意力。以下是自注意力機制的基本運作原理：

1. 詞向量的線性轉換：首先，將輸入的詞向量（或稱為詞嵌入）通過線性轉換映射到三種不同的向量：輸入的詞向量（Query）、輸入的詞向量（Key）和輸入的詞向量（Value）。這三種向量的轉換是通過不同的權重矩陣實現的。

2. 計算注意力分數：接下來，計算每個詞對於其他詞的注意力分 數。這是通過計算Query和Key之間的點積來實現的，並將結果除以 一個維度的平方根（通常是詞向量的維度），以確保梯度的穩定性。這 樣的計算過程捕捉了每對詞之間的相似性，這種相似性被表示為注意 力分數。

3. 計算加權和：將每個詞的Value向量乘以對應的注意力分數， 然後對所有詞的加權Value向量進行求和，得到最終的表示。這個表 示保留了對不同位置的詞的重要信息，並且在此過程中，每個詞的權 重由它與其他詞的相似性動態調整。

4. 多頭自注意力機制：為了進一步提高模型的表達能力，可以使 用多個自注意力機制組合成多頭注意力機制。每個頭注意力機制都學 習一種不同的詞彙表示，最後將這些多頭的結果進行線性映射和連接。

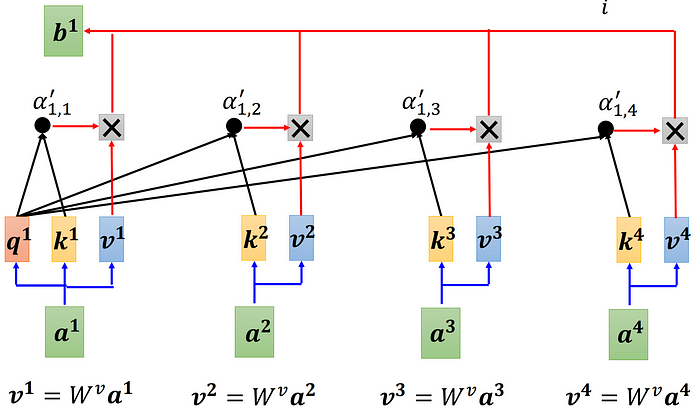


圖2.2，Attention架構圖

* + 1. 多頭注意力（Multi-Head Attention）

多頭注意力（Multi-Head Attention）是 Transformer 模型中的一個 關鍵概念，用於提高模型對序列數據的建模能力。多頭注意力引入了 多組不同的權重，每一組被稱為一個注意力頭。這種設計允許模型同 時學 習不同的關係，以更全面地理解輸入序列。在計算過程中，每個 注意力頭 都會生成一組加權和，這使得模型能夠在不同的注 意力空間中進行學習。

另一個重要特點是每個注意力頭都有自己的權重矩陣，用於對輸入進 行線 性投影。這意味著每個頭都可以專注於學習特定的特徵表示，而 不受其他 頭的影響。這種獨立的投影使得每個注意力頭能夠自主地捕 捉序列中的信 息。

多頭注意力的輸出會被組合在一起，通常通過串聯或加權和的方式。 這種組合確保了模型在多個注意力頭的協同作用下獲得更豐富的特徵 表示。這種設計使得模型能夠同時關注序列中的不同部分，捕捉不同 位置之間的依賴性。這種並行處理的方式有助於提高模型的訓練速度， 同時增強了模型的表達能力。

* + 1. 位置編碼（Positional Encoding）

在Transformer模型中，由於它是基於注意力機制的架構，缺少了序 列中詞的順序信息。為了將詞的相對位置納入考慮，引入了位置編碼 （Positional Encoding）。

位置編碼的主要目的是為模型提供一種關於輸入序列中詞的相對位置 的信息。由於注意力機制是無法區分詞的順序的，因此在不引入位置 編碼的情況下，模型無法理解序列中詞的先後順序。位置編碼通過在 詞向量中添加一些額外的信息，來表示詞的相對位置。

* + 1. 預訓練與微調(Pre-training and Fine-tuning)

預訓練(Pre-training)[15]是一種訓練模型前的準備動作，隨機給定 權重並在沒有任何經驗的情況下開始訓練。此種訓練方式建立在非常 大量的數據上，且訓練需要幾周的時間。微調(Fine-tuning)[15]是在 模型經過預訓練後完成的訓練，模型可以進行微調以適應特定任務。 在微調階段，模型使用標記的數據（有標籤的數據）進行進一步的訓 練，以提高在特定任務上的性能。預訓練與微調這樣的組合也被稱為 遷移學習（Transfer Learning），通過在大規模的通用任務上進行預 訓練，模型能夠學到通用的特徵，然後通過微調，這些特徵可以被轉 移到特定任務上，從而提高模型在少量樣本或特定任務上的性能。

* + 1. 生成和解碼

大型語言模型（LLM）的生成和解碼是指模型如何生成自然語言文本以 及如何解碼（生成）特定的輸出序列。這涉及到語言模型的基本運作， 尤其是生成序列的過程。

生成:語言模型的上下文中，生成通常指的是模型生成符合給定條件 （如輸入文本或特定任務要求）的自然語言文本。生成可以是有條件 的，即模型基於給定的上下文或信息生成文本。例如，給定一個開頭， 模型可以生成一個完整的句子或段落。

解碼:解碼是指在模型生成文本時選擇合適的單詞或子詞的過程。在生 成的過程中，模型需要從詞彙表中選擇應該出現在序列中的下一個元 素。解碼過程通常是基於機率分佈，模型會根據其內部學到的權重選 擇最可能的下一個詞。

* 1. GPT

生成式預訓練模型（Generative Pre-trained Transformer，簡稱GPT）[4][5][13]是由OpenAI開發的一系列自然語言處理模型，代表著當前深度學習領域中最先進的語言生成技術之一。GPT系列的模型以Transformer架構為基礎，通過大規模預訓練和微調的方式，具有強大的語言理解和生成能力，被廣泛應用於文本生成、機器翻譯、對話生成等多個NLP任務。

GPT模型的獨特之處在於其預訓練的方式。與傳統的監督式學習不同，GPT通過大規模的無監督預訓練來學習語言的統計結構和語境。模型通過預測下一個詞的方式，透過海量的文本數據訓練自身，學會了豐富的語言知識和上下文的把握能力。這種預訓練的方式使得GPT系列模型具有極高的通用性，能夠處理各種語言任務而無需重新訓練。

GPT使用Transformer架構，這種注意力機制使得模型能夠高效處理長文本序列，捕捉詞與詞之間的複雜依賴關係。這種能力使得GPT成為一個強大的生成式模型，能夠生成連貫且有意義的文本。

該模型在文本生成、文章摘要等任務上表現卓越，其擅長處理不同主題和風格的文本，並且能夠生成具有上下文一致性的長文本。這使得GPT在自然語言處理領域取得了顯著的成就。

GPT系列還包括了多個版本，如GPT-2和GPT-3，隨著模型規模的逐步擴大，性能也進一步提升。其中，GPT-3是目前模型參數規模最大的之一，擁有1750億個參數，這使得它具有更強大的表達能力和泛化能力。這種龐大的模型參數使得GPT-3在理解語言上更為深入，能夠處理更複雜的任務和更龐大的語料庫。

1. **系統架構**
   1. 系統架構說明

本次實驗以N篇文章作為輸入，首先將文章進行斷詞處理，將句子轉換為單詞的集合。接著，使用TFIDF（Term Frequency-Inverse Document Frequency）過濾關鍵詞的演算法，對斷詞後的單詞集合進行篩選和過濾，選出最具代表性的單詞。

在獲得關鍵詞後，通過注意力機制，系統可以選擇特別被注意的其中幾個關鍵詞，這些關鍵詞可能在文章中扮演著重要角色，進而有助於後續的文本摘要生成。

最後，透過ChatGPT生成具有前後文關係的摘要。ChatGPT是一種基於GPT架構的語言模型，能夠根據給定的文本生成具有一定邏輯性和連貫性的新文本。因此，透過ChatGPT生成的摘要將具有與原文相關的語義和內容，能夠幫助用戶更好地理解文章的核心內容。

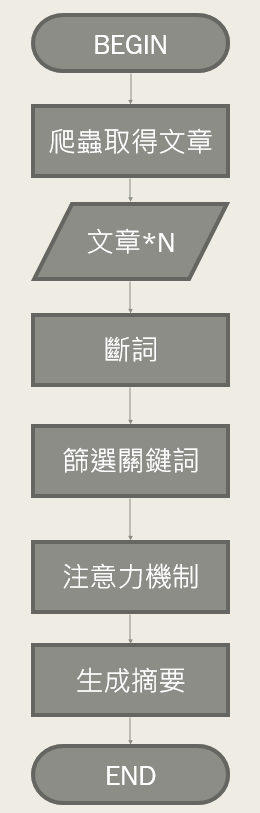


圖3.1，系統流程圖

* 1. 資料集

本研究透過網路爬蟲抓取網路上的新聞文章，以財經文章為主要研究對象。選擇udn、yahoo、天下雜誌三大新聞平台的文章作為研究目標，主要考量是這三大平台均有豐富的財經新聞資源，且是廣大網友熟悉且常使用的新聞閱讀平台。因此，選擇這三個平台作為爬取目標，有助於取得多元、廣泛的財經新聞資訊，提升研究的可信度與資料豐富度。

* 1. 資料前處理
     1. CKIP斷詞

選擇使用CKIP作為斷詞系統的原因之一是因為其針對繁體中文的支援相對較為穩定。由於CKIP是由中研院所開發的系統，因此在處理繁體中文時具有較高的準確性和可靠性。相比之下，Jieba是來自中國的斷詞系統，對於繁體中文的處理可能會有一些限制。

此外，CKIP相對於Jieba來說，功能更加全面，能夠更好地處理未知詞彙，這對於文本分析來說是非常重要的。因此，選擇CKIP作為斷詞系統可以更好地支援後續的文本分析工作，提高系統整體的效率和準確性。

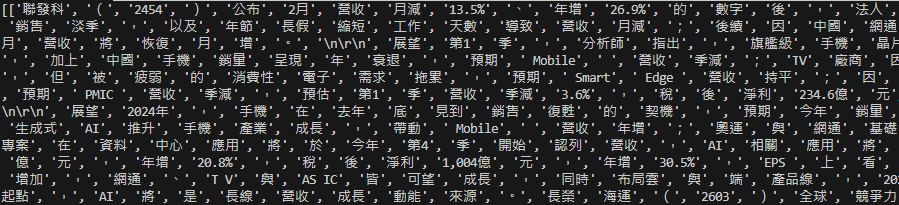


圖3.2，CKIP斷詞結果



圖3.3，系統流程圖

* + 1. 過濾停用詞

在自然語言處理中，斷詞是一個重要的步驟，它將一段文字切割成有意義的詞彙，方便後續的處理和分析。在斷詞的過程中，通常會將標點符號去除，因為它們對於詞彙的理解並沒有太大的幫助。不過，在某些特定情況下，保留標點符號也是必要的，例如在分析詞彙之間的關係時，標點符號可能包含重要的資訊。

除了標點符號外，還需要去除停用詞。停用詞通常是指在自然語言中使用頻率很高，但在文字內容分析中通常無需考慮的詞彙，例如代名詞、助動詞、介係詞、連接詞等。透過過濾停用詞，可以讓文字內容更加簡潔、易於分析。

建立停用詞列表的方法可以是透過建立字典，將不需要的詞彙逐一加入其中。然而，由於語言是活潑的，詞彙不斷變化，因此停用詞列表也需要不斷更新和維護。一種方法是透過詞性來選擇停用詞，因為大多數的停用詞都屬於特定的詞性，例如代名詞屬於名詞類、助動詞屬於動詞類等。利用詞性標註工具如CKIP可以幫助我們標註詞彙的詞性，進而選擇過濾掉哪些詞性的詞彙，使得分析更加精準。

3.4 關鍵詞提取

經由斷詞後，我們得到了單詞資料集。在本實驗中，我們將使用TFIDF進行單詞篩選。TFIDF通過詞頻來確定每個單詞的關鍵分數。我們將計算出每個單詞的分數，並將其排序，然後取前五十個作為後續注意力機制的輸入。



圖3.4，TFIDF輸出結果示意圖

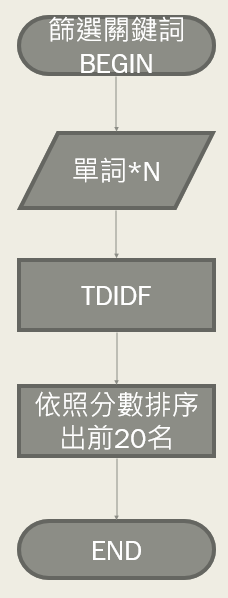


圖3.5，系統流程圖

3.5 詞向量轉換

經由TFIDF篩選出單詞後，需要將這些詞轉換成向量，讓注意力機制演算法能夠另一種形式看懂我們想要輸入的單詞，本實驗所使用的詞向量轉換工具為Word2Vec:

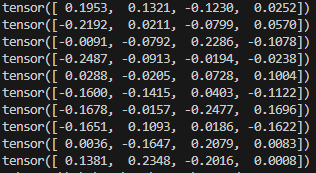


圖3.6，詞向量轉換結果示意圖

3.6 注意力機制

將文字經由詞向量轉換後，隨機產生兩組Q、K權重向量與詞向量做內積，因此會得到Q與K，在將Q與K內積得到阿法，本實驗對於Attention的做法僅到這部分，得到阿法後將阿法做排序，並取出前三十個被注意的詞，為最終將生成摘要的關鍵詞。

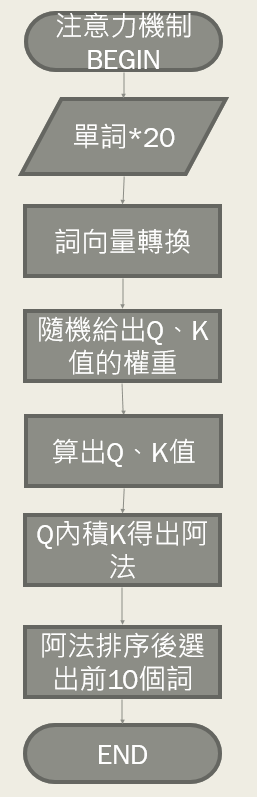


圖3.7，注意力機制流程圖

1. **實驗分析**
   1. 實驗方法
      1. 關鍵詞的數量

本研究透過控制關鍵詞的數量觀察摘要生成的狀況，關鍵詞給的越多 文章就越完整，贅詞減少與前後文關係更明確，並且十個詞的文章明 顯內容空洞，隨著詞的增加，文章的可參考性也隨之變高。然而在第 三十~五十個詞所產生的摘要來看，變化是越來越小，我們並不希望 用過多的詞影響摘要的品質，摘要總共兩百個字，若使用到五十個詞 也占據了整篇摘要的四分之一，違背自動摘要的真諦；因此本實驗取 中間值三十個詞作為生成詞數。從十個到五十個的結果如下圖:

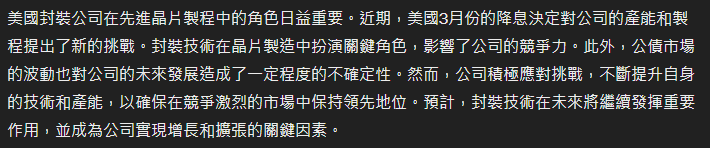


圖4.1，十個關鍵詞

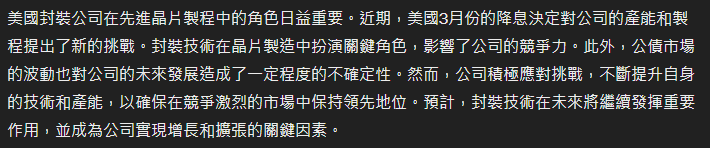


圖4.2，二十個關鍵詞

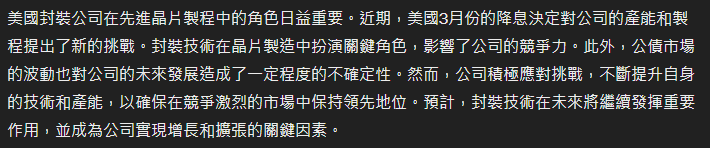


圖4.3，三十個關鍵詞

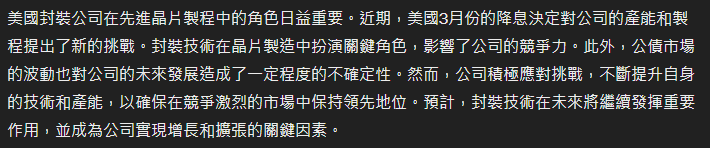


圖4.4，四十個關鍵詞

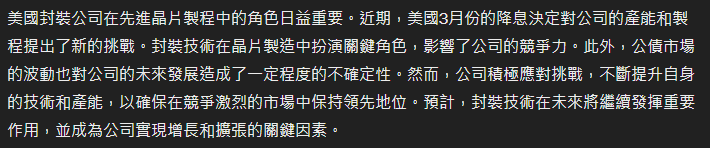


圖4.5，五十個關鍵詞

* 1. 實驗結果與評估
     1. ROUGE-N

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[16]為評估自動摘要的一種指標，通過自動生成的摘要與人工生成的參考摘要進行比較計算，得出分數以衡量兩者之間的相似度。ROUGE也衍伸出許多分類，本研究用來評估的方式為ROUGE-N。常見的為ROUGE-1與ROUGE-2，N即為單詞數，ROUGE-N簡單還說就是「產生一句句子，而句子裡頭的單字能夠召回多少個原本的Reference的單字」:

ROUGE-1(適合分析較短的摘要，亦是本研究所採用方式):

C1對應到R1的詞:共出現三次(police、the、gunman)

C1對應到R2的詞:共出現三次(the、gunman、police)

分母：R1、R2的字數相加

以上ROUGE-1的結果為:ROUGE-1(C1) = (3+3)/(4+7) = 6/11



圖4.6，Reference(人工標注)示例

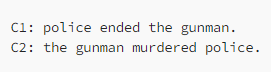


圖4.7，自動摘要(程式生成)示例

ROUGE-2(適合分析單一文件):

以上ROUGE-2的結果為:ROUGE-1(C1) = (1+1) / (3+6) = 2/9

C1對應到R1的詞:共出現一次(the gunman)

C1對應到R2的詞:共出現一次(the gunman)

分母：R1、R2的字數相加

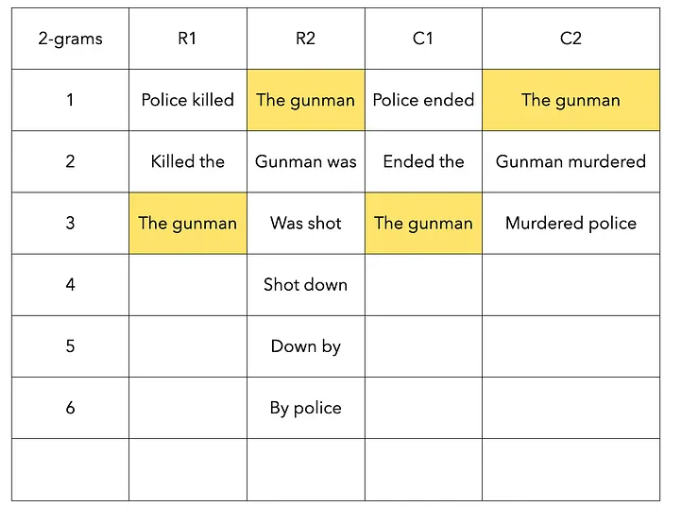


圖4.8，2-grams示例

* + 1. 自動生成之摘要：「最近台股市場出現了大跌，讓投資人感到擔憂。投資人可以考慮投資基金或債券來分散風險。投信在募集資金時可能會配息，而利率持續走高可能會影響投資方式。公司股票的籌碼情況也很重要，美國市場指出了質押問題可能導致股市震盪。目前資產配置是一個困難的問題，投資人需要謹慎選擇投資的方式。團隊合作對於投資人來說是很重要的。」

實驗者依照生成摘要之原文章人工生成摘要：「近期台股市場出現大跌，投資人對於台股逢高檔紛紛感到擔憂，基金經理人也藉此呼籲大眾可分配資金至基金債券方散風險，以穩定配息的方式保護短線上的振幅，近期利率的走勢也是個投資人關注的議題，利率的走勢直接影響了美國股市的震盪；大多投資人認為目前在市場的資產配置是很困難的，需要謹慎選擇投資標的。」

Rouge-1之結果：召回率為0.667、精準度為0.646、F1\_Score為0.656

來自其他系統繩成之自動摘要：「最近台灣股市出現了顯著的波動，先是大幅下跌，然後又大幅上漲，市場波動性增加。分析師指出這些波動的原因包括台灣股市距離兩萬點關卡的接近、美國股市收益的影響、美國可能的銀行倒閉以及美聯儲的利率決定。儘管美國經濟放緩，台灣的基本面仍然強勁，尤其是在高性能計算、人工智慧和雲端服務等產業上。」

實驗者依照生成摘要之原文章人工生成摘要：「近期台股市場出現大跌，投資人對於台股逢高檔紛紛感到擔憂，基金經理人也藉此呼籲大眾可分配資金至基金債券方散風險，以穩定配息的方式保護短線上的振幅，近期利率的走勢也是個投資人關注的議題，利率的走勢直接影響了美國股市的震盪；大多投資人認為目前在市場的資產配置是很困難的，需要謹慎選擇投資標的。」

Rouge-1之結果：召回率為0.354、精準度為0.34、F1\_Score為0.347

1. **結論**
   1. 研究總結

本論文研究主要因應如今資訊爆炸的年代，人們無法系統性的統整資料好讓自己看到需要的資訊，因此提供一個快速瀏覽的系統讓人們在翻閱資料時可以快速有效的知道重點。本論文結論有以下:

1. 本研究在關鍵字擷取的部分使用了TFIDF之方法，以TDIDF為基之方法可篩選出文件中頻率較高之單詞。
2. 本研究除了篩選關鍵詞外，還多了一層注意力機制，讓系統自動尋找這些關鍵詞中被注意的詞，這些詞即會是精華中之精華，利用這些詞生成出的摘要會乾淨簡潔許多。
3. 本研究利用近期最新的GPT3.5語言模型協助生成具有前後文關聯的摘要，方便人們進行閱讀。
4. 本研究主要以中文文件為主要研究對象，故不是用於英文或其他對象之解析。
   1. 未來展望

針對本研究之自動摘要技術，將提出以下幾項未來可改善的議題：

1. 希望即使在關鍵詞數量稀少的情況下也能拼湊出摘要
2. 前後文能夠更順暢
3. 更能順應目前時事拼湊出相對應的摘要

**參考文獻**

1. Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.(textrank)
2. G. Salton, C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management, Vol. 24, No. 5, 1988.(文字加權)
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” In Advances in Neural Information Processing Systems, pages 6000–6010.(注意力機制)
4. Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.(GPT1)
5. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.(GPT2)
6. Huang, SL., Chen, KJ., Ma, WY. et al. Semantic relation identification for consecutive predicative constituents in Chinese. lingua. sin. 3, 9 (2017).(中文斷詞)
7. Spetka, Scott. The TkWWW Robot: Beyond Browsing. NCSA. [21 November 2010].(網路爬蟲)
8. Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.(詞向量轉換)
9. Y. Zhang, M. Chen and L. Liu, "A review on text mining," 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2015, pp. 681-685, doi: 10.1109/ICSESS.2015.7339149.(文字探勘介紹)
10. Koopman, Philip. “How to Write an Abstract.” (2009).(摘要種類)
11. H. M. Mahedi Hasan, F. Sanyal, D. Chaki and M. H. Ali, "An empirical study of important keyword extraction techniques from documents," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, 2017, pp. 91-94, doi: 10.1109/ICISIM.2017.8122154.(關鍵詞提取)
12. Design of CKIP Chinese Word Segmentation System Wei-Yun Ma, Keh-Jiann Chen IJALP, Vol. 14, No. 3, pp. 235–249, May 2004(CKIP)
13. Proceedings of the 34th International Conference on Neural Information Processing SystemsDecember 2020Article No.: 159Pages 1877–1901(GPT3)
14. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.(LLM)
15. Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang, Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu, and Haifeng Wang. 2019. D-NET: A Pre-Training and Fine-Tuning Framework for Improving the Generalization of Machine Reading Comprehension. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 212–219, Hong Kong, China. Association for Computational Linguistics.(預訓練)
16. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.(ROUGE)