

自動摘要系統優化結案報告

陳育婷 Yu Ting Chen

Table of Content

1. August : Micro Revise for IASL and UDN
 1. Modify Kernel Functions
 2. Modify User Interface
 3. Others – Conjunctions, Compression, Code Refactoring
2. September – Mid October : User Interface Design for UDN
 1. Html, Css Design
 2. Flask & Javascript Implement
 3. UI & Algorithm Improvement – Highlight, Paragraph Display, Coherent, Sentence Compression -> 5 levels
3. Late October : Major Algorithm Improvements
 1. SQL Setting
 2. Generate & Integrate Keywords into Summary Generation

August :

Micro Revise for IASL and UDN

Original Result of the Example

PIESim 摘要生成系統 (designed by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的

使用者輸入-須加入重點(輸入單位:片段、單詞/需以換行分隔)

使用者輸入-不須加入重點(輸入單位:片段、單詞/需以換行分隔)

系統摘要

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，號召人的小約成眾矢之的；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，目前處境下，小約參加美網也打上問號，從3月開始賽季就停擺，ATP日前宣布修訂世界排名積分系統。

系統壓縮摘要

塞爾維亞世界球王巡迴賽，儘管檢測結果是陰性，心傷難癒，球員巡迴賽中擁抱一起打籃球，在夜店跳舞狂歡，流行期間受到許多批評，加上爆出名球員染疫，號召人的小約成眾矢之；小約事發發出聲明表示深深致歉，但他提到，他認為項計畫若是失敗，外界找名氣大的當罪魁禍首，目前處境下，參加美網打上問號，從3月開始賽季就停擺，ATP宣佈積分系統。

資訊覆蓋度 使用者輸入重點重要度 使用者輸入非重點重要度

1.0

0

0.0

位置重要度

冗餘去除重要度

0.1

0.3

字數

150

更新

1. Kernel – 1. 用標題資訊

- 標題資訊囊括入選擇考量之一，與其他指標分數合併，選最終整體分數高的部分 -> 選到獵巫

PIESim 摘要生成系統 (designed by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的訊息中，很多不只是批評；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，這就是自己目前狀態。

必要關鍵詞(句)

省略關鍵詞(句)

相關詞(句)覆蓋強度 必要關鍵詞(句)強度 省略關鍵詞(句)強度

最強 無 無

記憶相似強度 第一段落強度 省略重複詞(句)強度

無 中 弱

字數

150

更新

1. Kernel – 2. 以”原始” ”詞”為基礎做配對

- 以字為單位做配對 -> 以詞為單位做配對
- 不用EhowNet mapping表現較好 (以標題為答案做測試)

●Revised Formula:

$$s_{rep1} = sentrep_{term(char.)}(s_1), s_{rep2} = sentrep_{term(char.)}(s_2) \quad (4.11)$$

$$PIESim(s_1, s_2) = \frac{1}{3} \left(\frac{m_w}{\sum_{\{w_i | (t_i, w_i) \in s_{rep1}\}} w_k} + \frac{m_w}{\sum_{\{w_j | (t_j, w_j) \in s_{rep2}\}} w_k} + \frac{m_w - \frac{|t|}{2}}{m_w} \right) \quad (4.12)$$

$$m = \langle (c_{1i}, c_{2j}) | c_{1i} = c_{2j} \wedge |p_{1i} - p_{2j}| \leq \left\lfloor \frac{\max(|s_{rep1}|, |s_{rep2}|)}{2} \right\rfloor - 1 \wedge i \notin \bigcup_{k=1}^{i-1} k \rangle \quad (4.13)$$

$$m_w = \sum_{\substack{\{(w_{1i}, w_{2j}) | (c_{1i}, c_{2j}) \in m \wedge (c_{1i}, w_{1i}) \in s_{rep1} \wedge \\ (c_{2j}, w_{2j}) \in s_{rep2}\}}} \frac{w_{1i} + w_{2j}}{2} \quad (4.14)$$

$$m_{order} = \langle (u_{1k}, u_{2k}) \in m \wedge \exists 1 \leq i_1 \leq i_2 \leq \dots \leq l \rangle_{s.t.} \quad (4.15)$$

$$\forall (u_{1k}, u_{2k}), u_{1k} = c_{1i_k}, u_{2k} = c_{2i_k}$$

$$t = \langle (u_{1k}, u_{2k}) | (u_{1k}, u_{2k}) \in m_{order} \wedge u_{1k} \neq u_{2k} \rangle$$

c_{li} : unit i in s_l , p_{li} : position of c_{li} in s_l

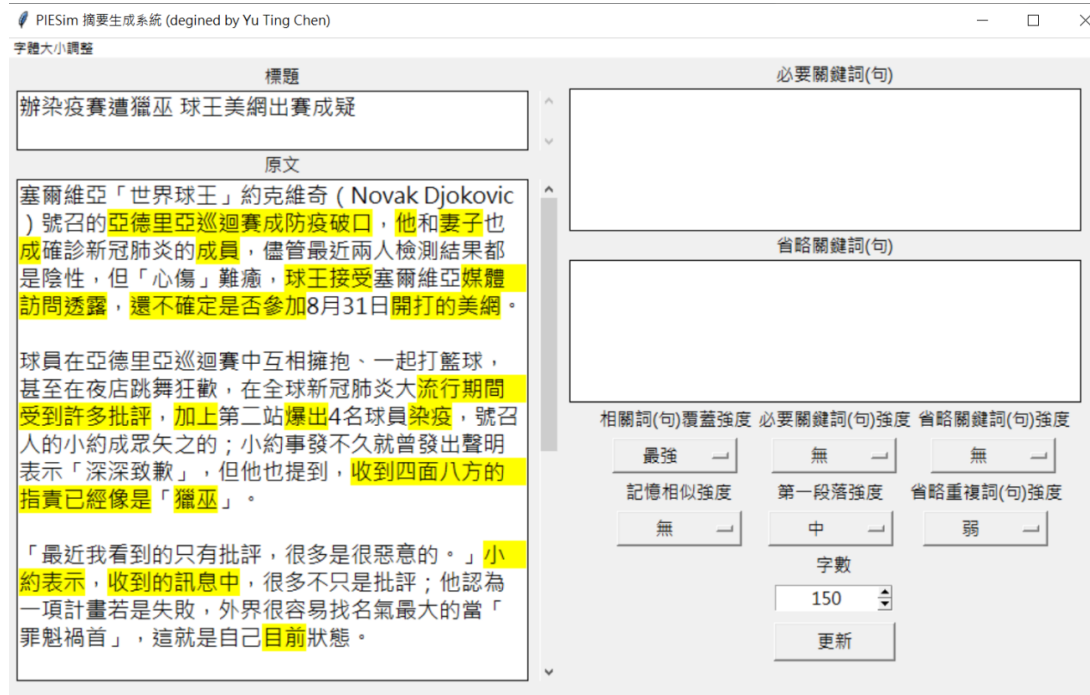
1. Kernel – 3. 位置重要度加入段落考量

- 原：
$$pos_i = 1 - \frac{(i-1)}{|\{s|s \in \{sent.s \text{ in document } d\}\}| - 1}$$

- 後：

$$pos_j = 1 - (j - 1) * 0.5 * posweight * \frac{1}{|\{p|p \in \{paragraph.s \text{ in doc}\}\}| - 1}$$
$$pos_{ji} = \max(pos_j - (i - 1) * \frac{1}{|\{s|s \in \{sent.s \text{ in doc}\}\}| - 1}, 0.1)$$

-> 相對於過度重視前面的部分(p.4)，會考慮各段落重要部分



2.User – 1. 方格、量度討論後重新命名

PIESim 摘要生成系統 (dedigned by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的訊息中，很多不只是批評；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，這就是自己目前狀態。

必要關鍵詞(句)

省略關鍵詞(句)

相關詞(句)覆蓋強度 必要關鍵詞(句)強度 省略關鍵詞(句)強度

最強 無 無

記憶相似強度 第一段落強度 省略重複詞(句)強度

無 中 弱

字數

150

更新

2.User – 2. 強度輸入部分改成5程度下拉式選單

PIESim 摘要生成系統 (degined by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的訊息中，很多不只是批評；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，這就是自己目前狀態。

必要關鍵詞(句)

省略關鍵詞(句)

相關詞(句)覆蓋強度 必要關鍵詞(句)強度 省略關鍵詞(句)強度

最強 無 無

無 第一段落強度 省略重複詞(句)強度

弱 中 弱

中

強

最強

字數

150

更新

2. User – 3. 刪除摘要輸出方框，直接在原文中標示壓縮摘要選到的片段

PIESim 摘要生成系統 (deigned by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞「世界球王」約克維奇 (Novak Djokovic) 號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的訊息中，很多不只是批評；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，這就是自己目前狀態。

必要關鍵詞(句)

省略關鍵詞(句)

相關詞(句)覆蓋強度 必要關鍵詞(句)強度 省略關鍵詞(句)強度

最強 無 無

無 強度 第一段落強度 省略重複詞(句)強度

弱 中 弱

字數

150

更新

2. User – 4. 將壓縮沒選到重要詞輸入必要關鍵詞方框，會產生含該詞的壓縮摘要

PIESim 摘要生成系統 (designed by Yu Ting Chen)

字體大小調整

標題

辦染疫賽遭獵巫 球王美網出賽成疑

原文

塞爾維亞 世界球王「約克維奇 (Novak Djokovic)」號召的亞德里亞巡迴賽成防疫破口，他和妻子也成確診新冠肺炎的成員，儘管最近兩人檢測結果都是陰性，但「心傷」難癒，球王接受塞爾維亞媒體訪問透露，還不確定是否參加8月31日開打的美網。

球員在亞德里亞巡迴賽中互相擁抱、一起打籃球，甚至在夜店跳舞狂歡，在全球新冠肺炎大流行期間受到許多批評，加上第二站爆出4名球員染疫，號召人的小約成眾矢之的；小約事發不久就曾發出聲明表示「深深致歉」，但他也提到，收到四面八方的指責已經像是「獵巫」。

「最近我看到的只有批評，很多是很惡意的。」小約表示，收到的訊息中，很多不只是批評；他認為一項計畫若是失敗，外界很容易找名氣最大的當「罪魁禍首」，這就是自己目前狀態。

必要關鍵詞(句)

塞爾維亞

省略關鍵詞(句)

相關詞(句)覆蓋強度 必要關鍵詞(句)強度 省略關鍵詞(句)強度

最強 無 無

記憶相似強度 第一段落強度 省略重複詞(句)強度

無 中 弱

字數

150

更新

3. Others

1. Consider conjunctions in summary generation
2. Separate Preprocessing procedure for traditional and simplified Chinese
3. Code, Layout Refactoring, Encapsulate in a Class


September – Mid October:
User Interface Design for UDN

1. Html, Css Design

- 外觀按照聯合報UI設計師提供的介面設計
- Html 方框類型以及 元素層級則由自己學習及設置



1. Html, Css Design - Demo

 AI Demo 2020

TTS寫作輔助新聞摘要One Piece 2.0新聞小幫手

新聞摘要

新聞標題

原文

[重置](#)[摘要](#)

摘要

☒ 進階設定

☐ 必要關鍵詞

請輸入必要關鍵詞

強度 無○●○●○高

☐ 排除關鍵詞

請輸入排除關鍵詞

強度 無○●○●○高

☒ 多次提及強度 無○●○●○高

☐ 首段文字強度 無○●○●○高

☐ 省略重複詞(句)強度 無○●○●○高

☒ 句子壓縮強度 無●○●○●○高

☐ 指定字數

請輸入字數

2. Flask & Javascript Implement

- 網頁互動部分用Javascript撰寫
 1. 進階設定按鈕互相連動綁定
 2. 摘要對應原文螢光標示
 3. 使用者行為動作觸發
- 透過 Flask 搭建本地伺服器回傳摘要，在瀏覽器上顯示
 1. 建立Router 導引到不同互動對應之不同功能
 2. Ajax 不同步更新資訊減少重載

2. Flask & Javascript – Implement Demo

- 使用者按下摘要按鈕後Javascript與Flask串接的程式

```
Summary > static > js > JS PIESim_interaction.js > ($) callback > click() callback > done() callback
75  $("button[type = submit]").click(function(event){
76      $.ajax({
77          type : 'POST',
78          url: '/abc',
79          data:{
80              text : $("div#context").text(),
81              titletext: $("textarea#title").val(),
82              querytext : $("textarea#necinput").val(),
83              querynotext : $("textarea#notnecinput").val(),
84              charnumval : $("textarea#charnuminput").val(),
85              queryval : $("input[type = 'radio'][class='neckkeywords[]']:checked").val(),
86              querynoval : $("input[type = 'radio'][class='notneckkeywords[]']:checked").val(),
87              centval : $("input[type = 'radio'][class='repeatmentionls[]']:checked").val(),
88              positionval : $("input[type = 'radio'][class='firstparals[]']:checked").val(),
89              redundantval : $("input[type = 'radio'][class='redundantls[]']:checked").val(),
90              compressionval : $("input[type = 'radio'][class='compressionls[]']:checked").val()
91          }
92      })
93      .done(function(data){
94          if (data.summary){
95              $("textarea#summary").val(data.summary);
96              alert('摘要產生!');
97          }
98      });
99  });
```

3. UI & Algorithm Improvement

1. 顯示摘要時按原始段落分割

摘要中所選相連片段若在原始文章中跨段落，則也會同步在摘要中跨段落顯示 -> 使用者理解及修改更容易。

2. 標示原文中被選到段落bugs修正

調整演算法與外觀設計，多次除錯，使壓縮部分也可以正確標示在原文。

3. 增強摘要通順度

考慮每句中結尾與開頭的詞與詞性，設計遞迴加入相鄰句的條件。

4. 壓縮分等級

設計不同層級的壓縮演算法，原本的壓縮為最高層級，最低層級的壓縮僅會刪除極少數部分(可能為標點符號、語助詞...)

3. Demo – ATS Software

- 前往該網址觀看：

<https://drive.google.com/file/d/1g0ilioDB3sdXrWh13hZInnOuoIwE3Zzl/view?usp=sharing>

Late October :
Major Algorithm Improvements

1. SQL Setting

- 根據聯合報資料庫應用軟體 MySQL ->將UDN2017-2018 Data由原來的csv轉存至MySQL或任何一種SQL軟體
- 用pymysql 做API 串接

```
import pymysql.cursors

# Connect to the database
connection = pymysql.connect(host='localhost',
                             user='user',
                             password='passwd',
                             db='db',
                             charset='utf8mb4',
                             cursorclass=pymysql.cursors.DictCursor)

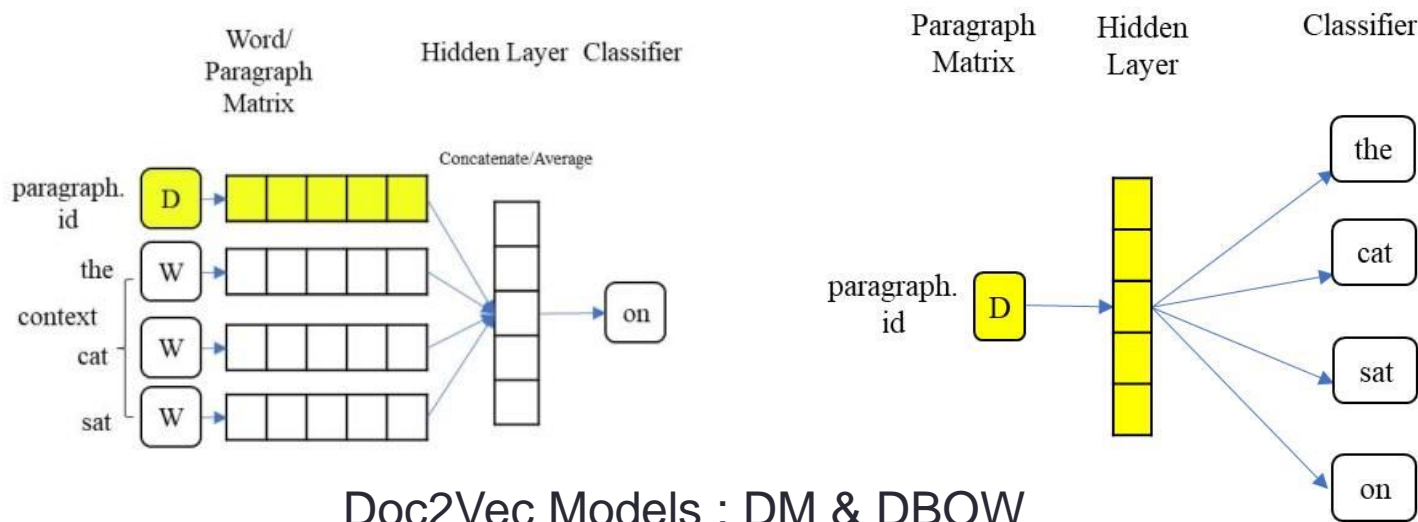
try:
    with connection.cursor() as cursor:
        # Create a new record
        sql = "INSERT INTO `users` (`email`, `password`) VALUES (%s, %s)"
        cursor.execute(sql, ('webmaster@python.org', 'very-secret'))

        # connection is not autocommit by default. So you must commit to save
        # your changes.
        connection.commit()

    with connection.cursor() as cursor:
        # Read a single record
        sql = "SELECT `id`, `password` FROM `users` WHERE `email`=%s"
        cursor.execute(sql, ('webmaster@python.org',))
        result = cursor.fetchone()
        print(result)
finally:
    connection.close()
```

2. + Keyword Alg. - Document Clustering

1. 事先分群：Kmeans/ DBSCAN/ K-medoids 之一，用SSE決定群數
2. 文章表達：Doc2Vec Network -> 可考慮語義，在許多實驗有良好表現
3. 新文章：找所屬群體 -> cosine similarity -> 找群體內m 篇最像成為領域文章



2. + Keyword Alg. - Keyword Generation

1. 找出領域文章的關鍵詞 -> LLR篩選 -> LLR 前 n%
2. 關鍵詞產生：LLR Formula

Document	in C_i	not in C_i	total
t exist	k_{11}	k_{10}	t_n
t not exist	k_{01}	k_{00}	t'_n
total	c_n	c'_n	D

$$\bullet H_{a,b} = \frac{a}{b} \log \left(\frac{a}{b} \right), \text{ where } a > 0$$

$$\bullet LLR_{t,i} = 2D \times \left(+H_{k_{11},D} + H_{k_{01},D} + H_{k_{10},D} + H_{k_{00},D} - H_{c_n,D} - H_{c'_n,D} - H_{t_n,D} - H_{t'_n,D} \right)$$

2. + Keyword Alg. – Integrate into Summary

1. 應用 1：壓縮時縱使沒有選到該群關鍵詞，關鍵詞會被囊括回來
2. 應用 2：Closed Sequential Pattern 篩選時，若有囊括關鍵詞，會將原本的support (frequency)乘上一定倍率，倍率大小為：
LLR比例正規化到100%-200% and 囊括關鍵詞詞數 之 函數
-> 用該weighted support 去做 pattern篩選 -> 後面步驟按原論文提出演算法進行。

Definition 1 Let $p_i = \langle t_1, t_2, \dots, t_m \rangle$ with $sup(p_i)$ - number of sentences (sequences) containing p_i , be a closed sequential pattern; t_j is j th term and m is number of terms in p_i . A “pattern-weight-pair set” of a pattern is

$$pw(p_i) = \{(t_1, w), (t_2, w), \dots, (t_m, w)\}, w = sup(p_i), t_j \in p_i \quad (4.3)$$

For example, a pattern $\{read, book\}$ with *support* 3 can be represented as

$$\{(read, 3), (book, 3)\}.$$

在此公式之下做調整

The End