



CSIS0801 Final Year Project 2015 – 2016

FYP15002

Chinese Character And Word Analysis In Daily Essays

Intermediate Project Report

Supervisor: Dr. Vincent Lau

Second Examiner: Dr. Beta Yip

Students:

Lau Tak Ming (3035042394)

Leung Ming Tak (3035053381)

Wong Man Wai (3035040683)

Table of Content

[CSIS0801 Final Year Project 2015 – 2016](#)

[Table of Content](#)

[Section 1: Project Overview](#)

[1.1 Project Background](#)

[1.2 Project Objective](#)

[Section 2: Technology used](#)

[2.1 Gensim -- Topic Modelling in Python](#)

[2.2 Word2Vec](#)

[2.3 Jieba](#)

[2.4 PHPCrawl](#)

[2.5 Scrapy](#)

[2.6 Chart.js](#)

[2.6 D3.js - Data-Driven Documents](#)

[Section 3: Project Methodology - Natural Language Processing](#)

[3.1 Overview](#)

[3.2 Data Cleaning and Normalization](#)

[3.3 Word Segmentation](#)

[3.4 Corpora Building](#)

[3.5 Analysis](#)

[3.5.1 Frequency Count with N-grams](#)

[3.5.2 Term weighting](#)

[3.5.3 Text Summarization](#)

[3.5.4 Documents Recommendation](#)



[3.5.5 Word Recommendation](#)

[Section 4: Milestones - Natural Language Processing](#)

[4.1 Word Segmentation](#)

[4.2 Frequency count and term weighting](#)

[4.3 Document Similarity](#)

[4.4 Word Similarity by word2Vec](#)

[Section 5: Milestones - Web Spider](#)

[5.1 Overview](#)

[5.2 System Design](#)

[5.3 Spiders for selected websites](#)

[5.4 Spiders for general web](#)

[5.5 Summary](#)

[Section 6: Summary and Further Development](#)

[6.1 Natural Language Processing](#)

[6.2 Web Application and Front-end Design](#)

[6.2.1 Composition](#)

[6.2.2 Goal](#)

[6.2.3 Styling of the Web App](#)

[6.2.4 Analysis Method Selection and Data Input](#)

[6.2.5 Data Transportation and Program Execution](#)

[6.2.6 Analysis Toolbox and Visualization Techniques](#)

[6.2.7 Summary of Analysis Toolbox](#)

[Section 7: Reference](#)

Section 1: Project Overview

1.1 Project Background

Chinese language is one of the most commonly used languages in our world, which covers approximately 1.2 billion people all over world. In addition, it is used by the majority of people living in Hong Kong, Mainland China and Taiwan. In the age of Internet, more and more online Chinese media and social media platforms have been arisen such that we can find many articles or online discussions written in Chinese. We believe that these materials reflect the cultural values and the trend in the society which is valuable to be studied.

However, unlike English, Chinese is a language written without spaces between words. This characteristic makes software difficult to retrieve every single word from an article and conducts subsequent analysis. In order to develop software which can process Chinese article effectively, we need to design special algorithm and script with a database to achieve this goal. Yet, we found that very few word retrieval tools exist for Chinese so we decided to work on it.


We believe that by developing software to analyze the pattern and usage of characters and words in daily use, we will be able to produce a lot of meaningful for subsequent studies like in the cultural area.

1.2 Project Objective

The ultimate goal of the project is to develop an online Chinese words analyzation tool which can display the statistical data (e.g. frequency of use, words relationship, domain origin, etc.) of each Chinese word.

To achieve this, we have the following sub-objectives:

- As the core of our project, we will design a segmentation algorithm that can effectively break down a Chinese essay into individual words that represent the closest meaning.
- We will develop a backend natural language processing tool that can using the segmentation algorithm to receive Chinese text from different systems and producing individual words from the text.
- We will develop a backend analyzation tool that works with the database to calculate the frequency count, word relationship, sources of words, etc.
- We will develop a web spider which is able to automatically fetch Chinese article / content from the Internet for subsequent analysis.
- We will design a database which can store Chinese words, statistic associated to words and web content retrieved from the spider in an organized and effective way.

- 
- We will design a front-end webpage to allow users to enter the website links for analysis, including Chinese character or word usage, and also the pattern of an essay.

Section 2: Technology used

2.1 Gensim -- Topic Modelling in Python

Gensim is a Python library for *topic modelling*, *document indexing* and *similarity retrieval* with large corpora. The library is widely used in the *natural language processing* (NLP) and *information retrieval* (IR) community.

2.2 Word2Vec

Released on 2013, word2vec is open source model of two-layer neural networks, that are trained to reconstruct linguistic contexts of words for grouping related words together. Currently the model has been implemented in Java, Python and Spark Mlib.

2.3 Jieba

"Jieba" (Chinese for "to stutter"): A open source Python Chinese word segmentation module, which

- Supports Traditional Chinese
- Supports customized dictionaries
- Unders MIT License

Other features such as part-of-speech tagging, adding custom dictionaries, parallel computation and keywords extraction are also included in the library. The library currently has been implemented in languages other than Python such as Java, C++, PHP etc..

Source: <https://github.com/fxsjy/jieba>

2.4 PHPCrawl

A framework under the GNU license for crawling/spidering websites written in the programming language PHP. It crawls websites and passes information about all found documents for further processing to users. Computation and keywords extraction are also included in the library.

2.5 Scrapy

A free and open source web crawling framework, written in Python. Originally designed for web scraping, and can also be used to extract data using APIs or as a general purpose web crawler.

Source: <https://github.com/scrapy/scrapy>



2.6 Chart.js

Chart.js is a JavaScript library which visualizes the data in different ways. Each of them is animated, with a load of customisation options and interactivity extensions. It also provides default simple support for canvas tooltips on hover/touch to the application

Source: <http://www.chartjs.org/>

2.6 D3.js - Data-Driven Documents

D3.js is a JavaScript library for manipulating documents based on data using HTML, SVG, and CSS. Moreover, D3.js combines powerful visualization components and a data-driven approach to DOM manipulation.

Source: <http://d3js.org/>

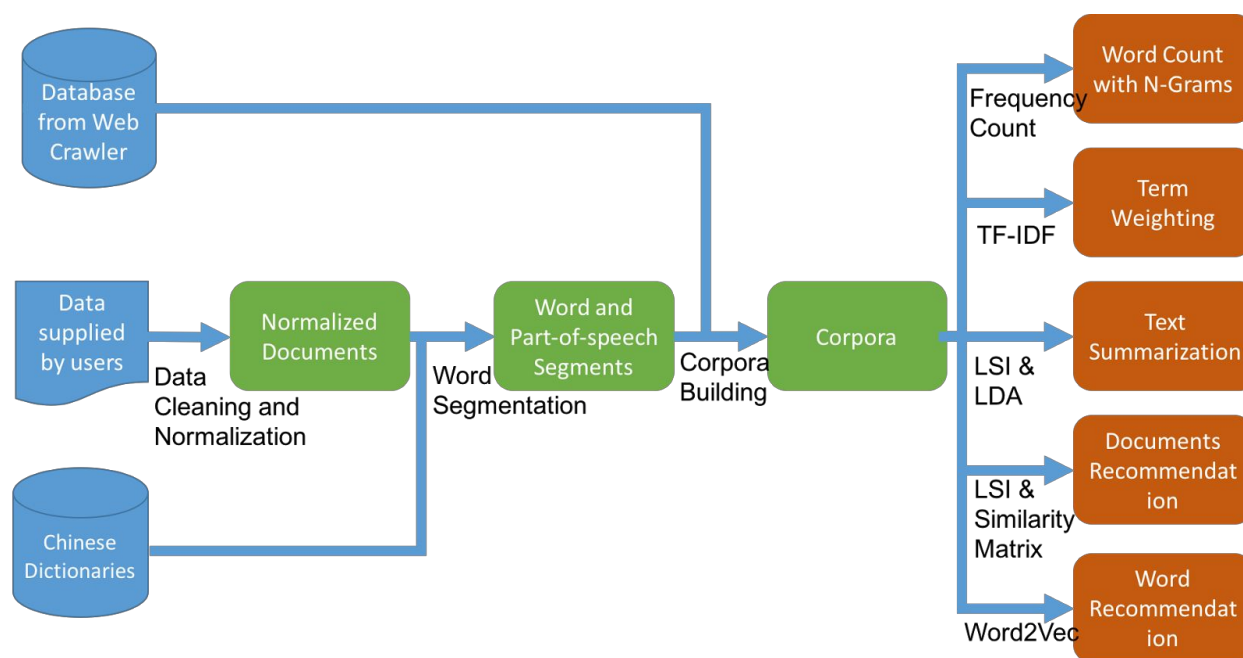
Section 3: Project Methodology - Natural Language Processing

3.1 Overview

The key process of the project would be manipulate and analyze language data. In this chapter, our team will introduce algorithms and techniques that are widely used in natural language processing (NLP) and analysis. We would also include brief discussion on the theory behind of each process.


The overview of NLP pipelines are shown below. The data generated in the process can be categorized to three types, including:

1. Preliminary data
2. Intermediate data
3. Analyzed data



Overview of NLP pipeline

Preliminary data, which is indicated by blue color, refers to data supplied at the beginning, data can be either local dictionaries, user input or database generated from web crawler. Intermediate data, which is indicated by blue color, refers to data generated in NLP pipeline, and refers to word segments with part-of-speech tagging, corpora and analysis data refers



to . Arrow connecting between two documents refers to processes or algorithms that will be discussed in the following sections.

3.2 Data Cleaning and Normalization

We assume that data supplied by user and database are not clean due invalid, excessive and inconsistent data. At the preliminary stage, we would normalize the data with the following stages:

1. Remove characters other than Chinese, punctuations, URL and tags etc.
2. Translate Simplified Chinese to Traditional Chinese in articles
3. Filter out short sentences

Text encoding would also be standardized at this stage.

3.3 Word Segmentation

Unlike other languages, segmentation of Chinese language is not trivial. Delimiters are not included in Chinese words and token cannot be determined easily by delimiters. In the project, we have explored algorithms and techniques that are used to segment a sentence into word segments. Consider a word “你好世界” (Hello world), we would illustrate methods that are used to segment the word into “你好” and “世界”:

1. Transform the sentence into prefix tree data structure. Word and values followed by each character indicates the weighting, which are defined by dictionaries.
2. Convert the tree to direct acyclic graph (DAG). The process aims to generate all possible word combinations from graph.
3. Find the path with highest weighting with dynamic programming.
4. For words or segments not included in dictionaries, a HMM-based model is used with the Viterbi algorithm.

TOP WORDS ⬇		BIGRAMS ⬇		TRIGRAMS ⬇	
Word	Frequency	bigram [®]	Frequency	trigram [®]	Frequency
政府	247	政府 繼續	20	一帶一路 沿線 國家	6
香港	163	沿線 國家	14	離岸 人民幣 業務	4
發展	154	一帶一路 策略	13	檢討 土地 用途	4
服務	107	初創 企業	11	政府 預留 億元	4
計劃	101	殘疾 人士	10	國家 十三 規劃	4
提供	78	持續 發展	10	創科 初創 企業	4
包括	58	政府 積極	10	提供 萬個 職位	4
長者	50	經濟 發展	9	保護 投資 協定	3
增加	49	約個 單位	9	持續 發展 政府	3

3.5.2 Term weighting

Term weighting is a central tool of search engines for scoring and ranking a document's relevance. In this project, we are going to evaluate the importance of each inputted words by term frequency inverse document frequency (TF-IDF), a product of two statistics.

For the condition of an important word, the word should be appear often in a document but not other documents. The formal definition of an important word are described as follows :

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}, \text{ where}$$

$n_{i,j}$: Occurrences of word from document d_j .

$\sum_k n_{k,j}$: Sum of occurrences from all documents

N : total number of documents in the corpus $N = |D|$

$|\{d \in D : t \in d\}|$: number of documents where the term t

Hence the weighting can be calculated by the product of two statistics:

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

Usage of TF-IDF will be illustrated in the following sections.

3.5.3 Text Summarization

Text summarization is an essential technique in NLP. For most cases, text summarization is useful for information and relation discovery, as well as document classification.

Given a Term-Document matrix, dimensionality of documents can be reduced by Latent Semantic Indexing (LSI), such that documents are “grouped” into topics. The following illustrates the matrix and its decomposition:

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Term-Document matrix: Element in the matrix represents the occurrence or weighting in each document.

Image source: <http://web.eecs.utk.edu/~mberry/sc95/sc95.html>

Topic modelling techniques like Latent Dirichlet Allocation (LDA) are also essential for text summarization.

3.5.4 Documents Recommendation

Given a word-document matrix from former section, similarity (correlation) matrix can be computed such that the similar documents can be indicated from the matrix, ranging from -1 to 1, higher magnitude would be more correlated to each other. In our case, we are interested in documents that are positively correlated to each other. The following matrix illustrates the similarity between topics, represented by its document index:

Document Index	40	20733	94363	199968	63281
40	1	0.952	0.952	0.948	0.949
20733		1	0.945	0.938	0.935
94363			1	0.961	0.945
199968				1	0.923
63281					1

3.5.5 Word Recommendation

Introduced by Google in late 2013, word2vec is vital for extracting similar words from dictionaries. For the model, word2vec model can be used to map each word into vector by two-layer neural network. Hence the model can predict the most similar word by cosine distance. The following diagram illustrate the workflow of word2vec:

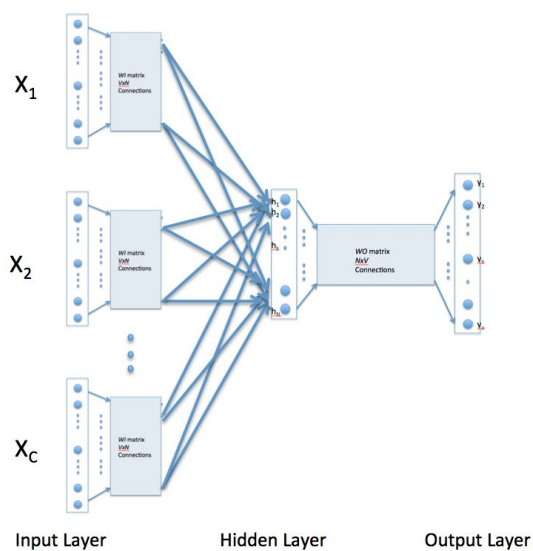


Image source from <https://iksinc.files.wordpress.com/2015/04/>

Section 4: Milestones - Natural Language Processing

In this chapter, our team will conclude milestones we achieved in previous phases. Summary and future works and deliverables will also be discussed. For NLP, we would demonstrate our analyses and works with abstract data (first paragraph) in Chinese Wikipedia.

4.1 Word Segmentation

Given an abstract of “數學” (Mathematics), the following shows the segmented Chinese words and punctuations.

```
segment2=" / ".join(jieba.cut(sentence))
print "After:\n"+ segment2
```

After:

數學 / (/ Mathematics /) / 是 / 利用 / 符號語言 / 研究 / 數量 / 、 / 結構 / 、
 / 變化 / 以及 / 空間 / 等 / 概念 / 的 / 一門 / 學科 / ， /
 / 從 / 某種 / 角度看 / 屬於 / 形式 / 科學 / 的 / 一種 / 。 / 數學 / 透過 / 抽象化
 / 和 / 邏輯推理 / 的 / 使用 / ， / 由 / 計數 / 、 / 計算 / 、 / 量度 / 和 / 對 /
 物體 / 形狀 / 及 / 運動 / 的 / 觀察 / 而 / 產生 / 。 /
 / 數學家 / 們 / 拓展 / 這些 / 概念 / ， / 為 / 了 / 公式化 / 新 / 的 / 猜想 / 以及
 / 從 / 選定 / 的 / 公理 / 及 / 定義 / 中 / 建立起 / 嚴謹 / 推導 / 出 / 的 / 定理 /
 。

4.2 Frequency count and term weighting

The following table shows the most commonly used words in Chinese Wiki abstracts:

```
# Sort by frequency count
!sort -k 3 -rn wikiDict.txt > sortedWikiDict.txt
!head sortedWikiDict.txt
```

28687	位於	54606
50179	一個	44078
13039	中國	37696
55647	分佈	31973
39366	一種	29991
5334	其中	21486
28668	平方公里	20865
13277	地區	18431
10507	學名	17967
29576	棲息	16356

First column refers to index of each Chinese word. From the above, “位於” (location), “一個” (one) and “中國” (China) are the most common words in Chinese Wikipedia with frequency 54406, 44078 and 37696 representatively.

Term Weighting by TF-IDF

The following tuples shows the importance of each word in abstract “數學” (Mathematics):

```
Term weighting by TF-IDF:
(概念, 0.23) (以及, 0.13) (數學, 0.23) (這些, 0.11) (運動, 0.11) (定義, 0.12)
(透過, 0.12) (觀察, 0.15) (使用, 0.08) (抽象化, 0.20) (公式化, 0.22)
(物體, 0.15) (學科, 0.13) (形式, 0.11) (建立起, 0.20) (一門, 0.14) (利用, 0.11)
(屬於, 0.06) (科學, 0.12) (量度, 0.19) (變化, 0.13) (產生, 0.10) (結構, 0.11)
(猜想, 0.20) (空間, 0.12) (公理, 0.18) (角度看, 0.23) (一種, 0.05) (選定, 0.18)
(研究, 0.09) (計數, 0.20) (某種, 0.15) (計算, 0.13) (嚴謹, 0.18) (定理, 0.15)
(拓展, 0.18) (邏輯推理, 0.23) (推導, 0.19) (數量, 0.14) (形狀, 0.14) (數學家, 0.14)
```

From the above wiki abstract “數學” (Mathematics), the most important words are “概念” (concept), “數學” (Mathematics), “邏輯推理” (Logical reasoning) and “角度看” (Perspective).

Text Summarization

The following shows the most relevant words in each summarized topic. Value followed by each word refers to its relevance.

```
#Latent Semantic Indexing (LSI)
for i,j in lsi.show_topic(23)[:5]: print i,j
```

```
冰川 -0.398759466345
冰原 -0.380382283655
島峰 -0.380199566592
群島 0.230939721847
山峰 0.209200546951
```

```
for i,j in lsi.show_topic(0)[:5]: print i,j
```

```
棲息 0.279671065614
可達 0.250394935748
體長 0.236198529056
公分 0.234641148863
習性 0.224247181259
```

Text summarization by Latent Semantic Indexing (LSI): Words that are more relevant in the topic would have a higher value magnitude.

4.3 Document Similarity

Three most relevant abstracts of abstract “遊戲” (Game), which returns abstract of “打錢” (Gold farming), “3D遊戲” (3D game) and “密碼(遊戲)” (Password (Video Gaming)).

```
# Input document index
documentIndex = 40
print " ".join(documents[documentIndex][:20])
```

遊戲 可以 指人 一種 娛樂活動 也可以 這種 活動 過程 遊戲 道具 可以 玩具 英語 體育比賽 遊
戲 一種 體育運動 遊戲 演變

```
sims = index[lsi[tfidf[corpus[documentIndex]]]]
# Sort the similarity values by their scores
sims = sorted(enumerate(sims), key=lambda item: -item[1])
```

```
# print sims
#[(40, 1.0), (20733, 0.95213395), (94363, 0.95173347), (199968, 0.95007229),
# (63281, 0.94930953), (55712, 0.94783223), (225236, 0.94558185), (7930, 0.94430
256),
# (63259, 0.94354367), (12311, 0.9380967), (2379, 0.93661648), (2254, 0.93655413
), .....

# Three most similiar topics
print "Index " + str(sims[1][0]) + ": " + " ".join(documents[sims[1][0]][:20])
print "Index " + str(sims[2][0]) + ": " + " ".join(documents[sims[2][0]][:20])
print "Index " + str(sims[3][0]) + ": " + " ".join(documents[sims[3][0]][:20])
```

Index 20733: 打錢 大型 多人 在線 角色扮演 遊戲 某個 通過 不斷 玩游 戲來 獲取 金錢 一
種 行為 打錢 眾多 大型 多人 在線
Index 94363: 遊戲 指以 三維 計算機 圖形 基礎 製作 立體 電子 遊戲 相對 傳統 遊戲 來說
帶給 玩家 更加 真實 遊戲 體驗
Index 199968: 密碼 一種 第三 第四 世代 許多 電子 遊戲 用於 返回 遊戲 階段 一串 資料
通過 關卡 接續 關卡 遊戲 給出

Document similarity extraction

3D游戏是指以三维计算机图形为基础制作的立体电子游戏，相对传统的2D游戏来说，会带给玩家更加真实的游戏体验。3D遊戲是指遊戲是以3D技術製成，而並不是指螢幕是以3D輸出，令人有覺得是立體的感覺。

密碼是一種在第三和第四世代的许多电子游戏中，用於返回遊戲階段的一串資料碼。当通过关卡或接續关卡时，游戏就会给出密码，玩家可以使用其返回此游戏阶段。简单情况下，密码只记录当前到达的关卡，而在角色扮演游戏等复杂情况下，密码会记录财产、事件等更多状态。随着现今存档储存器的广泛应用，密码逐渐罕见。

打钱是指在大型多人在线角色扮演游戏（MMORPG）中，某个人通过不断地玩游戏来获取金钱的一种行为。打钱也是众多大型多人在线角色扮演游戏所广为流传的一种方法。打钱工作者是指某个人通过收集游戏中的虛擬货币，再将其卖给其他游戏玩家来换取现实生活中的货币为目的的一种人。

Comparison with Wiki from internet.

4.4 Word Similarity by word2Vec

The following shows the most similar words compared to “計算機科學” (Computer Science). Value followed by word represents its cosine distance with input word.

```
for w, score in model.most_similar(u"計算機科學"): print w, score
```

```
信息科學 0.763425171375  
數理統計 0.724284648895  
理論物理 0.723909497261  
數學 0.719389021397  
統計學 0.712564647198  
材料科學 0.706905901432  
工程學 0.704143762589  
應用 0.702665686607  
計量經濟學 0.69039785862  
通信工程 0.687664985657
```

Section 5: Milestones - Web Spider

5.1 Overview

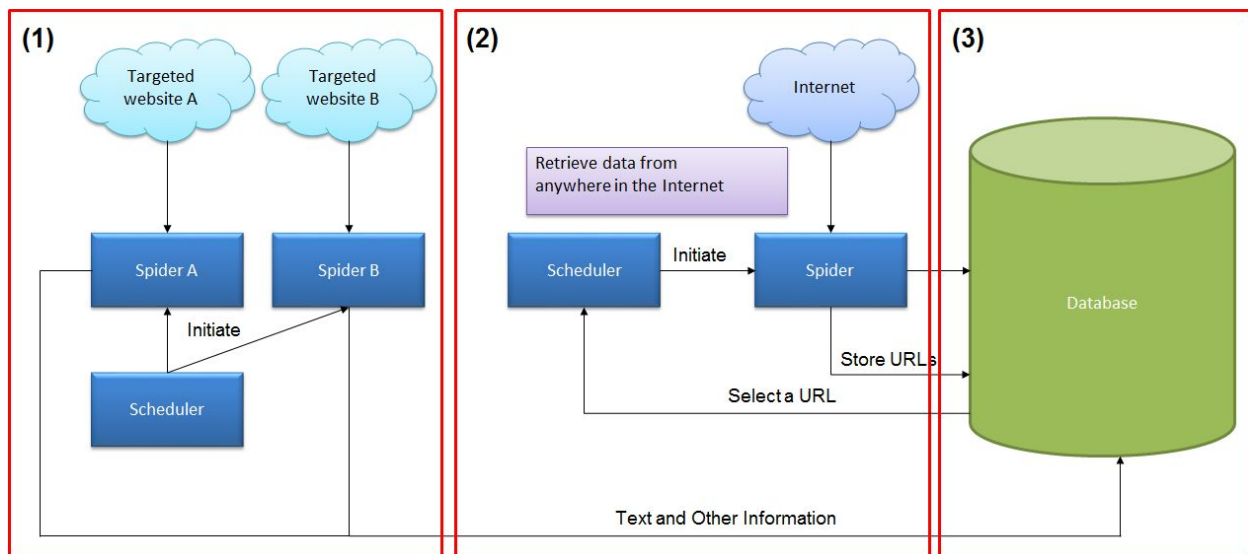
As we are working on analysis in daily Chinese essays, there are two channels of data collection:

1. Manual Input by user
2. Automatic data collection

Web spider is targeting to conduct automatic data collection. A huge database of Chinese essays is essential to conduct analysis and extract useful information (e.g. popular words in 2016, comparison between media with different political view)

Also, The performance of segmentation algorithm can be improved by running on more essays by adding more words to the dictionary

5.2 System Design



Part (1): Collect structural data from selected websites

Part (2): Collect Chinese plaintext from others websites

Part (3): MySQL database for data storage

5.3 Spiders for selected websites

For websites with high research value, we would develop specific spiders to crawl structural data from them. Currently, we have developed spiders for:

1. RTHK News Updates (<http://news.rthk.hk/rthk/ch/latest-news.htm>)
2. Chief Executive Blog Articles (<http://www.ceo.gov.hk/chi/blog/>)
3. Financial Secretary Blog Articles (<http://www.fso.gov.hk/chi/blog/>)

The spiders are able to retrieve XML documents from data sources regularly and extract information such as publication date, title and categories in addition to the essay by locating the specific XML tag of the desired data.

Here is a screenshot of XML document retrieved from RTHK:

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```

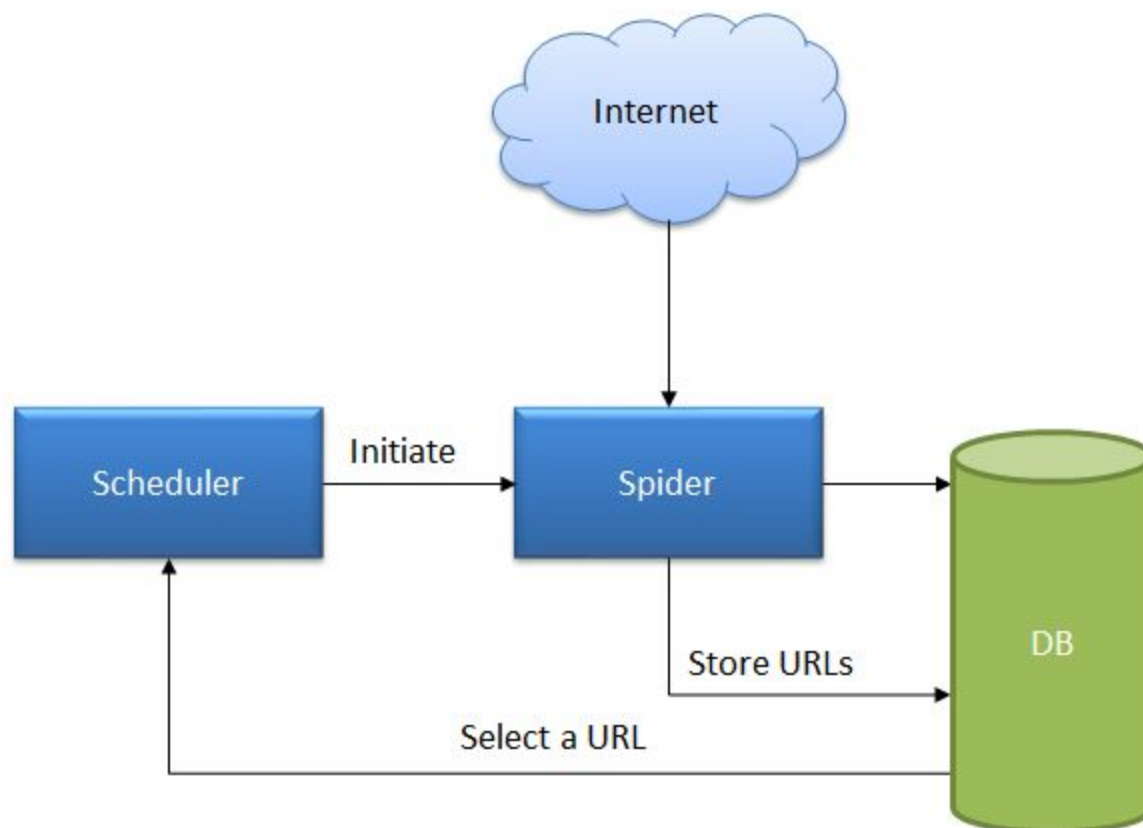
<?xml version="2.0"?>
<channel>
  <title>rthk.hk - 即時新聞: 本地</title>
  <link>http://www.rthk.hk/</link>
  <description>rthk.hk - 即時新聞: 本地</description>
  <language>en-us</language>
  <copyright>香港電台</copyright>
  <webmaster>webmaster@rthk.hk</webmaster>
  <pubDate>Fri, 15 Jan 2016 09:00:37 +0800</pubDate>
  <lastBuildDate>Fri, 15 Jan 2016 08:57:50 +0800</lastBuildDate>
  <category>新聞</category>
  <docs>http://blogs.law.harvard.edu/tech/rss</docs>
  <ttl>30</ttl>
  <image>
    http://rthk.hk/include2010/homepics/images/home_logo.png
  </image>
  <item>
    <title>
      <![CDATA[ 張建宗勞資關係非「你贏我輸」 盼市民對政府公道一點 ]]>
    </title>
    <guid>
      http://news.rthk.hk/rthk/ch/component/k2/1235947-20160115.htm
    </guid>
    <description>
      <![CDATA[
        勞工及福利局局長張建宗承認，新一份《施政報告》在勞資方面沒有大亮點，但他認為不能要求每年的施政報告都有大亮點，因為勞工和福利政策需要時間醞釀和推行。張建宗出席本台節目時又表示，標準工時和退休保障等重要政策的推動，需要經過一定過程，不能一蹴而就，一步登天，勞資關係亦非你贏我輸的零和遊戲，社會要建立共識，希望市民給予時間和空間，對政府公道一點。張建宗又說，今年《施政報告》深化多項惠及基層和長者的措施，顯示政府致力建立一個和諧的社會，他亦向警務處司長林炳輝致意，這是一份熱心的施政報告。
      ]]>
    </description>
    <pubDate>
      Fri, 15 Jan 2016 08:57:50 +0800
    </pubDate>
  </item>
  <item>
    <title>
      <![CDATA[ 張炳良指發展公屋房屋挑戰包括房委會營餘減少及欠缺土地 ]]>
    </title>
    <guid>
      http://news.rthk.hk/rthk/ch/component/k2/1235945-20160115.htm
    </guid>
    <description>
      <![CDATA[
        運輸及房屋局局長張炳良在一個電台節目中表示，發展公屋房屋有3方面挑戰，包括房委會營餘減少、欠缺土地，以及建造業人手不足。他說，當局不可隨意輸入外勞，長遠要努力解決土地以及人手等問題。《施政報告》提出，重新規劃將軍澳第137區填海發展成商住地區，被問到如何解決交通配套問題，張炳良說明白將軍澳區的交通需求，政府計劃在今個立法年度，向議會申請撥款，興建將軍澳至藍田隧道。他又提到當局在2014年提出計劃興建東九龍鐵路橋樑，舒緩區內的交通需要，但強調仍須詳細規劃和諮詢地區。
      ]]>
    </description>
  </item>
</channel>

```

A crontab with 1-min interval is set to call the spider to retrieve an updated XML document extract the specific information (enclosed by red rectangle). The crontab interval is determined by the update usual frequency of the data source (e.g. shorter for news but longer for blog).

5.4 Spiders for general web

Although we can get more structured data using spiders for specific website, it is time-consuming to develop as we need to study the structure of every single website. Thus, a spider for general web is developed to crawl Chinese essay from other sources which have relatively lower research value.

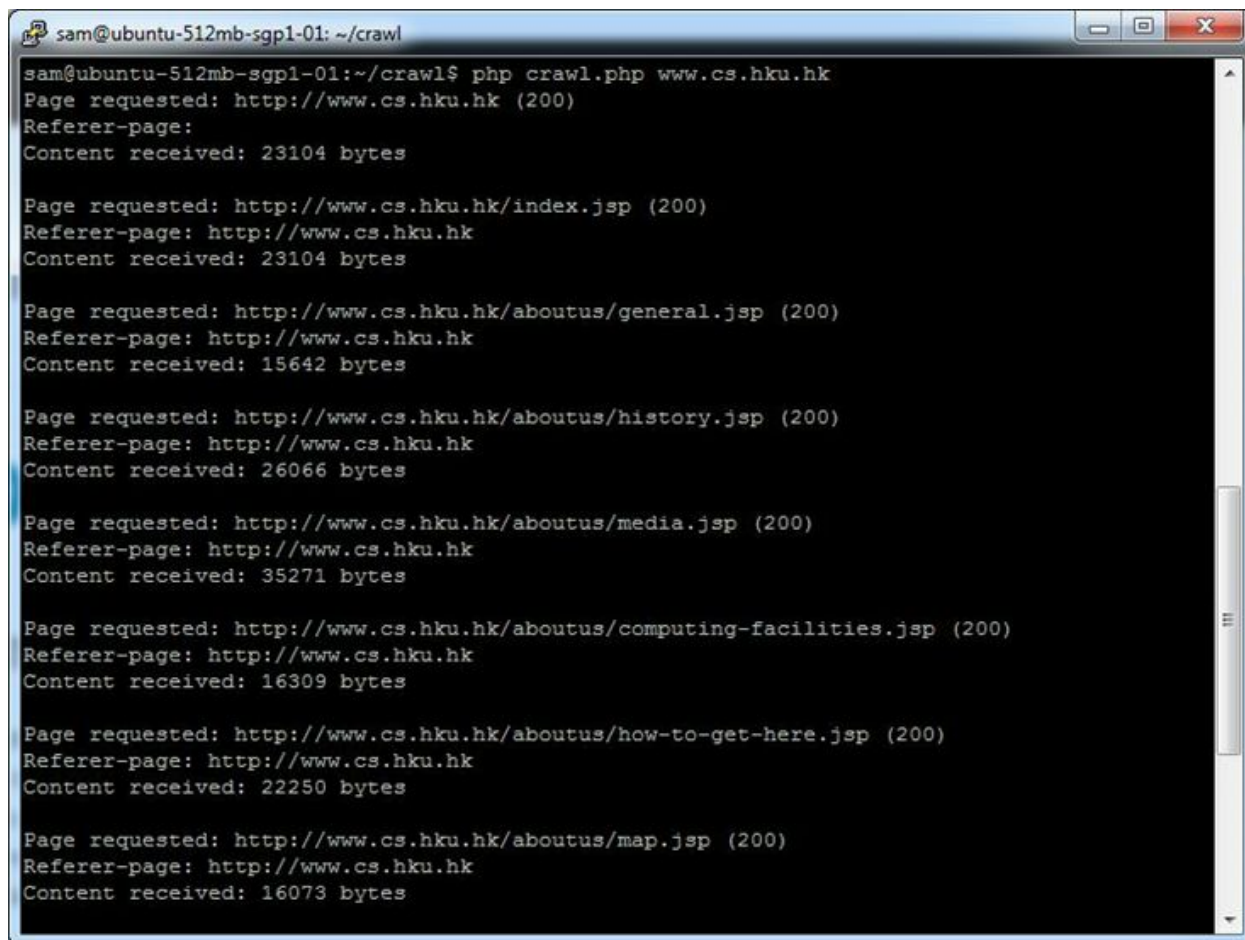


By using libraries PHPCrawl, Simple HTML DOM Parser and Scrapy. The Spider is able to locate hyperlinks in the document and extract plaintext from it.

We have provide a group of website as starting points and the scheduler will call the spider regularly to crawl through the web. Starting websites include common Chinese web portals such as Yellow Page, discuss.com.hk, yahoo.com.hk, etc.

During crawling, the spider would first follow links that lead to the same domain, then every link even leads to a different host or domain until a specific time period has been reached. Plain text and hyperlinks collect would sent to database for storage and subsequent visiting.

There is a screenshot when spider is running:

A terminal window titled 'sam@ubuntu-512mb-sgp1-01: ~/crawl' displays the output of a PHP crawler script. The script has successfully crawled several pages from the website www.cs.hku.hk. The output shows the URL of each page requested, the status code (200), the referer page, and the amount of content received in bytes. The pages crawled include the main page, index.jsp, aboutus/general.jsp, aboutus/history.jsp, aboutus/media.jsp, aboutus/computing-facilities.jsp, aboutus/how-to-get-here.jsp, and aboutus/map.jsp.

```
sam@ubuntu-512mb-sgp1-01:~/crawl$ php crawl.php www.cs.hku.hk
Page requested: http://www.cs.hku.hk (200)
Referer-page:
Content received: 23104 bytes

Page requested: http://www.cs.hku.hk/index.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 23104 bytes

Page requested: http://www.cs.hku.hk/aboutus/general.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 15642 bytes

Page requested: http://www.cs.hku.hk/aboutus/history.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 26066 bytes

Page requested: http://www.cs.hku.hk/aboutus/media.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 35271 bytes

Page requested: http://www.cs.hku.hk/aboutus/computing-facilities.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 16309 bytes

Page requested: http://www.cs.hku.hk/aboutus/how-to-get-here.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 22250 bytes

Page requested: http://www.cs.hku.hk/aboutus/map.jsp (200)
Referer-page: http://www.cs.hku.hk
Content received: 16073 bytes
```

5.5 Summary

Currently, web spiders have retrieved the following data and stored in the database:

1. 568,858 entries of plain text crawled by general web spider.
2. 254,940 hyperlinks found by general web spider.
3. 1,613 Chinese news article by RTHK.
4. 39 articles from Chief Executive's blog.
5. 8 articles from Financial Secretary's blog.

These data will be passed to the NLP component of our project and used for subsequent analysis. For example, the most concerned issue in Hong Kong can be found from processing RTHK's news.

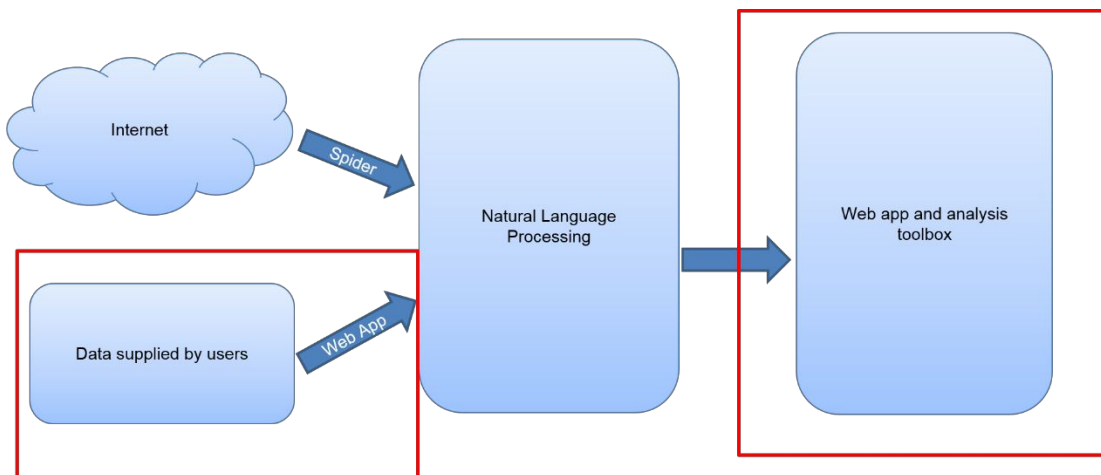
Section 6: Summary and Further Development

6.1 Neutral Language Processing

In the future phase, we will include the word frequency with bigrams and trigrams, also topic modelling methods like Latent Dirichlet Allocation (LDA) will be explored.

NLP and word analysis techniques	Methodology and algorithm	Package	Milestone	Future development
Word Segmentation	Prefix Tee, DAG, Dynamic Programming and HMM	Jieba	Chinese words segment program with part-of speech	/
Frequency Count	/	/	Program for word count	Frequency count with n-grams
Term Weighting	TF-IDF	Gensim	Program for computing TF-IDF	/
Text Summarization	LSI and LDA	Gensim	Program computing LSI	LDA
Document Similarity	LSI and Similarity Matrix	Gensim	Program computing similarity Matrix	/
Word Similarity	Neural Network	Word2Vec	Program for model training	/

6.2 Web Application and Front-end Design



Overview of project

6.2.1 Composition

A web app is implemented as the front-end application. The web app is divided into two parts:

1. User input front-end web app
2. Analysis toolbox with visualization

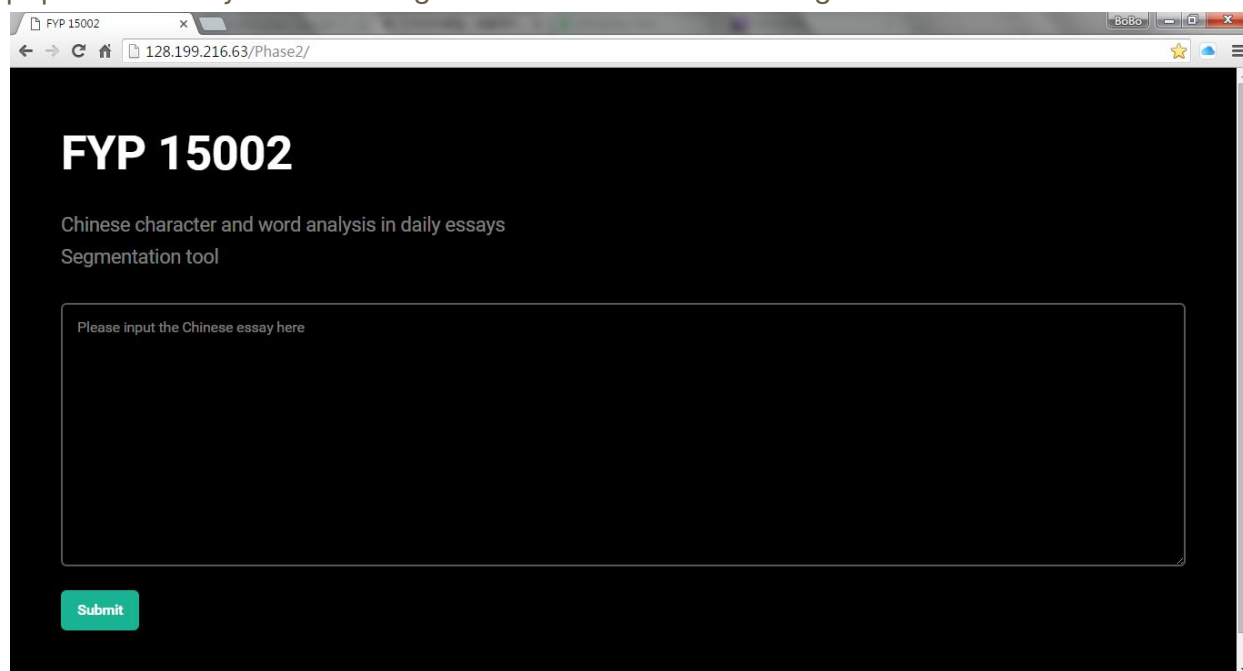
6.2.2 Goal

The goal of front-end web app is not only limited to attractive user interface, but also to provide great user experience.

6.2.3 Styling of the Web App

The web app is using HTML5 technology as the styling of itself. HTML5 provides a clear user interface for user to input the data.

Moreover, HTML5 provides responsive characteristic of the website. Responsive website is popular nowadays in web design. It makes the website look good on all devices.



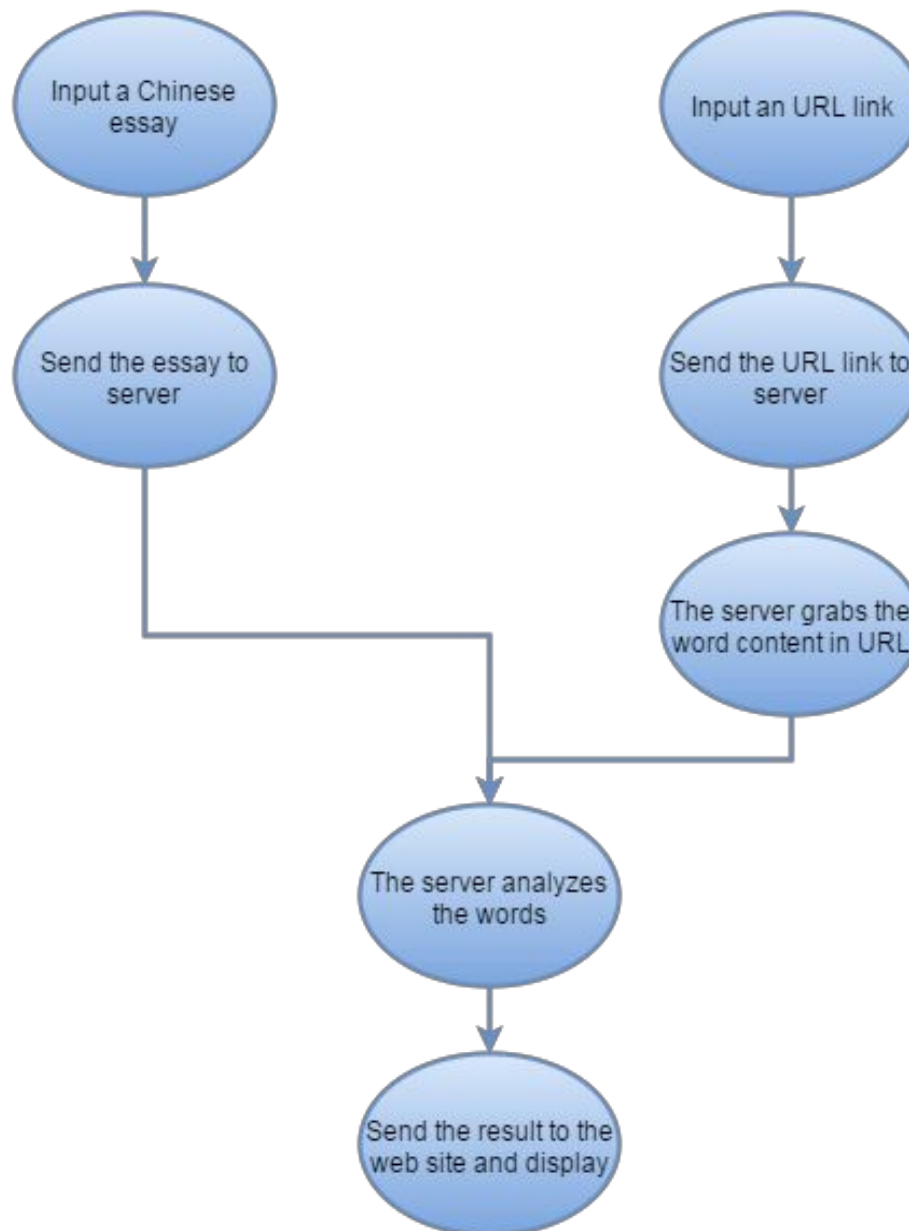
User Interface of the Segmentation Tool

6.2.4 Analysis Method Selection and Data Input

Before inputting the data, user selects the analysis function. Then, based on the analyzation selected, user inputs the correct type of data(i.e. Chinese essay plaintext or URL link). Further details will be discussed in the section **Analysis Toolbox and Visualization Techniques**.

6.2.5 Data Transportation and Program Execution

PHP script is used to retrieve the data input by user in the web app. The method used is POST.



Flow of Data Transportation

Next, the input data is sent as argument of the back-end analysis toolbox program. The program is executed by using the following script:

```
exec("python segmentation.py $input 2>&1");
```

6.2.6 Analysis Toolbox and Visualization Techniques

As the project needs to manipulate large amounts of data, visualization is vital for user to understand large and complex data. In order to provide great user experience to user, various visualization techniques and libraries are used in the application design.

The analyzation are

1. Segmentation Tool
2. Frequency Count & Word categorization
3. Word Trend
4. Author's Word
5. Similar Word
6. Similar Essay

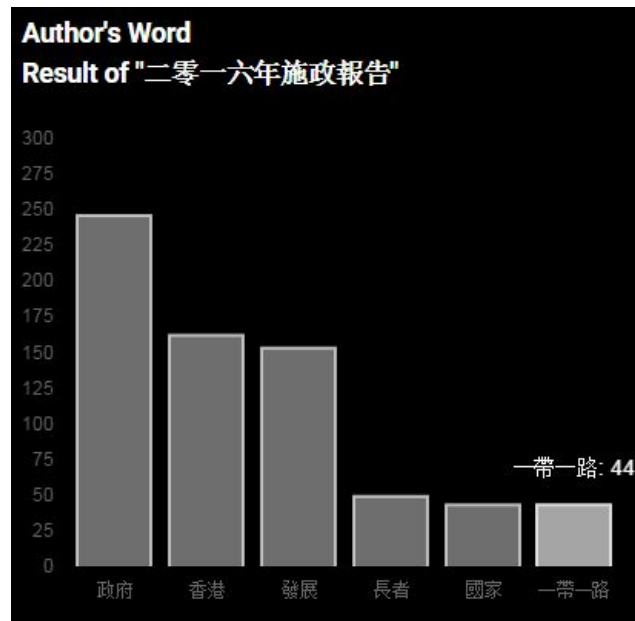
Segmentation Tool: It is a tool that break down a Chinese essay into individual words that represent closest meaning. The web app use a slash symbol “ / ” to separate the words in an essay and display it.



Segmentation Tool

Frequency Count & Word Categorization: The analysis tool will count the occurrence of Chinese word segments from incoming essay, and also categorize the words with similar meaning. Word bubbles will be used to present the analysis result as follow:

Author's Word: Every author has his/her own writing style. The user can input the author's name and the Chinese words which he/she uses frequently.



Sample of Author's Word analysis

Similar Word: The analysis tool can look for similar word and relevant word in the database. The sample is shown as below:

FYP 15002
Chinese character and word analysis in daily essays
Similar Word
Result of "計算機科學"

Word	Score
信息科學	76.34
數理統計	72.43
理論物理	72.39
數學	71.94
統計學	71.26
材料科學	70.69

Sample of Similar Word

Similar Essay: By extracting the keywords in a Chinese essay, the analysis tool can search for any essay which has the similar words of the keywords of original essay.

FYP 15002

Chinese character and word analysis in daily essays

Similar Essay

Result of 遊戲 <https://zh.wikipedia.org/wiki/%E6%B8%B8%E6%88%8F>

- 打錢 <http://www.wikiwand.com/zh-hk/%E6%89%93%E9%92%B1>
- 3D遊戲 <https://zh.wikipedia.org/wiki/3D%E6%B8%B8%E6%88%8F>
- 密碼 (電子遊戲) [https://wikipedia.kfd.me/zh-hk/%E5%AF%86%E7%A0%81_\(%E7%94%B5%E5%AD%90%E6%B8%B8%E6%88%8F\)](https://wikipedia.kfd.me/zh-hk/%E5%AF%86%E7%A0%81_(%E7%94%B5%E5%AD%90%E6%B8%B8%E6%88%8F))

Sample of Similar Essay

6.2.7 Summary of Analysis Toolbox

There are several analysis methods:

Analysis	Input	Output
Segmentation Tool	Chinese essay / URL link	Plain text segment
Frequency Count & Word Categorization	Chinese essay / URL link	Word bubbles
Word Trend	Chinese word	Line chart
Author's Word	Author's name	Bar chart
Similar Word	Chinese word	Table of similar words
Similar Essay	Chinese essay / URL link	List of similar essay

Section 7: Reference

Introduction to Latent Dirichlet Allocation

<http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>

Dimensionality Reduction and Latent Topic Models

<http://pages.cs.wisc.edu/~jerryzhu/cs769/latent.pdf>

Google's word2vec

<https://code.google.com/p/word2vec/>

Latent Semantic Analysis and Topic Modeling: Roads to Text Meaning

<http://www.jaist.ac.jp/~bao/Writings/TopicModeling2.pdf>

Latent Semantic Indexing (LSI) An Example

<http://www1.se.cuhk.edu.hk/~seem5680/lecture/LSI-Eg.pdf>

word2vec 中的数学原理详解

<http://suanfazu.com/t/word2vec-zhong-de-shu-xue-yuan-li-xiang-jie-duo-tu-wifixia-yue-du/178>

中英文维基百科语料上的Word2Vec实验

<http://www.52nlp.cn/中英文维基百科语料上的word2vec实验>

Matrix decompositions and latent semantic indexing

<http://nlp.stanford.edu/IR-book/pdf/18lsi.pdf>

Chinese NLP with Open Source Tools in Python

<https://github.com/albertaueyung/pyconhk2015-chinese-nlp>

JIEBA 結巴中文斷詞

<https://speakerdeck.com/fukuball/jieba-jie-ba-zhong-wen-duan-ci>

Ultimate guide for scraping JavaScript rendered web pages

<https://impythonist.wordpress.com/2015/01/06/ultimate-guide-for-scraping-javascript-rendered-web-pages/>

Crawler Approaches And Technology

<http://www.iicm.tugraz.at/cguetl/courses/isr/uarchive/uews2008/Ue01%20-%20Crawler-Approaches-And-Technology.pdf>

PHP Simple HTML DOM Parser

<http://simplehtmldom.sourceforge.net/>



Google Trends

<https://www.google.com.hk/trends/>

Chart.js

<http://www.chartjs.org/>

D3.js - Data-Driven Documents

<http://d3js.org/>