FYP 15002

CHINESE CHARACTER AND WORD ANALYSIS IN DAILY ESSAYS

Department of Computer Science, The University of Hong Kong

Supervisor
Dr. Vincent Lau

Group
Lau Tak Ming (3035042394)
Leung Ming Tak (3035053381)
Wong Man Wai (3035040683)

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.

Project Objective

The ultimate goal of the project is to develop an online Chinese words analyzation tool which can display the statistic 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.

Project Methodology

The major tasks of the project include raw data collection, natural language processing, database design, word analysis and front-end design. The role of each task, as well as its implementation details will be discussed in this section.

Raw Data Collection

The most important data source of daily essays is undoubtedly the web. The team will implement suite of programs that automatically collects daily essays, targeting news, social media and forums from web, known as web spider. After crawling raw data from web, the team will normalize the data by removing unnecessary tags and words. The remaining information will be prepared for natural language processing.

Natural Language Processing

The key process of the project would be manipulate and analyze language data, therefore the team will explore techniques that are widely used in natural language processing (NLP). Some important algorithms, such as word segmentation, word tagging and categorizing, word classification and chunking will be explored and implemented in the project. The processed data will be stored into the database for further analysis.

Database Design

The team will implement a database that stores all the essays in a structural with the following attributes: region, date time, source, author, essay, tags etc.. By adopting word segmentation algorithm in NLP, the team plans to develop a word bank that not only stores historical words but also automatically inserts segmented words from essays and articles. To optimize query performance, the team will index the data before proceeding to word analysis stage.

Word Analysis

The group explored some possible features and analysis in the project. The idea will be further explored, evaluated and implemented:

Frequency Count: System will count the occurrence of Chinese word segments from incoming essay, and extracts the most common words from social media.

Behaviour Analysis: After tagging author name from essay, the group will explore author's writing style: What words does he/she most likely to say? What type of issues does he/she is interested? etc.

Pattern Analysis: The group will develop a user-interactive tool. User can enter a essay so as to analysis the Chinese character or word, and also search for similar essay by its pattern.

Spatial Analysis: The group will categorize media or forum from web into different regions or countries. After that, the group will study the behaviour of Chinese character usage among different regions or countries.

Time Series Analysis: The group will further explore the usage of Chinese word by finding the trend of specific words. It may useful for exploring the growth and decay of some social issues.

Association Analysis: The group will discover the relation between words. For example, given a Chinese essay with tagged keywords, the system will extract all the related keywords from database. User can thus discover more similar topics from tagged keywords.

Front-end Design

The team will implement a web application. By providing analysis toolbox, user are able to discover analyzed patterns and trends from the app. Our design goal is not only limited to attractive user interface, but also to provide great user experience. 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, we would also explore various visualization techniques and libraries in the application design.

Project Schedule and Milestones

Milestone	Date of Completion	Deliverables
1. Project Initialization	10 July 2015	-
2. Project Analysis and Design	26 September 2015	-
3. Phase 1 Deliverables	4 October 2015	Project website Project plan
4. Raw Data Collection	10 January 2016	Implementation: ● Web spider ● Raw text preprocessing
5. Algorithm Design and Implementation	10 January 2016	Implementation: ● Word segmentation ● Word tagging and categorization ● Information extraction
6. Phase 2 Deliverables	11 January 2016	Preliminary Implementation Interim presentation materials Interim report
7. Database Design and Implementation	28 February 2016	Implementation: ● Structural database ● Word tree bank
8. Front-end Design and Implementation	31 March 2016	Implementation: ● Web app ● Word analysis toolbox
9. Phase 3 Deliverables	17 April 2016	Finalized Implementation Final report
10. Final Presentation	22 April 2016	Final presentation materials