웹 환경에서 100 논문에 대한 텍스트 마이닝, 데이터 분석과 시각화

이효맹* · 이가베* · 이현창* · 신성윤**

*원광대학교

**군산대학교

100 Article Paper Text Minning Data Analysis and Visualization in Web Environment

Xiaomeng Li[1] · Jiapei Li[1*] · HyunChang Lee[1],· SeongYoon Shin[2]

*Wonkwang University

**Kunsan National University

E-mail : *jiapei@gmail.com *hclglory@gmail.com

ABSTRACT

There is a method to analyze the big data of the article and text mining by using Python language. And Python is a kind of programming language and it is easy to operating. Reaserch and use Python to creat a Web environment that the research result of the analysis can show directly on the browser. In this thesis, there are 100 article paper frrom Altmetric, Altmetric tracks a range of sources to capture. It is necessary to collect and analyze the big data use an effictive method, After the result coming out, Use Python wordcloud to make a directive image that can show the highest frequency of words.

Keywords

co-word analysis, python, wordcloud

Ⅰ. Introduction

Python is a computer engineering languageand and it is a widely used high-level programming language for general-purpose programming. An interpreted language, Python has a design philosophy that emphasizes code readability, and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages. The language provides constructs intended to enable writing clear programs on both a small and large scale.

Python is the most popular language to make statistics. Python itself has data source connection, and it read data, the operation of the system, or the regular expression and word processing have a clear advantage. Nowadays it is a populor tool to make data analysis. Specially, in a Web-based environment could make the result more and more directive just like show on a browser, this is mainly to achieve..

Ⅱ. Research method

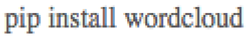
Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, fuctional programming, and procedural styles. It has a large and comprehensive standard library.

Word cloud is a visual representation of text data, typically used to depict keyword meta data(tag) on web sites, or to visualize free form text. Python can use the function of wordcloud make a directive visualization that the most important information could be got. That is the reason for the wordcloud function's exist.

Ⅲ. Analysis result

To install the wordcloud package firstly call to the terminal and install it. This is the most important step. If it was missed that the function of wordcloud would not be operated .

.



```
pip install wordcloud
```

Secoundly, programming the code of the wordcloud , Python will get the data from the download text, in this thesis we downloaded the data from Altmetric, and the next step is operating the code,

And in this link, we use a web enviornment Python programming tool to achieve the operation, this environment is named 'Jupyter' , It is Web-based, interactive computing notebook environment, the function is edit and run human-readable docs while describling the data analysis

Ⅳ.Instantiation

Use the data from the Altmetric that The Top100 article papers in 2016, There are 100 titles of the article papers that could use to make a wordcloud image. Programming the wordcloud code and then operate it, Finally, the wordcloud image show the result of the most important information and the highest frequency of words of the 100 article papers.

Ⅴ.Conclusion

This thesis inroduced the wordcloud and a programming tool named 'Jupyter' which could programming the Python code in a web environment. It means 'Jupyter' could edict the codes and operate the codes in a web environment. And wordcloud is a convinient and popular form to make data analysis nowadays. Finally, the result of a wordcloud image showed with a broeser in a web enviroment has been achieved.

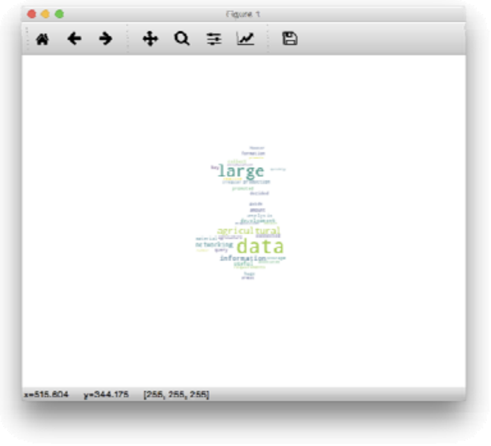


Fig 1. Result of the wordcloud for 100 articles

Reference

[1] Li X, Thelwall M, Giustini D. Validating online reference managers for scholarly impact measurement. Scientometrics. 2011;91(2):461-71.

[2]Haustein S, Larivière V. A multidimensional analysis of Aslib proceedings – using everything but the impact factor. Aslib Journal of Information Management. 2014;66(4):358-80.

[3]Costas R, Zahedi Z, Wouters P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology. 2015;66(10):2003-19.