

# Wikipedia Analysis Projects

---

Title : **Is the quantity/quality of Wikipedia document of each university correlated with University Level?**

*This project is carried out for the Marketing Research Class(fall semester of 2019) by Prof. Shali Wu(School Of Management, Kyung Hee University, Seoul, Korea).*

## Purpose

---

- The purpose of this project is to investigate the correlation between the quantity/quality of Wikipedia document and the University Ranking/Score(QS World University Rankings by Quacquarelli Symonds).

### Notes

- Subject of study is the top 500 universities selected by QS.
  - There are many of rankings from various institutions. However, we select QS as our data because, according to Alexa data, QS is the world's most-viewed global university rankings.
- In this project, the total number of Header, Sub-Header, Image, and words in the document will be collected by Web Crawler(Made of Python BeautifulSoup4, Selenium).
  - In this project, a regression analysis(Simple Linear Regression, Multiple Linear Regression, Polynomial Regression) between QS ranking/Score Data and collected data from Wikipedia will be carried out(If enough data is available, diverse Deep Learning algorithms will also be applied).
  - Also, the purpose of this project is to visualize the analysis above.

## Contributing

---

- Pull requests are welcome. For major changes, please open an issue first to discuss what you would like to change.
- Please make sure to update tests as appropriate.
- Contact : [saanghyuk@gmail.com](mailto:saanghyuk@gmail.com)(Sanghyuk Son/孙祥赫, RUCY TECH Co., Ltd, Shenzhen, China)

## Using Libraries

---

- Pandas
- Numpy
- Matplotlib
- Seaborn
- WordCloud

- BeautifulSoup4
- TensorFlow(if possible)