# Final Homework

# The JLU news Spider (60%)

* You should design and implement a web spider to crawl the OA system.
* Starter URL: <https://oa.jlu.edu.cn/defaultroot/PortalInformation!jldxList.action?channelId=179578>
* The date range: 2017-01-01~2018-06-01
* The information you crawled should include the following information.
  + Title
  + Submission date
  + Submission department
  + Main body of the news
* The technology may use
  + RegExp
  + Multi-thread
  + String Handle
  + File I/O
* The results are indexed by the submission date
  + One date one directory
    - Named by the data like 2018-05-21
  + One news one file
    - named by the title
    - saved in the same directory
* Simple analysis of the results
  + The total amount of the crawled news. The more news your crawl, the more score you will get.
  + The total amount news of each week and shown by curve [1]. If possible, divided by department.
  + The average amount news of each day
  + The average amount news of each weekday, it’s better to give a boxplot [2].
  + The average amount news of each department, it’s better to give a boxplot 2].
  + Other Statistics data you interested…

# The Word Cloud Plot of the results (40%)

1. Segment the news by Jieba [3].
2. Delete the stop words
3. Extract the keywords of each news by TF-INF and TextRank (Jieba).
4. Demonstrate one day’s news by word cloud plot and the scores (D3 [4-5]).
5. Design a web page to demonstrate all the results indexed by date.

# The Final thesis

* You should submit the following result to our system.
  + The spider codes
  + The demonstration web page
  + The thesis
  + The results files should be uploaded to your private cloud server and submit the downloading url!
* It’s a teamwork.
  + Each team has 1~5 students.
  + Every group has a leader.
  + The leader should specific each member’s contribution and give the percentage.
* Your thesis for should include:
  + Title Page
  + Abstract
  + Table of Contents (optional)
  + Chapter One - Introduction
  + Chapter Two - Review of Literature (optional)
  + Chapter Three - Methods
  + Chapter Four - Data Analysis and Results
  + Chapter Five - Conclusion
  + References

# Reference

1. <https://bl.ocks.org/mbostock/3884955>
2. <https://bl.ocks.org/mbostock/4061502>
3. <https://github.com/anderscui/jieba.NET/>
4. <https://www.jasondavies.com/wordcloud/>
5. <https://d3js.org/>
6. "GitHub - zlzforever/DotnetSpider." https://github.com/zlzforever/DotnetSpider.
7. "Web scraping - Wikipedia." https://en.wikipedia.org/wiki/Web\_scraping.
8. "GitHub - code4craft/webmagic: A scalable web crawler framework for ...." https://github.com/code4craft/webmagic.
9. "Scrapy." https://scrapy.org/.
10. "Plagiarism - Wikipedia." https://en.wikipedia.org/wiki/Plagiarism.
11. "Programming style - Wikipedia." https://en.wikipedia.org/wiki/Programming\_style.
12. "Viewing the history of your project - GitHub Help." https://help.github.com/desktop/guides/contributing/viewing-the-history-of-your-project/.
13. "Webscraping with C# - CodeProject." 20 Oct. 2015, https://www.codeproject.com/Articles/1041115/Webscraping-with-Csharp.