|  |  |
| --- | --- |
| Project Status Report |  |

# Project Name: Science & Technology Search Engine

January 31, 2017 Team Members: Liyi Li, Yunyan Wu, Jie He, Jinjin Xiao, Yi Luo

|  |  |
| --- | --- |
| The project, during the week of 1/23/2017 – 1/31/2017, is due to the following: | 1, Start to learn AWS and Search Engine development  2, Try to deal with big data problem  3, Collect domain URLs using python  4, Extract plaintext from WET |
| Issues: | 1, The URLs we have collected need to revised so that we can download the text using URLs directly.  2, Need to write code in order to download text from URLs automatically.  4, When downloaded text from URLs found the files were too big to save in the local, which is estimated to be 400 million G  5, Meet the big data problem |
| Milestones accomplished the week of 1/23/2017 - 1/31/2017: | 1, Successfully used python to get all the needed URL list from common crawl  2, Successfully used python to get a sample of wet file which stored all the text information on web page from common crawl  3, Discussed with Pro. Zhou on how to handle big data |
| Milestones planned this week, but not achieved with variance: | 1, Get all the wet file files for target domain |
| Milestones planned for next week: | 1, Find a way to handle big data problem  2, Continue to learn open sources for search engine  3, Start to learn could service and Search Engine development, and try to deal with big data problem |
| Areas/questions for discussion: | How to get all the wet file without storing in local laptop |
| Last week’s issues forwarded to this week: | How to handle big data problem |