Progress Report

Team

- Xiaoyu Zhu, xiaoyu23 (Captain)
- Cheng Wei, chengw5
- Jiaxi Wu, jiaxiwu4
-

1) Which tasks have been completed?

Task	**Status**	**Details**
Research for implementation details	60% completion	Found the common words data source online for the ranking function: https://www.wordfrequency.info/samples.asp Found the Facebook children story datasets: http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz
Clean and compile datasets	20% completion	Reviewed the Facebook children story datasets and performed preliminary data cleansing.
Index by sentence	10% completion	Set up the environment and drafted the index script.
Ranking function	10% completion	Completed a preliminary ranking function with the top common words. Completed the web scraping function to scrape the word definition from the online dictionary.
Compile evaluation dataset	0% completion	Not started.
Project Report and Demo	0% completion	Not started.

2) Which tasks are pending?

See above table for details.

3) Are you facing any challenges?

For the datasets, we managed to find a Facebook children story datasets as a starting point and reviewed the data structure. We have compiled the datasets and completed some preliminary data cleansing, including removing short sentences, removing numbers and removing duplicate sentences.

For the ranking function, we are still researching to determine the best approach. For now, we have decided to use the top 5000 common words and sentence structure difficulty to determine the difficulty of the resulting sentences and rank the results.