> Jiahui Tang

1. MapReduce Programming (75 points)

MapReduce is more of a framework than a tool. You have to fit your application into the execution pattern of map and reduce, which in some situations might be challenging. Design patterns can make application design and development easier allowing problems to be solved in a reusable and general way. In these exercises we will practice some of the most frequent **MapReduce programming patterns that have been described in class**. Exercises from 1.1 to 1.5 can be tried locally, 1.6 requires a EMR AWS cluster. Your scripts will be tested using python version 2.7, and do not use additional modules.

1.1. Distributed Grep (10 points)

Develop a distributed version of the grep tool to search for words in very large documents. Use the design patterns explained in class. The output should be the lines that match a given pattern. You can use as an input file the input text used in the word count example described in class (eBook of Moby Dick). We expect to be able to run your scripts with the following command, but you can specify any necessary modifications in your writeup along with a concise explanation on your solution:

```
$ ./P11_mapper.py Web < input.txt | sort | ./P11_reducer.py
```

Submission

- `P11_mapper.py`: Mapper script
- `P11_reducer.py`: Reducer script
- `P11.pdf`: The command line that you used to execute the job and any information required to reproduce the execution

Command Line Used:

```
./P11_mapper.py Web < input.txt | sort | ./P11_reducer.py
```

It supports Regex pattern matching as well:

```
./P11_mapper.py "Web|web|WEB" < input.txt | sort | ./P11_reducer.py
```

```
admin@C02D36V0ML85 HWC % ./P11_mapper.py Web < input.txt | sort | ./P11_reducer.py
"Nantucket itself," said Mr. Webster, "is a very striking and peculiar
Most people start at our Web site which has the main PG search facility:
Please check the Project Gutenberg Web pages for current donation
This Web site includes information about Project Gutenberg-tm,
never attained to that erudition; Noah Webster's ark does not hold it.
admin@C02D36V0ML85 HWC %
```

```
admin@C02D36V0ML85 HWC % ./P11_mapper.py "Web|web|WEB" < input.txt | sort | ./P11_reducer.py
"Nantucket itself," said Mr. Webster, "is a very striking and peculiar
--REPORT OF DANIEL WEBSTER'S SPEECH IN THE U. S. SENATE, ON THE
Most people start at our Web site which has the main PG search facility:
Please check the Project Gutenberg Web pages for current donation
The entire member seems a dense webbed bed of welded sinews; but cut
This Web site includes information about Project Gutenberg-tm,
and the Foundation web page at http://www.pglaf.org.
cable, a telegraph wire, or a strand of cobweb, it is all the same.
information can be found at the Foundation's web site and official
never attained to that erudition; Noah Webster's ark does not hold it.
posted on the official Project Gutenberg-tm web site (www.gutenberg.org),
rolling; for in Dan. HVALT is arched or vaulted." --WEBSTER'S DICTIONARY
admin@C02D36V0ML85 HWC %
```