

mrjob: part I

steve johnson - steve@steveasleep.com

Reference slide comes first

- <http://mrjob.readthedocs.org/> (very latest)
- <http://github.com/yelp/mrjob> (read the source if you get bored)
- <http://www.python.org/dev/peps/pep-0008/> (if you want to be nice to me when I read your code)

Setup

1. Have a Github account (email your username to steve+bigdive@steveasleep.com)
2. clone https://github.com/irskep/mrjob_course on Github
3. `git clone https://github.com/YOURNAME/mrjob_course.git && cd mrjob_course`
4. `pip install -r requirements.txt`
5. http://www.yelp.com/dataset_challenge: Get The Data
6. Put it in `mrjob_course/data/yelp`,
run `make` from `mrjob_course`

MapRewhaat?

“map”

- Input looks like this:

```
<key>\t<value>\n
```

- In Python, your function looks like this:

```
def mapper(self, key, value):  
    # as many times as you want  
    yield <new key>, <new value>
```


“reduce”

- Input looks like this:

```
<key1>\t<value>\n
<key1>\t<value>\n
<key2>\t<value>\n
<key2>\t<value>\n
```

- In Python, your function looks like this:

```
def reducer(self, key, values):
    # values for the same key
    for value in values:
        # do something
    yield <new key>, <new value>
```


Exercises!



Protocols

Forward dive\n

Twisting dive\n

Small dive\n

- No key
- Raw lines of text
- Use RawValueProtocol in mrjob
 - first task sees None for key

Protocols

```
dive_type\tForward dive\n  
dive_type\tTwisting dive\n  
dive_type\tSmall dive\n
```

- Raw string key separated from value by \t
- Raw string value
- Use RawProtocol in mrjob

Protocols

```
“dive”\t{“type”: “forward”}\n“dive”\t{“type”: “twisting”}\n“dive”\t{“type”: “small”}\n
```

- JSON-encoded keys and values separated by \t
- Use JSONProtocol in mrjob
- JSONValueProtocol exists for reading or writing JSON data without keys

Exercises!

