Running hadoop

Hadoop má pouze python verze 2.7 (viz níže)
Velice se klade důraz na to, co je uvedeno v deklaraci interpreteru:
`#!/usr/bin/python2.7` a nesmí být pak znak „r“ (CR) ale rovnou „n“ (LF)
=> provést konverzi z windows řádek na linuxové

1.1 Nakopírování dat

```
$ hadoop fs -mkdir /data
$ hadoop fs -cp file:///mnt/DAT500/oneline.txt /data/
```

1.2 Mazání dat

```
$ hadoop fs -rm /data/data.txt
$ hadoop fs -rmr /data
```

1.3 Spuštění úlohy provedeme

```
$ hadoop jar  /contrib/streaming/hadoop-streaming.jar -mapper mapper.py -reducer reducer.py
-input /data/bg.txt -output /data/out10 -file /mnt/DAT500/mapper.py -file /mnt/DAT500/reducer.py
```

- mapper nemusí být uvedený s cestou
- reducer taky ne
- input jsou data, která jsme předtím zkopírovali na hadoop
- output je adresář, který se vytvoří na hdfs
- file mapuje soubory, které bude hadoop používat

1.4 Uložení dat

```
$ hadoop fs -cp /data/output/data s3://uisbucket/group-4/
```

1.5 Testování mapperu a reduceru lokálně

```
$ echo "věta" | ./mapper.py | ./reducer.py
```

1.6 Čtení UTF-8 stdin v python2.7

```
import codecs
sys.stdin = codecs.getreader("utf-8")(sys.stdin)
```

1.7 Psaní UTF-8 stdout v python2.7

```
print ( "%s %s" % ( str ( elem ) .  encode ( "utf-8" ), ngrams[elem] ) )
```