

# Insert Assignment Title Here

## 02807 Computational Tools for Big Data

Anonymous authors

Insert hand in date here

### 1 Exercise 1.1

The following pipeline:

- 1 Deletes all punctuation, commas and quotes from file
- 2 Translates whitespace to newline
- 3 Sorts it
- 4 Counts occurrence of each word
- 5 Sorts it numerically in reverse (largest number first)
- 6 Prints the top 10 lines

```
tr -d ",.'" < test | tr ' ' '\n' | sort | uniq -c | sort -n -r | head -n 10
```

### 2 Exercise 1.2

The following unix script deletes all lines that contains a number with 5 or more digits

```
sed "/[0-9]\{5,\}/d" < test2
```

### 3 Exercise 1.3

The following pipeline:

- 1 Translates all tabs into spaces in the shakespeare.txt file
- 2 Removes all characters satisfying [ ^ a-zA-Z ]
- 3 Translates all spaces to newlines
- 4 Translates upper case to lower case
- 5 Sorts the lines
- 6 Keeps only unique lines
- 7 Uses dict file as plain string to match on the entire individual lines and print only the lines that don't match anything in dict.
- 8 counts the lines i.e. the misspelled words.

```
tr '\t' ' ' < shakespeare.txt | sed 's/[ ^a-zA-Z ]//g' | tr ' ' '\n' | tr A-Z a-z |  
sort | uniq | grep -F -x -v -f dict | wc -l
```

## 4 Exercise 1.4

We chose to use gbar instead of AWS.

## 5 Exercise 1.5

Git pull on gbar:

```
gbarlogin1(s152165) $ git pull
remote: Counting objects: 15, done.
remote: Compressing objects: 100% (10/10), done.
remote: Total 15 (delta 4), reused 1 (delta 1), pack-reused 2
Unpacking objects: 100% (15/15), done.
From https://github.com/ttsoftware/computational-tools
   1d522e1..407cabb  master    -> origin/master
Merge made by recursive.
 ENyYffaq.txt | 40 ++++++
 comm         | 1 +
 matrix3      | 3 +++
 week2.py     | 52 ++++++
 week3.py     | 15 ++++++
5 files changed, 111 insertions(+), 0 deletions(-)
create mode 100644 ENyYffaq.txt
create mode 100644 comm
create mode 100644 matrix3
create mode 100644 week2.py
create mode 100644 week3.py
~/computational-tools
```

## 6 Exercise 2.1

## 7 Exercise 2.2

## 8 Exercise 2.3

## 9 Exercise 3.1

## 10 Exercise 3.2

## 11 Exercise 3.3

## 12 Exercise 3.4

## 13 Exercise 3.5

## 14 Exercise 4.1