Section 2:Read in the data

The following code assumes that the dataset `spooky.csv` lives in a `data` folder (and that we are inside a `doc` folder).

Step 1: Using spooky

```
spooky<-read.csv('../data/spooky.csv',as.is=T)
```

An overview of the data structure and content

Let's first remind ourselves of the structure of the data.

```
dim<-dim(spooky)
dim
```

```
## [1] 19579     3
```

```
head(spooky)
```

```
##        id
## 1 id26305
## 2 id17569
## 3 id11008
## 4 id27763
## 5 id12958
## 6 id22965
##
## 1
## 2
## 3
## 4
## 5
## 6 A youth passed in solitude, my best years spent under your gentle and feminine fosterage, has so re
##   author
## 1    EAP
## 2    HPL
## 3    EAP
## 4    MWS
## 5    HPL
## 6    MWS
```

```
summary(spooky)
```

```
##       id                 text              author
##  Length:19579       Length:19579       Length:19579
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
sum(is.na(spooky))
```

```
## [1] 0
```

```
spooky$author<-as.factor(spooky$author)
unique(spooky$author)
```

```
## [1] EAP HPL MWS
## Levels: EAP HPL MWS
```

When we look into spooky data set, it is a 19579 rows and 3 columns dataset. Each row correspoding a unique id number, an excerpt of texts, and author name. Addtionally, there are no missing values. There are three authors, Like `HPL` is Lovecraft, `MWS` is Shelly, and `EAP` is Poe.