ETC3555 2018 - Lab 11

Text mining

Cameron Roach and Souhaib Ben Taieb

06 October, 2018

In this lab, you will perform text mining using the Harry Potter books available in the *harrypotter* package.

Note: The book "Text Mining with R" is available at the following link: https://www.tidytextmining.com/tidytext.html

Harry Potter books

You can load the harrypotter package with the following code:

```
if (packageVersion("devtools") < 1.6) {
  install.packages("devtools")
}

devtools::install_github("bradleyboehmke/harrypotter")
```

You now have access to the following books:

1. **philosophers_stone**: Harry Potter and the Philosophers Stone, published in 1997
2. **chamber_of_secrets**: Harry Potter and the Chamber of Secrets, published in 1998
3. **prisoner_of_azkaban**: Harry Potter and the Prisoner of Azkaban, published in 1999
4. **goblet_of_fire**: Harry Potter and the Goblet of Fire, published in 2000
5. **order_of_the_phoenix**: Harry Potter and the Order of the Phoenix, published in 2003
6. **half_blood_prince**: Harry Potter and the Half-Blood Prince, published in 2005
7. **deathly_hallows**: Harry Potter and the Deathly Hallows, published in 2007

Each book is represented as a vector of strings where each string is one chapter of the book. The number of chapters for each book is given below:

```
library(harrypotter)
length(philosophers_stone)
```

```
## [1] 17
```

```
length(prisoner_of_azkaban)
```

```
## [1] 22
```

```
length(goblet_of_fire)
```

```
## [1] 37
```

```
length(order_of_the_phoenix)
```

```
## [1] 38
```

```
length(half_blood_prince)
```

```
## [1] 30
```

```
length(deathly_hallows)
```

```
## [1] 37
```

Question 1

For each book, plot the frequency of each word as a function of its rank (in log-log scale) to illustrate the Zipf's law.

Question 2

Remove stop words and perform word frequency analysis for the *half_blood_prince* book by identifying:

1. the top 10 most common words across the entire book
2. the top 5 most common words for each chapter
3. the top 3 rarest words in the third chapter

Question 3

Produce one plot that shows the top 10 most common trigrams in each book (witout removing stopwords).

Question 4

The same question as Question 3 but now remove stopwords as explained in Chapter 4 of the book "Text mining in R". Briefly explain the difference with your solution to Question 3.

TURN IN

- Your `.Rmd` file (which should knit without errors and without assuming any packages have been pre-loaded)
- Your Word (or pdf) file that results from knitting the Rmd.
- DUE: October 21, 11:55pm (late submissions not allowed), loaded into moodle.