Pdf\_text

Vijay Mudivedu

2018-06-11

## PDF to text

## Loading required package: RColorBrewer

## ── Attaching packages ────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 2.2.1 ✔ purrr 0.2.4  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.4  
## ✔ tidyr 0.8.0 ✔ stringr 1.3.0  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

* Download the data from google drive

## Auto-refreshing stale OAuth token.

## File downloaded:  
## \* JavaBasics-notes.pdf  
## Saved locally as:  
## \* JavaBasics-notes.pdf

* Store the data in the

## chr [1:23] " Java Basics\nJava Basi"| \_\_truncated\_\_ ...

### Data Cleaning

* Remove the “Java Basics” header, special characters, the Footers
* Combining all the text from different pages.

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 1

#### Further Cleaning of text

* Lower case
* Removing punctuations
* Removing white spaces
* Removing numbers
* Removing stopwords
* Creating Document term-matirx for the cleaned text
* Most frequent terms are:

## [1] "allocate" "applet" "applets" "array" "browser"   
## [6] "button" "can" "class" "code" "comments"   
## [11] "data" "example" "following" "garbage" "int"   
## [16] "java" "language" "may" "memory" "method"   
## [21] "new" "null" "object" "objects" "pointer"   
## [26] "primitive" "program" "public" "return" "runtime"   
## [31] "stack" "string" "types" "use" "void"

* Frequency of the most frequent terms

## java new data code int   
## 75 52 39 38 34   
## applet class button array objects   
## 30 29 28 25 24   
## method object public example string   
## 23 23 22 19 19   
## language memory null return types   
## 15 15 15 15 14   
## use void comments browser may   
## 14 14 13 12 12   
## primitive program allocate can applets   
## 12 12 11 11 10   
## following garbage pointer runtime stack   
## 10 10 10 10 10   
## adata applications called file make   
## 9 9 9 9 9   
## methods operator reference system two   
## 9 9 9 9 9   
## byte collection init programs stat   
## 8 8 8 8 8

* creating the word dataframe

## words freq  
## 1 java 75  
## 2 new 52  
## 3 data 39  
## 4 code 38  
## 5 int 34  
## 6 applet 30

* Printing the Keywords In Java\_basic in the text file. 