Project 2

Suman Paudel

2024-03-24

```
# Load the packages
library(pdftools)
library(tm)
library(magrittr)
library(tibble)
library(wordcloud, warn.conflicts = F)
library(Rgraphviz, warn.conflicts = F)
library(graph, warn.conflicts = F)
```

```
files <- list.files(pattern = "pdf$")
files
```

```
## [1] "33-Project-2.pdf"
## [2] "Suman paudel - Project 2_ Unit 2.pdf"
```

```
# load the pdf files into list
pdf_files <- lapply(files, pdf_text)
```

```
# create a corpus from vector source i.e from list pdf_files
corpus <- Corpus(VectorSource(unlist(pdf_files)))

# copy the loaded corpus
corpus_copy <- corpus

# inspect first few texts of corpus
inspect(corpus[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1]                                          Project 2\n\n                                                  Su
## [2] Project 2: Unit 2 (Deadline 23 March 2024 11:59pm)\n\nThis project has 3 parts and you need to d
```