fb_babi_data

R Markdown

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Download here: https://venturebeat.com/2016/02/18/facebook-releases-1-6gb-data-set-of-childrens-stories-for-training-its-ai/

Preposition data Look at read me and file sizes “Data is in the included”data" folder. Questions are separated according to whether the missing word is a named entity (NE), common noun (CN), verb (V) or preposition (P). The POS/NER was done by Stanford CoreNLP and nothing else. Thus the dataset consists of the following files:"

“Data is in the included”data" folder. Questions are separated according to whether the missing word is a named entity (NE), common noun (CN), verb (V) or preposition (P). The POS/NER was done by Stanford CoreNLP and nothing else. Thus the dataset consists of the following files:

cbtest_NE_train.txt : 67128 questions cbtest_NE_valid_2000ex.txt : 2000 cbtest_NE_test_2500ex.txt : 2500

cbtest_CN_train.txt : 121176 questions cbtest_CN_valid_2000ex.txt : 2000 cbtest_CN_test_2500ex.txt : 2500

cbtest_V_train.txt : 109111 questions cbtest_V_valid_2000ex.txt : 2000 cbtest_V_test_2500ex.txt : 2500

cbtest_P_train.txt : 67128 questions cbtest_P_valid_2000ex.txt : 2000 cbtest_P_test_2500ex.txt : 2500

These files, the text is in the format of CBT questions (see below). We also release the raw text from test / valid / train sets, again tokenised by Stanford Core NLP. These files have the form cbt_{train, valid, test}.txt

Detailed stats of all question types are given in the “stats” folder."

```
cb_data_test <- "~/Downloads/CBTest/data/cbtest_P_train.txt"
prep_data <- read_tsv(cb_data_test, col_names = F)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_character()
## )
```

```
## Warning: 334030 parsing failures.
## row col  expected    actual                                          file
```

```
##  21  -- 1 columns 4 columns '~/Downloads/CBTest/data/cbtest_P_train.txt'
##  42  -- 1 columns 4 columns '~/Downloads/CBTest/data/cbtest_P_train.txt'
##  63  -- 1 columns 4 columns '~/Downloads/CBTest/data/cbtest_P_train.txt'
##  84  -- 1 columns 4 columns '~/Downloads/CBTest/data/cbtest_P_train.txt'
## 105  -- 1 columns 4 columns '~/Downloads/CBTest/data/cbtest_P_train.txt'
## ... ... ......... ......... ........................................
## See problems(...) for more details.
```

```
prep_data
```

```
## # A tibble: 7,014,630 x 1
##    X1
##    <chr>
##  1 1 CHAPTER I. -LCB- Chapter heading picture : p1.jpg -RCB- How the Fairies we~
##  2 2 Once upon a time there reigned in Pantouflia a king and a queen .
##  3 3 With almost everything else to make them happy , they wanted one thing : t~
##  4 4 This vexed the king even more than the queen , who was very clever and lea~
##  5 5 However , she , too in spite of all the books she read and all the picture~
##  6 6 The king was anxious to consult the fairies , but the queen would not hear~
##  7 7 She did not believe in fairies : she said that they had never existed ; an~
##  8 8 Well , at long and at last they had a little boy , who was generally regar~
##  9 9 Even her majesty herself remarked that , though she could never believe al~
## 10 10 Now , the time drew near for the christening party , and the king and que~
## # ... with 7,014,620 more rows
```

Load the wordbank words

```
base::load("~/GitHub_C/wordbank-book/data/psychometrics/eng_ws_mods_4pl.Rds")
```

```
## Warning: namespace 'mirt' is not available and has been replaced
## by .GlobalEnv when processing object 'mod_4pl'
```

```
wordbank_words <- coefs_4pl %>% pull(definition)
```

Filter for rows containing wordbank words

```
prep_data %>%
  filter(
    str_detect(X1, wordbank_words)
  )
```

```
## Warning in stri_detect_regex(string, pattern, negate = negate, opts_regex =
## opts(pattern)): longer object length is not a multiple of shorter object length
```

```
## # A tibble: 181,069 x 1
##    X1
##    <chr>
##  1 1 CHAPTER I. -LCB- Chapter heading picture : p1.jpg -RCB- How the Fairies we~
##  2 11 It was a splendid room , hung with portraits of the royal ancestors .
##  3 12 There was Cinderella , the grandmother of the reigning monarch , with her~
##  4 20 `` People are so touchy on these occasions , '' said his majesty .
##  5 2 However , she , too in spite of all the books she read and all the picture~
##  6 12 replied the queen ; for the king 's aunts were old-fashioned , and did no~
##  7 13 `` They are very old friends of our family , my dear , that 's all , '' s~
##  8 16 `` Your grandmother ! ''
##  9 13 One , in particular , was most kind and most serviceable to Cinderella I.~
## 10 17 If anyone puts such nonsense into the head of my little Prigio -- '' But ~
## # ... with 181,059 more rows
```