Data Pre-processing II (Text Data)

Applied Machine Learning for Educational Data Science

true

08/23/2021

Contents

[Updated: Wed, Sep 08, 2021 - 22:04:42]

1. Importing Data

```
readability <- read.csv('https://raw.githubusercontent.com/uo-datasci-specialization/c4-ml-fall-2021/ma
                        header=TRUE)

str(readability)
```

Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the udpipe and quanteda

Morphological annotation

Morphological features

Syntactic annotation

Word length

Measures of lexical variety

Measures of readability

```
require(udpipe)
require(quanteda)
require(nsyllable)

model  <- udpipe_download_model(language = "english")

text <- as.character(train_df[i,]$excerpt)


  tokenized <- tokens(text)
  dm <- dfm(tokenized)


  annotated <- udpipe_annotate(ud_eng, x = text)
  annotated <- as.data.frame(annotated)

  # cbind_morphological(annotated, term = "feats", which = "lexical")

  # Morphological annotation (universal POS tags, https://universaldependencies.org/u/pos/index.html)

  temp     <- data.frame(table(annotated$upos))
  temp[,2] <- temp[,2]

  words <- annotated[!annotated$upos%in%c('PUNCT','SYS','X'),]$token

  temp[,1] <- as.character(temp[,1])
  temp <- rbind(temp,data.frame(table(annotated$xpos)))

  temp <- rbind(temp,data.frame(Var1 ='nwords',Freq=length(words)))
  temp <- rbind(temp,data.frame(Var1 ='nchars',Freq=sum(nchar(annotated$token))))
  temp <- rbind(temp,data.frame(Var1 ='nchars',Freq=sum(nchar(annotated$token))/length(words)))
  temp <- rbind(temp,data.frame(Var1 ='wdiv',Freq=length(unique(words))/length(words)))
  temp <- rbind(temp,data.frame(Var1 ='nsent',Freq=length(unique(annotated$sentence_id))))


  # Morphologicla features (https://universaldependencies.org/u/feat/index.html)
```

```
feats  <- na.omit(annotated$feats)
feats1 <- unlist(strsplit(feats,split='\\|'))
feats2 <-  unlist(strsplit(feats1,split='='))[c(TRUE,FALSE)]

feats1        <- table(feats1)
names(feats1) <- gsub('=','.',names(feats1))

feats2        <- table(feats2)
names(feats2) <- names(feats2)


temp <- rbind(temp,data.frame(feats1))
temp <- rbind(temp,data.frame(feats2))


# Syntactic Annotation (https://universaldependencies.org/u/dep/index.html)

temp <- rbind(temp,data.frame(table(annotated$dep_rel)))

# Word Length distribution

wl <- table(nchar(tokens(text,
                         remove_punct = TRUE,
                         remove_numbers = TRUE,
                         remove_symbols = TRUE,
                         remove_separators = TRUE)[[1]])
)

names(wl) <- paste0('l',names(wl))


label_let <- names(wl)

ifelse('l1'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l1',Freq=wl['l1'])),
       temp <- rbind(temp,data.frame(Var1 ='l1',Freq=0)))

ifelse('l2'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l2',Freq=wl['l2'])),
       temp <- rbind(temp,data.frame(Var1 ='l2',Freq=0)))

ifelse('l3'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l3',Freq=wl['l3'])),
       temp <- rbind(temp,data.frame(Var1 ='l3',Freq=0)))

ifelse('l4'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l4',Freq=wl['l4'])),
       temp <- rbind(temp,data.frame(Var1 ='l4',Freq=0)))

ifelse('l5'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l5',Freq=wl['l5'])),
       temp <- rbind(temp,data.frame(Var1 ='l5',Freq=0)))
```

```
ifelse('l6'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l6',Freq=wl['l6'])),
       temp <- rbind(temp,data.frame(Var1 ='l6',Freq=0)))

ifelse('l7'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l7',Freq=wl['l7'])),
       temp <- rbind(temp,data.frame(Var1 ='l7',Freq=0)))

ifelse('l8'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l8',Freq=wl['l8'])),
       temp <- rbind(temp,data.frame(Var1 ='l8',Freq=0)))

ifelse('l9'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l9',Freq=wl['l9'])),
       temp <- rbind(temp,data.frame(Var1 ='l9',Freq=0)))

ifelse('l10'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l10',Freq=wl['l10'])),
       temp <- rbind(temp,data.frame(Var1 ='l10',Freq=0)))

ifelse('l11'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l11',Freq=wl['l11'])),
       temp <- rbind(temp,data.frame(Var1 ='l11',Freq=0)))

ifelse('l12'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l12',Freq=wl['l12'])),
       temp <- rbind(temp,data.frame(Var1 ='l12',Freq=0)))

ifelse('l13'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l13',Freq=wl['l13'])),
       temp <- rbind(temp,data.frame(Var1 ='l13',Freq=0)))

ifelse('l14'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l14',Freq=wl['l14'])),
       temp <- rbind(temp,data.frame(Var1 ='l14',Freq=0)))

ifelse('l15'%in%label_let,
       temp <- rbind(temp,data.frame(Var1 ='l15',Freq=wl['l15'])),
       temp <- rbind(temp,data.frame(Var1 ='l15',Freq=0)))

# Measures of Lexical variety)

lexical <- textstat_lexdiv(tokenized,measure = 'all')[1,2:16]

temp <- rbind(temp,data.frame(Var1 = colnames(lexical),Freq = as.numeric(lexical[1,])))

# Measures of Readability

readable <- textstat_readability(text, measure = 'all')[1,2:49]
temp <- rbind(temp,data.frame(Var1 = colnames(readable),Freq = as.numeric(readable[1,])))


list1[[i]] <- data.frame(t(temp[,2]))
```

```
  colnames(list1[[i]]) <- temp[,1]

print(i)
}
```

Word Embeddings

Text package installation and info

https://www.r-text.org/

https://www.r-text.org/articles/Word%20embeddings.html

Preparing Environment

```
require(reticulate)

# List the available Python environments

virtualenv_list()

# Import the modules

reticulate::import('torch')
reticulate::import('numpy')
reticulate::import('transformers')
reticulate::import('nltk')
reticulate::import('tokenizers')

# Load the text package

require(text)
```

```
txt1 <- as.character('This is a great class! The lectures are very well organized. Grading is fair. I l
txt1

txt2 <- as.character('I hate this class. There is no organization and lectures are too boring. The assi
txt2

# first hidden layer

tmp1 <- textEmbed(x = txt1,model = 'roberta-base',layers = 1,contexts=TRUE)
tmp2 <- textEmbed(x = txt2,model = 'roberta-base',layers = 1,contexts=TRUE)

tmp1$x
tmp2$x

# Concatenating the last four hidden layer

tmp1 <- textEmbed(x = txt1,model = 'roberta-base',layers = 9:12,contexts=TRUE)
```

```
tmp2 <- textEmbed(x = txt2,model = 'roberta-base',layers = 9:12,contexts=TRUE)
tmp1$x
tmp2$x

txt1 <- as.character('phone')
tmp1 <- textEmbed(x = txt,model = 'roberta-base',layers = 9:12,contexts=TRUE)
tmp1$x
```