Contiguous_Sequence_Mining

Gautham Meenakshisundaram

12/16/2019

Mining Contiguous Sequential Patterns in Text

Data

The provided input file ("reviews_sample.txt") consists of 10,000 online reviews from Yelp users. The reviews have been stemmed (to remove the postfix of each word so words with similar semantics can have the same form), and most of the punctuation has been removed. Therefore, each line is basically a list of strings separated by spaces.

```
filepath = "C:/my_repo/Data Science/44468423_Non-
Formal_Unspecified_RPL/Assignments/Pattern Discovery in Data
Mining_Assignment2/reviews_sample.txt"
```

Output

An algorithm is to be implemented to mine contiguous sequential patterns that are frequent in the input data. A contiguous sequential pattern is a sequence of items that frequently appears as a consecutive subsequence in a database of many sequences.

```
library('CSeqpat')
```

Extract all the frequent contiguous sequential patterns that have an absolute support no smaller than 100.

```
output <- CSeqpat(filepath, phraselenmin = 1, phraselenmax = 99999,
minsupport = 100, "\n", stopword = FALSE, stemword = FALSE, lower = FALSE,
removepunc = FALSE)

head(output)
```

```
##   Freq_Phrases Support
## 1          100     121
## 2         able     402
## 3   absolutely     392
## 4       across     261
## 5       actual      98
## 6     actually     570
```

Please write all the frequent contiguous sequential patterns along with their absolute supports into a text file named “patterns.txt”. Every line corresponds to exactly one pattern you found and should be in this format: support:item_1;item_2;item_3

```
library(splitstackshape)

df1 <- data.frame(output[1],output[2])
head(df1)

##   Freq_Phrases Support
## 1          100     121
## 2         able     402
## 3   absolutely     392
## 4       across     261
## 5       actual      98
## 6     actually     570

df2 <- cSplit(df1, 'Freq_Phrases', sep=" ")

df2$concat <-
paste(df2$Support,":",df2$Freq_Phrases_1,";",df2$Freq_Phrases_2)

df2$concat <- gsub("; NA","", df2$concat)
df2$concat <- gsub(" ","", df2$concat)

head(df2)

##    Support Freq_Phrases_1 Freq_Phrases_2         concat
## 1:     121            100           <NA>        121:100
## 2:     402           able           <NA>       402:able
## 3:     392     absolutely           <NA> 392:absolutely
## 4:     261         across           <NA>     261:across
## 5:      98         actual           <NA>      98:actual
## 6:     570       actually           <NA>   570:actually
```

Saving as text file

```
writeLines(df2$concat,"patterns.txt")
```