Scrap Discuss HK

Paine (HE Peiwei, 54471952)

First, read the URL by using rvest, and the encoding is GBK. “Utf-8” is for OS system, but for Windows, the text should be encode in GBK.

```
library(httr)
library(rvest)
url = "http://www.discuss.com.hk/viewthread.php?tid=26488701"
r = GET(url)
text = iconv(list(r$content), from="big5", to="GBK", sub="")
page = read_html(text)
```

Second, get comments to this post by using the html node, and check the css.

```
posts = page %>% html_nodes(".viewthread")

contents = posts %>%
  html_node("span[id*='postorig']") %>%
  html_text %>% trimws
```

Third, save the content’s text to txt file.

```
write.table(contents, file = "disshk.txt", sep = ",", col.names = colnames(alltext)
)
```

Fourth, get segment the Chinese text and calculate the frequency. And convert the frequency into dataframe.

```
library(rJava)
library(Rwordseg)
```

```
## # Version: 0.2-1
```

```
library(wordcloud2)
text=segmentCN("disshk.txt",returnType = "tm")
```

```
## Output file:  disshk.segment.txt
```

```
text1=readLines("disshk.segment.txt", encoding = "UTF-8")
word = lapply(X = text1, FUN = strsplit, "\\s")
word1=unlist(word)
df=table(word1)
df=sort(df, decreasing = T)

d=data.frame(df)
newd = head(d, n=100)
newd<-newd[-1,]
head(newd)
```

```
##   word1 Freq
## 2    奴   75
## 3    是   75
## 4     3   70
## 5    阿   70
## 6     4   65
## 7    仙   65
```

Finally, output a wordcloud.

```
library(wordcloud2)
wordcloud2(newd, color = "random-light", backgroundColor = "white")
```