Information about arabicStemR
Rich Nielsen
5/28/2019

The manual: https://cran.r-project.org/web/packages/arabicStemR/arabicStemR.pdf.
It is somewhat helpful, but not exhaustive.

Transliteration

<table>
<thead>
<tr>
<th>Arabic letter</th>
<th>Transliterated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ا</td>
<td>a</td>
</tr>
<tr>
<td>ئ</td>
<td>A</td>
</tr>
<tr>
<td>ب</td>
<td>b</td>
</tr>
<tr>
<td>ت</td>
<td>t</td>
</tr>
<tr>
<td>ث</td>
<td>U</td>
</tr>
<tr>
<td>ج</td>
<td>j</td>
</tr>
<tr>
<td>ح</td>
<td>7</td>
</tr>
<tr>
<td>خ</td>
<td>K</td>
</tr>
<tr>
<td>د</td>
<td>d</td>
</tr>
<tr>
<td>ذ</td>
<td>i</td>
</tr>
<tr>
<td>ر</td>
<td>r</td>
</tr>
<tr>
<td>ز</td>
<td>z</td>
</tr>
<tr>
<td>س</td>
<td>s</td>
</tr>
<tr>
<td>ش</td>
<td>W</td>
</tr>
<tr>
<td>ص</td>
<td>S</td>
</tr>
<tr>
<td>ض</td>
<td>D</td>
</tr>
<tr>
<td>ط</td>
<td>T</td>
</tr>
<tr>
<td>ظ</td>
<td>Z</td>
</tr>
<tr>
<td>ع</td>
<td>3</td>
</tr>
<tr>
<td>غ</td>
<td>G</td>
</tr>
<tr>
<td>ف</td>
<td>f</td>
</tr>
<tr>
<td>ق</td>
<td>Q</td>
</tr>
<tr>
<td>ك</td>
<td>K</td>
</tr>
<tr>
<td>ل</td>
<td>l</td>
</tr>
<tr>
<td>م</td>
<td>m</td>
</tr>
<tr>
<td>ن</td>
<td>n</td>
</tr>
<tr>
<td>ه</td>
<td>h</td>
</tr>
<tr>
<td>و</td>
<td>w</td>
</tr>
<tr>
<td>ي</td>
<td>Y</td>
</tr>
<tr>
<td>أ</td>
<td>a</td>
</tr>
<tr>
<td>إ</td>
<td>a</td>
</tr>
<tr>
<td>ؤ</td>
<td>o</td>
</tr>
</tbody>
</table>
Stop words removed by default in stem() and removeStopWords()

Prepositions

<table>
<thead>
<tr>
<th>تا</th>
<th>تا</th>
<th>تا</th>
<th>تا</th>
<th>تا</th>
</tr>
</thead>
<tbody>
<tr>
<td>تا</td>
<td>تا</td>
<td>تا</td>
<td>تا</td>
<td>تا</td>
</tr>
</tbody>
</table>

Pronouns

<table>
<thead>
<tr>
<th>تا</th>
<th>تا</th>
<th>تا</th>
<th>تا</th>
<th>تا</th>
</tr>
</thead>
<tbody>
<tr>
<td>تا</td>
<td>تا</td>
<td>تا</td>
<td>تا</td>
<td>تا</td>
</tr>
</tbody>
</table>
Particles and connectors

Prefix removal

Only one prefix is removed. They are evaluated in this order – after the first prefix match is found and removed, the stemmer moves to the next word.
"ال" if >= 4 characters  
"و café" if >= 5 characters  
"بال" if >= 5 characters  
"كما" if >= 5 characters  
"قال" if >= 5 characters  
"ال" if >= 5 characters  
"و" if >= 4 characters  

Prefix removal

Only one prefix is removed. They are evaluated in this order – after the first prefix match is found and removed, the stemmer moves to the next word.

"ها" if >= 4 characters  
"آن" if >= 4 characters  
"أات" if >= 4 characters  
"ون" if >= 4 characters  
"أين" if >= 4 characters  
"ية" if >= 4 characters  
"ية" if >= 4 characters  
"د" if >= 3 characters  
"س" if >= 3 characters  
"ي" if >= 3 characters

Is there a way to turn off parts of the stemmer without programming?
Sort of. In the commands “removePrefixes()” and “removeSuffixes()”, you can specify how long a word must be in order to remove the stem. If you set this number very high for a specific prefix or suffix (i.e, Inf), it will not remove that prefix or suffix. However, this is inside the stem() function. You would have to combine your own custom stemmer from the internal parts, which does require programming. I have an example with code here: http://www.mit.edu/~rnielsen/r%20stemmer%20example_website.R

How can I see what the stemmer is doing?

From the arabicStemR help files for stem()...

```r
# Load data
Library(arabicStemR)
data(aljazeera)

## stem and return the stemlist
out <- stem(aljazeera,returnStemList=TRUE)
```
This allows you to see which words are being combined
Interpret this as follows:
  i <- 1
  This is the i'th stem in quotes (with the original word as the label)
  out$stemlist[i]
  These are all the words that resolve to the same stem.
  names(out$stemlist)[out$stemlist==out$stemlist[i]]
  And this will provide a count.
  mytab <- table(names(out$stemlist)[out$stemlist==out$stemlist[i]])
  for(i in 1:length(mytab)){print(mytab[i])}
  Note that if you just look at "mytab", it will appear incorrect because
  R displays the Arabic labels from right to left but the numbers from left
to right (thanks R!).

This can be done for all of the stems
result <- sapply(out$stemlist,
  function(x){table(names(out$stemlist)[out$stemlist==x])})
for(i in 1:length(result)){
  cat(paste("stemmed:",out$stemlist[i],"\n"))
  cat("unstemmed:")
  print(result[[i]])
  cat("\n")
}

display the results correctly for the i'th stem
i <- 1
for(j in 1:length(result[[i]])){print(result[[i]][j])}