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

Arabic letter	Transliterated
ا	a
ى	A
ب	b
ت	t
ث	U
ج	j
ح	7
خ	K
د	d
ذ	i
ر	r
ز	z
س	s
ش	W
ص	S
ض	D
ط	T
ظ	Z
ع	3
غ	G
ف	f
ق	Q
ك	K
ل	l
م	m
ن	n
ه	h
و	w
ي	y
أ	a
إ	a
ؤ	o

ئ	5
ء	q
آ	a
ة	0
پ	p
ه	h
ک	k
ٹ	t
ں	n
ے	y
ی	y
ٱ	a
ڶ	l
ھ	h
ؿ	y
ؾ	y
ػ	k
ؼ	k

Stop words removed by default in stem() and removeStopWords()

Prepositions

"في", "فيه", "فيها", "فيهم", "على"
"عليك", "عليكم", "علينا", "عليه", "عليها"
"عليهم", "علي", "به", "بها", "بهم", "بهذا"
"بذلك", "بك", "بكم", "بكل", "بما", "بمن"
"بنا", "له", "لها", "لهم", "مع", "معه"
"معها", "معهم", "عن", "عنا", "عنه", "عنها"
"عنهم", "تحت", "حتى", "فوق", "فوقَ"
"بجانب", "أمام", "أمامَ", "امام", "خارج"
"بالخارج", "حولَ", "حول", "رغم", "بالرغم"
"رغمَ", "منذ", "منذُ", "من", "خلال"
"خلالَ", "حول", "حولَ", "قبل", "قبلَ"
"وفقا", "إلى", "الوراءَ", "وراء"
"بينَ", "بين", "بينهم", "بينهما", "بينكم"
"بينما", "بدون", "لكن", "باتجاه", "أقل"
"اقل", "اكثر"

Pronouns

"هذا", "هذه", "ذلك", "تلك", "هؤلَاء"
"هؤلاء", "اولائك", "هذان", "هذينهتان"
"هتينأنا", "انا", "أنت", "هما", "أنتَ"
"انت", "أنت", "أنتِ", "انتهو", "هوَ"

"هو", "هي", "هيَ", "نحن", "أنتم", "انتم",
"أنتم", "انتم", "هُم", "هم", "لهم", "منهم",
"وهم", "التي", "الذي", "اللذان", "اللذين",
"اللتان", "اللتين"

Particles and connectors

"ان", "وان", "إن", "إنه", "إنها",
"إنهم", "إنهما", "إني", "وإن", "وأن",
"ان", "انه", "انها", "انهم", "انهما",
"اني", "أنك", "إنك", "انك", "أنكم", "إنكم",
"انكم", "اننا", "وان", "أن", "ان",
"ألا", "بأن", "ان", "الا", "بان", "بانهم",
"أنه", "أنها", "أنهم", "أنهما", "انه",
"انها", "انهم", "انهما", "أذ", "اذ",
"اذا", "إذ", "إذا", "وإذ", "وإذا", "اذ",
"اذ", "اذا", "اذ", "اذا", "فاذا", "ماذا",
"واذ", "واذا", "لولا", "لو", "ولوسوف",
"لن", "ما", "لم", "ولم", "أما", "اما",
"لا", "ولا", "إلا", "الا", "أم", "أو",
"ام", "او", "بل", "قد", "وقد", "لقد", "أنما",
"إنما", "بل", "انما", "انما", "و"
"بما", "كما", "لما", "لأن", "لان",
"لي", "لى", "لهذأ", "لذأ", "لأنه", "لأنها",
"لأنهم", "لان", "لانه", "لانها", "لانهم",
"ثم", "أيضا", "ايضا", "كذلك", "قبل",
"بعد", "لكن", "ولكن", "لكنه", "لكنها",
"لكنهم", "فقط", "رغم", "بالرغم", "بفضل",
"حيث", "بحيث", "لكي", "هنا", "هناك",
"بسبب", "ذات", "ذو", "ذي", "ذى", "وه",
"يا", "انما", "فهذا", "فهو", "فما", "فمن",
"فيما", "فهل", "وهل", "فهؤلاء", "كذا",
"لذلك", "لماذا", "لمن", "لنا", "منا",
"منك", "منكم", "منهما", "لك",
"ولو", "مما", "وما", "ومن", "عند", "عندهم",
"عندما", "عندنا", "عنهما", "عنك", "اذن",
"الذي", "فانا", "فانهم", "فهم", "فه",
"فكل", "لكل", "لكم", "فلم", "فلما", "فيك",
"فيكم", "لهذا"

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
"وال" if >= 5 characters
"بال" if >= 5 characters
"كال" if >= 5 characters
"فال" if >= 5 characters
"لل" 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
"ان" if >= 4 characters
"ات" if >= 4 characters
"ون" if >= 4 characters
"ين" if >= 4 characters
"يه" if >= 4 characters
"ية" if >= 4 characters
"ه" if >= 3 characters
"ة" 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()…

```
# Load data
Library(arabicStemR)

data(aljazeera)

## stem and return the stemlist
out <- stem(aljazeera,returnStemList=TRUE)
```

```
out$text
out$stemlist


## 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])}
```