15BCE0517

M.S.SANJAY

L3+L4

EXPERIMENT 7

DOCUMENT SIMILARITY TEHNIQUES AND MEASUREMENTS

>library(“stringr”)

>library(“teext2vec”)

> data("movie\_review")

>

> movie\_review = movie\_review[1:500, ]

>

> prep\_fun = function(x) {

+ x %>%

+ # make text lower case

+ str\_to\_lower %>%

+ # remove non-alphanumeric symbols

+ str\_replace\_all("[^[:alnum:]]", " ") %>%

+ # collapse multiple spaces

+ str\_replace\_all("\\s+", " ")

+ }

> movie\_review$review\_clean = prep\_fun(movie\_review$review)

>

> doc\_set\_1 = movie\_review[1:300, ]

> it1 = itoken(doc\_set\_1$review\_clean, progressbar = FALSE)

>

> doc\_set\_2 = movie\_review[301:500, ]

>

> it2 = itoken(doc\_set\_2$review\_clean, progressbar = FALSE)

> it = itoken(movie\_review$review\_clean, progressbar = FALSE)

>

> v = create\_vocabulary(it) %>% prune\_vocabulary(doc\_proportion\_max = 0.1, term\_count\_min = 5)

>

> vectorizer = vocab\_vectorizer(v)

**JACCARD SIMILARITY**

> dtm1 = create\_dtm(it1, vectorizer)

>

> dim(dtm1)

[1] 300 2338

>

>

> dtm2 = create\_dtm(it2, vectorizer)

>

> dim(dtm2)

[1] 200 2338

> dim(dtm2)

[1] 200 2338

> d1\_d2\_jac\_sim = sim2(dtm1, dtm2, method = "jaccard", norm = "none")

> dim(d1\_d2\_jac\_sim)

[1] 300 200

> d1\_d2\_jac\_sim[1:2, 1:5]

2 x 5 sparse Matrix of class "dgCMatrix"

1 2 3 4 5

1 0.02142857 . 0.02362205 0.007575758 0.02597403

2 0.01219512 . 0.02941176 0.013888889 0.02083333

> dtm1\_2 = dtm1[1:200, ]

>

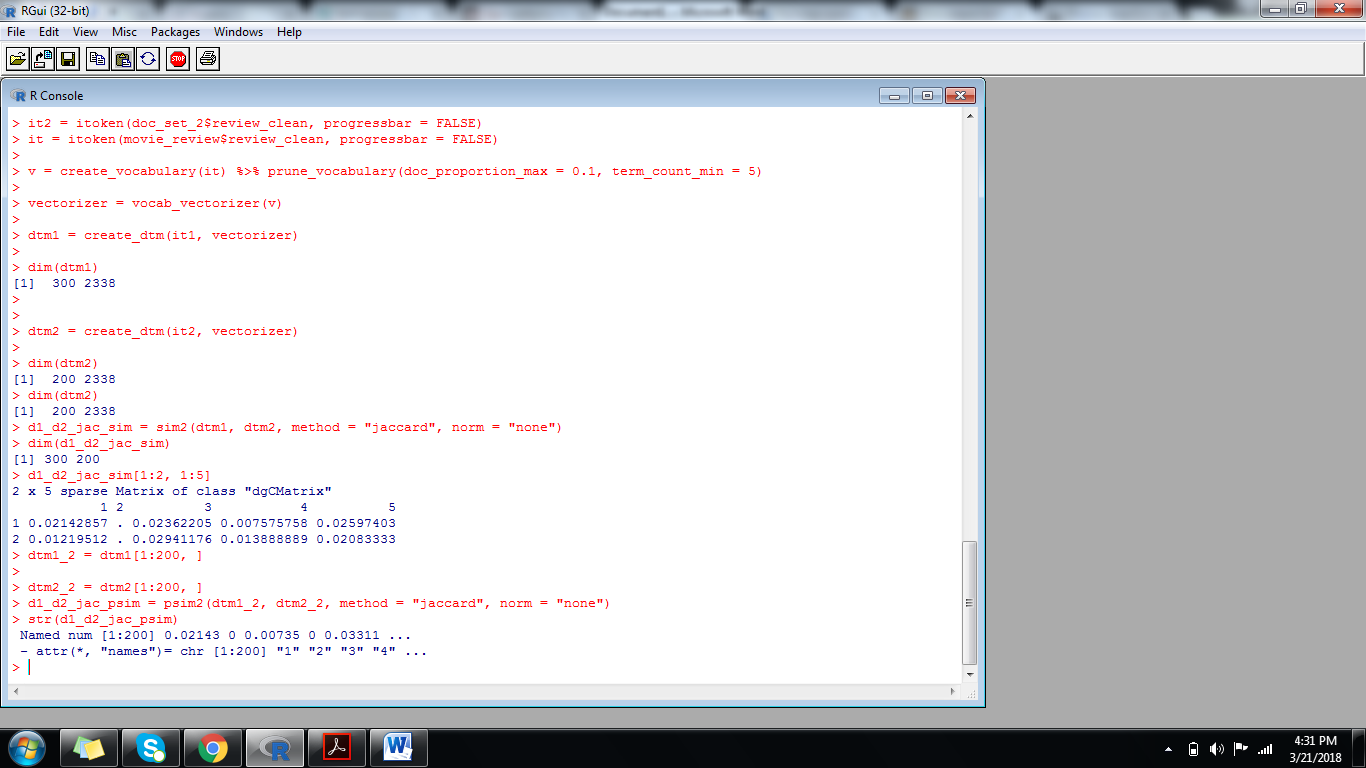
> dtm2\_2 = dtm2[1:200, ]

> d1\_d2\_jac\_psim = psim2(dtm1\_2, dtm2\_2, method = "jaccard", norm = "none")

> str(d1\_d2\_jac\_psim)

Named num [1:200] 0.02143 0 0.00735 0 0.03311 ...

- attr(\*, "names")= chr [1:200] "1" "2" "3" "4" ...



**COSINE SIMILARITY**

> d1\_d2\_cos\_sim = sim2(dtm1, dtm2, method = "cosine", norm = "l2")

> dim(d1\_d2\_cos\_sim)

[1] 300 200

> d1\_d2\_cos\_sim[1:2, 1:5]

2 x 5 sparse Matrix of class "dgCMatrix"

1 2 3 4 5

1 0.02703999 . 0.05063299 0.009500143 0.02753954

2 0.02455143 . 0.06567587 0.034503278 0.04000800

**COSINE SIMILARITY WITHTF-IDF(tern frequency invertd document frequency)**

> dtm = create\_dtm(it, vectorizer)

>

> tfidf = TfIdf$new()

> dtm\_tfidf = fit\_transform(dtm, tfidf)

> d1\_d2\_tfidf\_cos\_sim = sim2(x = dtm\_tfidf, method = "cosine", norm = "l2")

> d1\_d2\_tfidf\_cos\_sim[1:2, 1:5]

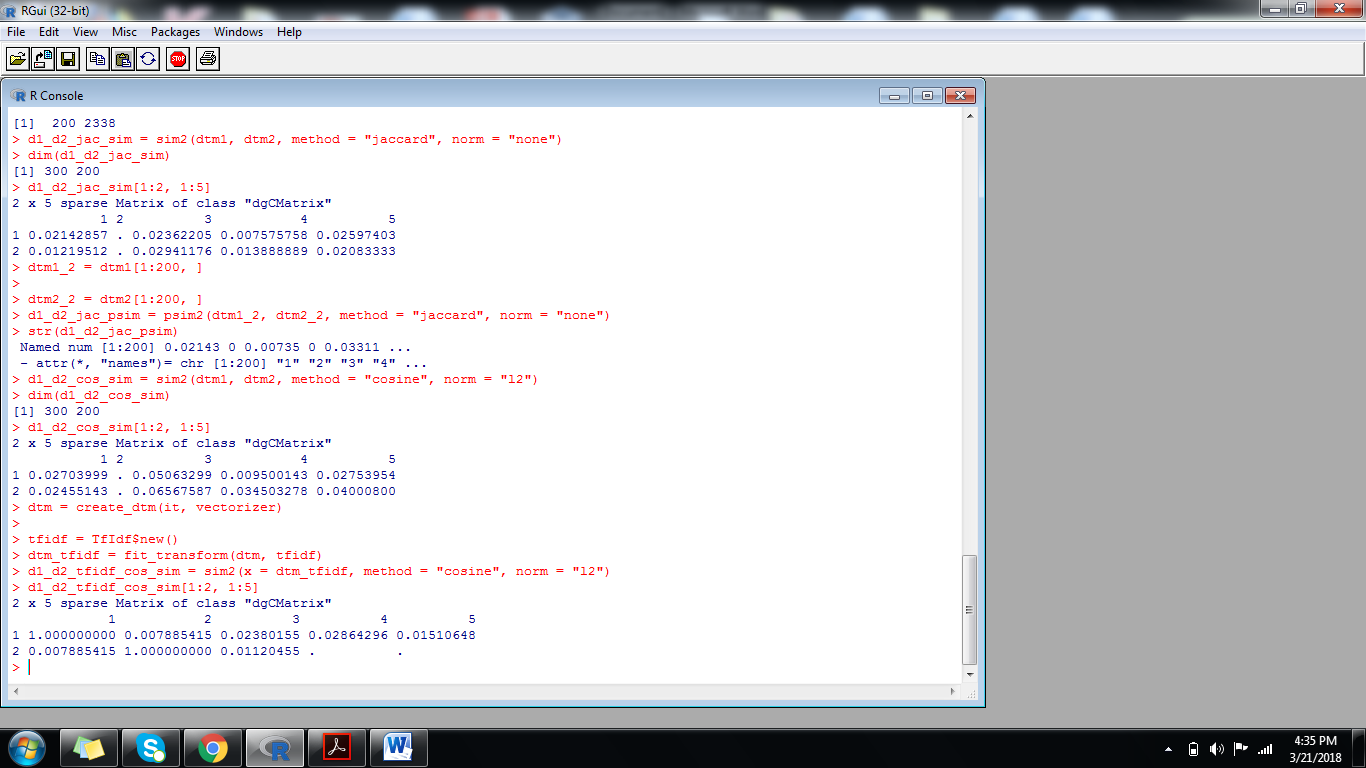
2 x 5 sparse Matrix of class "dgCMatrix"

1 2 3 4 5

1 1.000000000 0.007885415 0.02380155 0.02864296 0.01510648

2 0.007885415 1.000000000 0.01120455 . .

>.



**Cosine similarity with lsa:**

> lsa = LSA$new(n\_topics = 100)

> dtm\_tfidf\_lsa = fit\_transform(dtm\_tfidf, lsa)

> d1\_d2\_tfidf\_cos\_sim = sim2(x = dtm\_tfidf\_lsa, method = "cosine", norm = "l2")

> d1\_d2\_tfidf\_cos\_sim[1:2, 1:5]

1 2 3 4 5

1 1.0000000 0.1131024 0.2441907 0.22305870 0.27371567

2 0.1131024 1.0000000 0.1196754 -0.02698779 -0.04750231

> x = dtm\_tfidf\_lsa[1:250, ]

> y = dtm\_tfidf\_lsa[251:500, ]

> head(psim2(x = x, y = y, method = "cosine", norm = "l2"))

1 2 3 4 5 6

0.11315509 0.12302477 0.19959300 0.07237192 0.18237174 0.02230621

>

