Data-analysis and Retrieval
Assignment 2

Ad Feelders

Universiteit Utrecht

Assignment 2

- Data from Home Depot.
- Learn a function to predict relevance of products to queries.
- Based on match between query and product name or product description.

Query-Product Pairs

This table contains the following attributes:

1. `id`: unique identifier of query-product pair
2. `product_uid`: unique identifier of the product (foreign key)
3. `product_title`: name of the product (text)
4. `search_term`: the query (text)
5. `relevance`: (numeric) relevance of product to query; average score of at least 3 judges. Every judge gives a score of 1 (not relevant), 2 (relevant) or 3 (highly relevant).

Product Descriptions

This table contains the following attributes:

1. `product_uid`: unique product identifier
2. `product_description`: description of the product (text)

Assignment in a Nutshell

1. Construct features that could be predictive for the relevance of a product to a query.

   For example: do all query terms occur in the product title?

2. Compute the feature values on the data.
3. Learn regression / classification models on the training set.
4. Evaluate performance of the models on the test set.

To get you started

```
# read in the query-product table

> query_product.dat <- read.csv("D:/Home Depot/query_product.csv",stringsAsFactors=FALSE)

> str(query_product.dat)
'data.frame':   74067 obs. of  5 variables:
 $ id           : int  2 3 9 16 17 18 20 21 23 27 ...
 $ product_uid  : int  100001 100001 100002 100005 100005 100006 100006 100006 100007 100009 ...
 $ product_title: chr  "Simpson Strong-Tie 12-Gauge Angle" "Simpson Strong-Tie 12-Gauge Angle"
 $ search_term  : chr  "angle bracket" "l bracket" "deck over" "rain shower head" ...
 $ relevance    : num  3 2.5 3 2.33 2.67 3 2.67 3 2.67 3 ...

# descriptive statistics of "relevance"

> summary(query_product.dat$relevance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.330   2.382   3.000   3.000

> summary(as.factor(query_product.dat$relevance))
    1  1.25  1.33   1.5  1.67  1.75     2  2.25  2.33   2.5  2.67  2.75     3
 2105     4  3006     5  6780     9 11730    11 16060    19 15202    11 19125
```

To get you started

```
> library(tau)

> query_product.dat[1,3]
[1] "Simpson Strong-Tie 12-Gauge Angle"

> wc <- textcnt(query_product.dat[1,3],method="string",n=1L)

# show word counts

> wc
  angle   gauge simpson  strong     tie
      1       1       1       1       1

# which words occur in the product title?

> names(wc)
[1] "angle"   "gauge"   "simpson" "strong"  "tie"
```

Computing a feature

```
all.queryterms <- function (queries,docs)
{
n <- length(queries)
feature <- vector(length=n)
for(i in 1:n){
 query <- queries[i]
 document <- docs[i]
 a <- textcnt(query,method="string",n=1L)
 b <- textcnt(document,method="string",n=1L)
 c <- intersect(names(a), names(b))
 feature[i] <- as.numeric(length(a)==length(c))}
feature
}
```

```
> allterms <- all.queryterms(query_product.dat$search_term,
                             query_product.dat$product_title)
> summary(allterms)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2433  0.0000  1.0000
```

Performing the regression

```
> qp.dat <- data.frame(relevance=query_product.dat$relevance,allterms=allterms)

> tr.index <- sample(74067,50000)
> qp.lm <- lm(relevance~allterms,data=qp.dat[tr.index,])
> summary(qp.lm)

Call:
lm(formula = relevance ~ allterms, data = qp.dat[tr.index, ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.61913 -0.30683  0.02317  0.38087  0.69317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.306835   0.002656  868.61   <2e-16 ***
allterms    0.312298   0.005389   57.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5167 on 49998 degrees of freedom
Multiple R-squared:  0.06294,   Adjusted R-squared:  0.06292
F-statistic:  3358 on 1 and 49998 DF,  p-value: < 2.2e-16
```

The Report

- The report must contain:
  1. Introduction and problem description
  2. Description of data (including descriptive statistics)
  3. Description of data pre-processing performed
  4. Description of constructed features (8 in total)
  5. Results and discussion of regression and classification models
  6. Conclusion
- The report should be approximately 10-15 pages long (indication, not hard constraint).
- You may use any tools you want (support for R and Python).
- Hand in ultimately June 24, 2022 in pdf-format by e-mail.