Ready Biodegradability

2019-07-24

```
library(reticulate)
```

```
## Warning: package 'reticulate' was built under R version 3.6.1
```

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages -------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts ----------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
getwd()
```

```
## [1] "C:/Users/us16120/Projects/Cheminformatics/readybiodegradability/models"
```

read data

```
import pandas as pd
import numpy as np
from rdkit import Chem

df = pd.read_csv('C:/Users/us16120/Projects/Cheminformatics/readybiodegradability/data/processed/alles.
df = df.drop(['Unnamed: 0'], axis = 1)

for i, row in df.iterrows():
     df.loc[i, 'InChI'] = Chem.MolToInchiKey(Chem.MolFromSmiles(row['SMILES']))

df.sample(5).head()
```

```
##                                      SMILES EndPt                        InChI
## 2647                      Oc1cc(Cl)cc(Cl)c1   NRB  VPOMSPZBQMDLTM-UHFFFAOYSA-N
## 1041    C(OP(OCc1ccccc1)OCc2ccccc2)c3ccccc3    RB  KKFOMYPMTJLQGA-UHFFFAOYSA-N
## 1841  O=P(OC(CCl)CCl)(OC(CCl)CCl)OC(CCl)CCl   NRB  ASLWPAWFJZFCKF-UHFFFAOYSA-N
## 1733                            Cc1ccccc1NC   NRB  GUAWMXYQZKVRCW-UHFFFAOYSA-N
## 1694               OC(=O)c1cccc([n]1)C(O)=O    RB  WJJMNDUMQPNECX-UHFFFAOYSA-N
```

duplicate records

```
df <- py$df
head(df)
```

```
##                         SMILES EndPt                       InChI
## 1                C[S](O)(=O)=O    RB AFVFQIVMOAPDHO-UHFFFAOYSA-N
## 2   OC(=O)c1ccccc1[N+]([O-])=O    RB SLAMLWHELXOEJZ-UHFFFAOYSA-N
## 3        CC(C)=CCC\\C(C)=C/C=C    RB WTEVQBCEXWBHNA-YFHOEESVSA-N
## 4 CCCCCCCC\\C=C/CCCCCCCC(=O)OC    RB QYDYPVFESGNLHU-KHPPLWFESA-N
## 5                COC(=O)C(C)=C    RB VVQNEPGJFQJSBK-UHFFFAOYSA-N
## 6      CCOC(=O)\\C=C\\C(=O)OCC    RB IEPRKVQEAMIZSS-AATRIKPKSA-N
```

```
dim(df)
```

```
## [1] 2990    3
```

```
length(unique(df$InChI))
```

```
## [1] 2092
```

```
df$ReadyBiodeg <- ifelse(df$EndPt == 'RB', 1, 0)
# grps: molecules with discrepancies in reported bidegradablity
grps <- df %>%
  group_by(InChI) %>%
  summarise(count = n(), qaz = sum(ReadyBiodeg), remainder = qaz %% count) %>%
  filter(remainder > 0)
# remove grps from df
df <- anti_join(df, grps)
```

```
## Joining, by = "InChI"
```

```r
# keep unique molecules
df <- df[!duplicated(df$InChI), ]
df <- df[ , c('SMILES', 'InChI', 'EndPt', 'ReadyBiodeg')]
dim(df)
```

```
## [1] 2064    4
```

```r
head(df)
```

```
##                          SMILES                       InChI EndPt
## 1                C[S](O)(=O)=O AFVFQIVMOAPDHO-UHFFFAOYSA-N    RB
## 2   OC(=O)c1ccccc1[N+]([O-])=O SLAMLWHELXOEJZ-UHFFFAOYSA-N    RB
## 3        CC(C)=CCC\\C(C)=C/C=O WTEVQBCEXWBHNA-YFHOEESVSA-N    RB
## 4 CCCCCCCC\\C=C/CCCCCCCC(=O)OC QYDYPVFESGNLHU-KHPPLWFESA-N    RB
## 5                COC(=O)C(C)=C VVQNEPGJFQJSBK-UHFFFAOYSA-N    RB
## 6      CCOC(=O)\\C=C\\C(=O)OCC IEPRKVQEAMIZSS-AATRIKPKSA-N    RB
##   ReadyBiodeg
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

```r
write.csv(df, 'C:/Users/us16120/Projects/Cheminformatics/readybiodegradability/data/processed/alles02.c
```