Automated Readability Index Assignment

October 29, 2024

[19]:

```
import polars as pl
import pandas as pd
```

[26]:

```
pl.Config.set_thousands_separator(",")
pl.Config.set_tbl_hide_column_data_types(True)
pl.Config.set_float_precision(3)
pl.Config.set_tbl_rows(20)
```

[26]:

```
polars.config.Config
```

[27]:

```
!wget -nc https://ling583.s3.amazonaws.com/books.parquet
```

```
File ‘books.parquet’ already there; not retrieving.
```

[28]:

```
c = pl.read_parquet('books.parquet')
```

[29]:

```
c.head()
```

[29]:

```
shape: (5, 6)

  tok      norm     tag   pos     fileid        sentid

  [        [        XX    X       austen-emma   austen-emma_0000
  Emma     emma     NNP   PROPN   austen-emma   austen-emma_0000
  by       by       IN    ADP     austen-emma   austen-emma_0000
  Jane     jane     NNP   PROPN   austen-emma   austen-emma_0000
  Austen   austen   NNP   PROPN   austen-emma   austen-emma_0000
```

[30]:

```
c.with_columns(chars = pl.col('tok').str.len_chars())
```

[30]:

```
shape: (1_439_763, 7)


  tok        norm       tag      pos      fileid                sentid
  chars
```

```
 [         [          XX      X       austen-emma          austen-emma_0000
 1
 Emma      emma       NNP     PROPN   austen-emma          austen-emma_0000
 4
 by        by         IN      ADP     austen-emma          austen-emma_0000
 2
 Jane      jane       NNP     PROPN   austen-emma          austen-emma_0000
 4
 Austen    austen     NNP     PROPN   austen-emma          austen-emma_0000
 6
 1816      1816       CD      NUM     austen-emma          austen-emma_0000
 4
 ]         ]          -RRB-   PUNCT   austen-emma          austen-emma_0000
 1
 VOLUME    volume     NN      NOUN    austen-emma          austen-emma_0000
 6
 I         i          PRP     PRON    austen-emma          austen-emma_0000
 1
 CHAPTER   chapter    NN      NOUN    austen-emma          austen-emma_0000
 7
 …         …          …       …       …                    …
 …
 Exeunt    exeunt     NNP     PROPN   shakespeare-macbeth  shakespeare-
macbeth_1442   6
 Omnes     omnes      NNP     PROPN   shakespeare-macbeth  shakespeare-
macbeth_1442   5
 .         .          .       PUNCT   shakespeare-macbeth  shakespeare-
macbeth_1442   1
 FINIS     finis      NNP     PROPN   shakespeare-macbeth  shakespeare-
macbeth_1443   5
 .         .          .       PUNCT   shakespeare-macbeth  shakespeare-
macbeth_1443   1
 THE       the        DT      DET     shakespeare-macbeth  shakespeare-
macbeth_1444   3
 TRAGEDIE  tragedie   NNS     NOUN    shakespeare-macbeth  shakespeare-
macbeth_1444   8
 OF        of         IN      ADP     shakespeare-macbeth  shakespeare-
macbeth_1444   2
 MACBETH   macbeth    NNP     PROPN   shakespeare-macbeth  shakespeare-
macbeth_1444   7
 .         .          .       PUNCT   shakespeare-macbeth  shakespeare-
macbeth_1444   1
```

[32]:

```
c.group_by('fileid').agg(
    n_words = pl.col('tok').count(),
    n_chars = pl.col('tok').str.len_chars().sum(),
    n_sents = pl.col('sentid').n_unique()
).with_columns(
    ari = 4.71 * pl.col('n_chars')/pl.col('n_words') + 0.5 * pl.col('n_words')/
 ↪pl.col('n_sents') - 21.43
).sort(by='ari')
```

[32]:

```
shape: (15, 5)

  fileid                 n_words   n_chars   n_sents   ari

  shakespeare-caesar     25,155    88,972    1,591     3.134
  shakespeare-hamlet     36,308    129,405   2,320     3.182
  shakespeare-macbeth    22,200    80,188    1,445     3.265
  burgess-busterbear     18,739    66,688    1,002     4.683
  carroll-alice          34,503    116,009   1,624     5.029
  bryant-stories         56,047    194,498   2,719     5.221
  edgeworth-parents      209,856   739,177   10,026    5.626
  chesterton-thursday    69,927    261,229   3,568     5.965
  chesterton-ball        97,936    370,970   4,619     7.012
  chesterton-brown       86,456    326,248   3,703     8.017
  austen-emma            189,305   722,615   7,376     9.382
  melville-moby_dick     258,411   999,044   9,781     9.989
  austen-persuasion      98,407    380,023   3,653     10.228
  austen-sense           140,313   549,509   4,790     11.662
  milton-paradise        96,200    376,389   1,834     23.225
```

[ ]:

```
```