products-csv-gen

April 24, 2019

1 Products-to-CSV-Generator

In this Jupyter-Notebook the reviews were transformed into product data. The summarized star and review data is not used actually in the final service, but it *can* be used, to make it more performant when there is a bigger amount of data.

Content: - Importing necessary packages - Read the data - Create product dataset - Calculate star ratings, review counts and average ratings - Add product names (from bonprix.de) - Add product pictures as django paths (from bonprix.de) - Saving to CSV

1.1 Importing necessary packages

```
In [1]: import pandas as pd
        import json
        import nltk
        import random
        #import numpy as np
        #nltk.download("punkt")
        import string
        exclude = set(string.punctuation)
        import time
        from tqdm import tqdm
        import re
```

1.2 Read the data

```
In [121]: data = pd.read_csv('Datensatz_Coding_Challenge.csv', delimiter=";")
          corpus = data.copy()
          corpus.head()
```

Out[121]:

	StyleID	text	rating
0	1709054	Die sind okay und dann für den Preis.	5
1	1709054	Qualität und Preis sind gut. Leider sind sie z...	3
2	8623725	lässt schlanker aussehen	5
3	8623725	Material und Farbe gut. Da einige Kundinnen in...	3
4	9743730	Material ist schön zum verdunkel. Leider doch...	5

1.3 Create product dataset

```
In [130]: product = pd.DataFrame(index=corpus["StyleID"].unique(), columns=['review_count'])

In [131]: agg = corpus[["StyleID", "rating"]].groupby(["rating", "StyleID"]).size()

In [132]: agg = agg.to_frame()
```

1.3.1 Calculate star ratings, review counts and average ratings

```
In [134]: star1_count = agg.xs(1, level='rating', axis=0, drop_level=True)
          star2_count = agg.xs(2, level='rating', axis=0, drop_level=True)
          star3_count = agg.xs(3, level='rating', axis=0, drop_level=True)
          star4_count = agg.xs(4, level='rating', axis=0, drop_level=True)
          star5_count = agg.xs(5, level='rating', axis=0, drop_level=True)

In [135]: star1_count = star1_count[0]
          star2_count = star2_count[0]
          star3_count = star3_count[0]
          star4_count = star4_count[0]
          star5_count = star5_count[0]

In [138]: product.insert(1, "star1_count", star1_count, True)
          product.insert(1, "star2_count", star2_count, True)
          product.insert(1, "star3_count", star3_count, True)
          product.insert(1, "star4_count", star4_count, True)
          product.insert(1, "star5_count", star5_count, True)

In [141]: col_list = ["star5_count", "star4_count", "star3_count", "star2_count", "star1_count"
          product["review_count"] = product[col_list].sum(axis=1)

In [142]: product
```

Out[142]:

	review_count	star5_count	star4_count	star3_count	star2_count	\
1709054	2039	1043	648	168	71	
8623725	1911	850	559	241	113	
9743730	1998	1364	427	93	31	
655046	1935	1013	573	129	78	
553018	1828	859	556	197	90	
434886	2025	895	786	165	63	
654563	1797	849	530	166	93	
709229	2032	1123	548	175	95	
515928	1620	1169	368	38	20	
44970574	1965	714	682	268	109	

	star1_count
1709054	109
8623725	148
9743730	83
655046	142

```
553018            126
434886            116
654563            159
709229             91
515928             25
44970574          192
```

1.3.2 Add product names (from bonprix.de)

```
In [143]: names = {44970574: "Unknown",
           9743730: "Verdunkelungsvorhang 'Uni' (1er-Pack)",
           8623725: "Shape Badeanzug",
           1709054: "Basic Baumwollshirt Shirt Single-Jersey",
           709229: "Stretch-Kleid in Hängerchen-Optik",
           655046: "Sport-BH Level 1",
           654563: "BH (3er-Pack) Bio-Baumwolle",
           553018: "Ballerina",
           515928: "Slip",
           434886: "Jeans Classic Fit Straight"}

In [144]: name = pd.DataFrame.from_dict(names, orient='index')
           name

Out[144]:                                              0
           44970574                            Unknown
           9743730     Verdunkelungsvorhang 'Uni' (1er-Pack)
           8623725                          Shape Badeanzug
           1709054   Basic Baumwollshirt Shirt Single-Jersey
           709229          Stretch-Kleid in Hängerchen-Optik
           655046                        Sport-BH Level 1
           654563               BH (3er-Pack) Bio-Baumwolle
           553018                             Ballerina
           515928                                  Slip
           434886                Jeans Classic Fit Straight

In [145]: product.insert(1, "name", name, True)

In [146]: product

Out[146]:           review_count                                    name  star5_count  \
           1709054          2039  Basic Baumwollshirt Shirt Single-Jersey         1043
           8623725          1911                          Shape Badeanzug          850
           9743730          1998    Verdunkelungsvorhang 'Uni' (1er-Pack)         1364
           655046           1935                          Sport-BH Level 1         1013
           553018           1828                                 Ballerina          859
           434886           2025              Jeans Classic Fit Straight          895
           654563           1797             BH (3er-Pack) Bio-Baumwolle          849
           709229           2032       Stretch-Kleid in Hängerchen-Optik         1123
           515928           1620                                    Slip         1169
```

```
44970574          1965                                  Unknown          714

          star4_count  star3_count  star2_count  star1_count
1709054           648          168           71          109
8623725           559          241          113          148
9743730           427           93           31           83
655046            573          129           78          142
553018            556          197           90          126
434886            786          165           63          116
654563            530          166           93          159
709229            548          175           95           91
515928            368           38           20           25
44970574          682          268          109          192
```

1.3.3 Add product pictures as django paths (from bonprix.de)

```
In [149]: pictures = {434886: "static/images/434886.jpg",
          515928: "static/images/515928.jpg",
          553018: "static/images/553018.jpg",
          654563: "static/images/654563.jpg",
          655046: "static/images/655046.jpg",
          709229: "static/images/709229.jpg",
          1709054: "static/images/1709054.jpg",
          8623725: "static/images/8623725.jpg",
          9743730: "reviews/static/images/9743730_GOuMyL6.jpg",
          44970574: "reviews/static/images/image_1_8L1UOQX.jpg"}

In [150]: picture = pd.DataFrame.from_dict(pictures, orient='index')
          picture

Out[150]:                                                  0
          434886                     static/images/434886.jpg
          515928                     static/images/515928.jpg
          553018                     static/images/553018.jpg
          654563                     static/images/654563.jpg
          655046                     static/images/655046.jpg
          709229                     static/images/709229.jpg
          1709054                   static/images/1709054.jpg
          8623725                   static/images/8623725.jpg
          9743730   reviews/static/images/9743730_GOuMyL6.jpg
          44970574  reviews/static/images/image_1_8L1UOQX.jpg

In [151]: product.insert(1, "picture", picture, True)

In [159]: product

Out[159]:                                          name  \
          1709054   Basic Baumwollshirt Shirt Single-Jersey
          8623725                            Shape Badeanzug
```

```
9743730     Verdunkelungsvorhang 'Uni' (1er-Pack)
655046                            Sport-BH Level 1
553018                                  Ballerina
434886                  Jeans Classic Fit Straight
654563                 BH (3er-Pack) Bio-Baumwolle
709229           Stretch-Kleid in Hängerchen-Optik
515928                                        Slip
44970574                                   Unknown

                                            picture  review_count  \
1709054                  static/images/1709054.jpg          2039
8623725                  static/images/8623725.jpg          1911
9743730   reviews/static/images/9743730_GOuMyL6.jpg          1998
655046                    static/images/655046.jpg          1935
553018                    static/images/553018.jpg          1828
434886                    static/images/434886.jpg          2025
654563                    static/images/654563.jpg          1797
709229                    static/images/709229.jpg          2032
515928                    static/images/515928.jpg          1620
44970574  reviews/static/images/image_1_8L1U0QX.jpg          1965

          star1_count  star2_count  star3_count  star4_count  star5_count
1709054           109           71          168          648         1043
8623725           148          113          241          559          850
9743730            83           31           93          427         1364
655046            142           78          129          573         1013
553018            126           90          197          556          859
434886            116           63          165          786          895
654563            159           93          166          530          849
709229             91           95          175          548         1123
515928             25           20           38          368         1169
44970574          192          109          268          682          714
```

```
In [158]: product = product[["name", "picture", "review_count", "star1_count", "star2_count",
```

1.4 Saving to CSV

```
In [163]: product.to_csv("products.csv", sep=";")
```