cb-01-szymanskir-book_descriptions_analysis

January 22, 2019

1 Important

`make models/content-based-models/tf-idf-nouns-model.pkl` has to be run in the main directory before as the notebook uses the result data from that process.

2 Imports

```
In [1]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline

        from langdetect import detect, detect_langs
```

3 Visualization settings

```
In [2]: sns.set(context='paper', font_scale=1.2, style='ticks', palette='muted',
                rc={"axes.labelsize":16, "ytick.labelsize": 14, "xtick.labelsize":14,
                    "font.family": "sans-serif"})
```

4 Analysis

```
In [3]: original_data = pd.read_csv('../data/processed/book.csv', index_col='book_id')
```

There are 315 books with missing descriptions:

```
In [4]: original_data['description'].isna().sum()
```

```
Out[4]: 315
```

The are 199/200 descriptions that are written in different languages than english.

```
In [5]: non_english_desc = original_data['description'].dropna().apply(
            lambda desc: detect(desc) != 'en')
        non_english_desc.sum()
```

```
Out[5]: 200
```

Missing description example

```
In [6]: original_data.loc[9973]
```

```
Out[6]: goodreads_book_id                                                   849380
        best_book_id                                                        849380
        work_id                                                               4370
        books_count                                                             52
        isbn                                                             609805797
        authors                                        John M. Gottman, Nan Silver
        original_publication_year                                             1999
        original_title               The Seven Principles for Making Marriage Work:...
        title                        The Seven Principles for Making Marriage Work:...
        language_code                                                          NaN
        average_rating                                                        4.19
        ratings_count                                                         8868
        work_ratings_count                                                   10017
        work_text_reviews_count                                                749
        ratings_1                                                              126
        ratings_2                                                              334
        ratings_3                                                             1604
        ratings_4                                                             3446
        ratings_5                                                             4507
        image_url                    https://images.gr-assets.com/books/1320521960m...
        small_image_url              https://images.gr-assets.com/books/1320521960s...
        isbn13                                                         9.78061e+12
        description                  John Gottman has revolutionized the study of m...
        Name: 9973, dtype: object
```

Description not in english example

```
In [7]: original_data.loc[9966]
```

```
Out[7]: goodreads_book_id                                                     9864
        best_book_id                                                          9864
        work_id                                                            3279710
        books_count                                                             72
        isbn                                                             312254997
        authors                                                     Salman Rushdie
        original_publication_year                                             1999
        original_title                                 The Ground Beneath Her Feet
        title                                          The Ground Beneath Her Feet
        language_code                                                          eng
        average_rating                                                        3.77
        ratings_count                                                         8673
        work_ratings_count                                                    9541
        work_text_reviews_count                                                535
        ratings_1                                                              264
        ratings_2                                                              803
        ratings_3                                                             2450
        ratings_4                                                             3360
        ratings_5                                                             2664
        image_url                    https://s.gr-assets.com/assets/nophoto/book/11...
```

```
small_image_url            https://s.gr-assets.com/assets/nophoto/book/50...
isbn13                                                           9.78031e+12
description                The ground shifts repeatedly beneath the reade...
Name: 9966, dtype: object
```

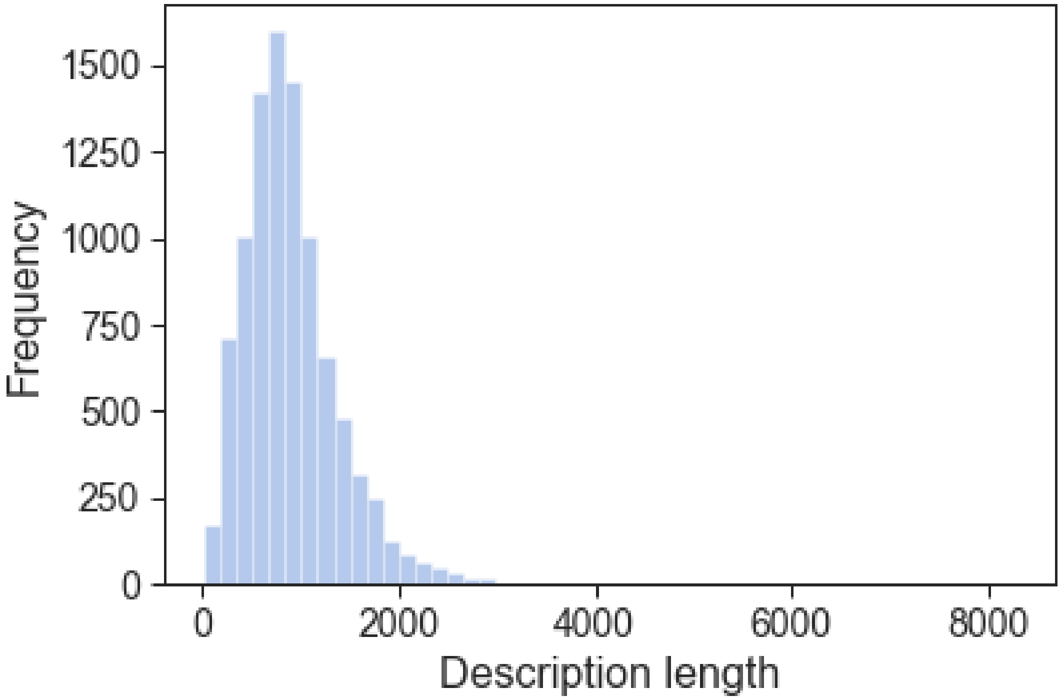
```
In [8]: reduced_data_descriptions = original_data['description'].dropna()[~non_english_desc]
```

4.0.1 Description length analysis

```
In [9]: description_lengths = reduced_data_descriptions.str.len()
        ax = sns.distplot(description_lengths, kde=False)
        ax.set(xlabel='Description length', ylabel='Frequency')
        # ax.get_figure().savefig('description-length-distribution.pdf', bbox_inches='tight')
```

```
/home/szymanskir/Documents/Inzynierka/Recommendation-system/rs-
venv/lib/python3.7/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a
non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]`
instead of `arr[seq]`. In the future this will be interpreted as an array index,
`arr[np.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

```
Out[9]: [Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Description length')]
```



```
In [10]: reduced_data_descriptions.str.len().describe()
```

```
Out[10]: count    9485.000000
         mean      903.555930
         std       495.163786
         min        18.000000
         25%       575.000000
         50%       828.000000
         75%      1125.000000
         max      8271.000000
         Name: description, dtype: float64
```

4.1 Noticed issues

- There are missing descriptions in the data
- Some descriptions are not in english

4.2 Description content analysis

```
In [11]: original_data['description'].dropna().head()
```

```
Out[11]: book_id
         1     Winning will make you famous. Losing means ce...
         2    Harry Potter's life is miserable. His parents ...
         3     About three things I was absolutely positive...
         4    The unforgettable novel of a childhood in a sl...
         5     THE GREAT GATSBY , F. Scott Fitzgeralds thir...
         Name: description, dtype: object
```

The descriptions need cleaning regarding removing punctuation and stopwords. Additionally stemming and lemmatization will be performed.

5 Cleaning results

Descriptions have been cleaned using the following operations: - transforming to lower case - lemmatization - stemming

Two approaches regarding nouns have been implemented: - nouns are kept in the description - nouns are deleted from the description

The reason why there are two approaches is the fact that on the one hand expressions like `Harry Potter` is a very important feature. But if there is another book in which the main character is named `Harry` then even though this book might be completely different it might get classified as similar.

```
In [12]: cleaned_data_with_nouns = pd.read_csv('../data/interim/cb-tf-idf/book_with_nouns.csv',
         index_col='book_id')
         cleaned_data_with_nouns['description'].head()
```

```
Out[12]: book_id
         1    win make lose mean certain the nation form nor...
         2    harri life his parent dead stuck heartless for...
         3    about three thing i absolut edward part i know...
         4    the unforgett novel childhood sleepi southern ...
         5    the great gatsbi scott third stand suprem achi...
         Name: description, dtype: object
```

6 Example results

```
In [13]: print(cleaned_data_with_nouns.loc[1, 'description'])
```

```
win make lose mean certain the nation form north countri consist wealthi capitol
region surround poorer earli rebellion led district capitol result destruct creation
annual televis event known hunger in remind power grace district must yield one boy
one girl age lotteri system particip the chosen annual reap forc fight leav one
survivor claim when young select district femal katniss volunt take she male
counterpart pit stronger train whole see death but katniss close death for surviv
second
```

```
In [14]: print(original_data.loc[1, 'description'])
```

```
 Winning will make you famous. Losing means certain death. The nation of Panem, formed
from a post-apocalyptic North America, is a country that consists of a wealthy Capitol
region surrounded by 12 poorer districts. Early in its history, a rebellion led by a
13th district against the Capitol resulted in its destruction and the creation of an
annual televised event known as the Hunger Games. In punishment, and as a reminder of
the power and grace of the Capitol, each district must yield one boy and one girl
between the ages of 12 and 18 through a lottery system to participate in the games.
The 'tributes' are chosen during the annual Reaping and are forced to fight to the
death, leaving only one survivor to claim victory. When 16-year-old Katniss's young
sister, Prim, is selected as District 12's female representative, Katniss volunteers
to take her place. She and her male counterpart Peeta, are pitted against bigger,
stronger representatives, some of whom have trained for this their whole lives. , she
sees it as a death sentence. But Katniss has been close to death before. For her,
survival is second nature.
```

6.1 Comparison of nouns removal

```
In [15]: clean_data_with_nouns = pd.read_csv('../data/interim/cb-tf-idf/book_with_nouns.csv',
         index_col='book_id')
         clean_data_without_nouns = pd.read_csv('../data/interim/cb-tf-
         idf/book_without_nouns.csv', index_col='book_id')
```

```
In [16]: harry_potter_description_with_nouns = clean_data_with_nouns.loc[2, 'description']
         harry_potter_description_without_nouns = clean_data_without_nouns.loc[2, 'description']
```

```
In [17]: print(harry_potter_description_with_nouns)
```

```
harri life his parent dead stuck heartless forc live tini closet but fortun chang
receiv letter tell truth a mysteri visitor rescu relat take new hogwart school
witchcraft after lifetim bottl magic harri final feel like normal but even within
wizard he boy person ever surviv kill cur inflict evil lord launch brutal takeov
wizard vanish fail kill though first year hogwart best everyth there danger secret
object hidden within castl harri believ respons prevent fall evil but bring contact
forc terrifi ever could full sympathet wild imagin countless excit first instal seri
assembl unforgett magic world set stage mani adventur
```

```
In [18]: print(harry_potter_description_without_nouns)
```

```
life his parent dead stuck heartless forc live tini closet but fortun chang receiv
letter tell truth mysteri visitor rescu relat take new after lifetim bottl magic final
feel like normal but even within he boy person ever surviv kill cur inflict evil
launch brutal takeov vanish fail kill first year best everyth there danger secret
object hidden within castl believ respons prevent fall evil but bring contact forc
terrifi ever could sympathet wild imagin countless excit first instal seri assembl
unforgett magic world set stage mani adventur
```

6.1.1 Example of books with short descriptions

Unfortunately when some descriptions are very short the cleaning results in an empty description.

```
In [19]: original_data.loc[4210, 'description']
```

```
Out[19]: 'Kiss of the Highlander (The Highlander Series, Book 4)'
```

```
In [20]: clean_data_with_nouns.loc[4210, 'description']
```

```
Out[20]: 'kiss highland highland book'
```

```
In [21]: clean_data_without_nouns.loc[4210, 'description']
```

```
Out[21]: nan
```

However this occurs only 2 times in case of the proper noun removal approach.

```
In [22]: clean_data_without_nouns['description'].isna().sum()
```

```
Out[22]: 2
```

```
In [23]: clean_data_with_nouns['description'].isna().sum()
```

```
Out[23]: 0
```

6.2 Descriptions length after cleaning

```
In [24]: desc_len_with_nouns = clean_data_with_nouns['description'].str.len()
         desc_len_without_nouns = clean_data_without_nouns['description'].str.len()
```

```
In [25]: desc_len_with_nouns.describe()
```

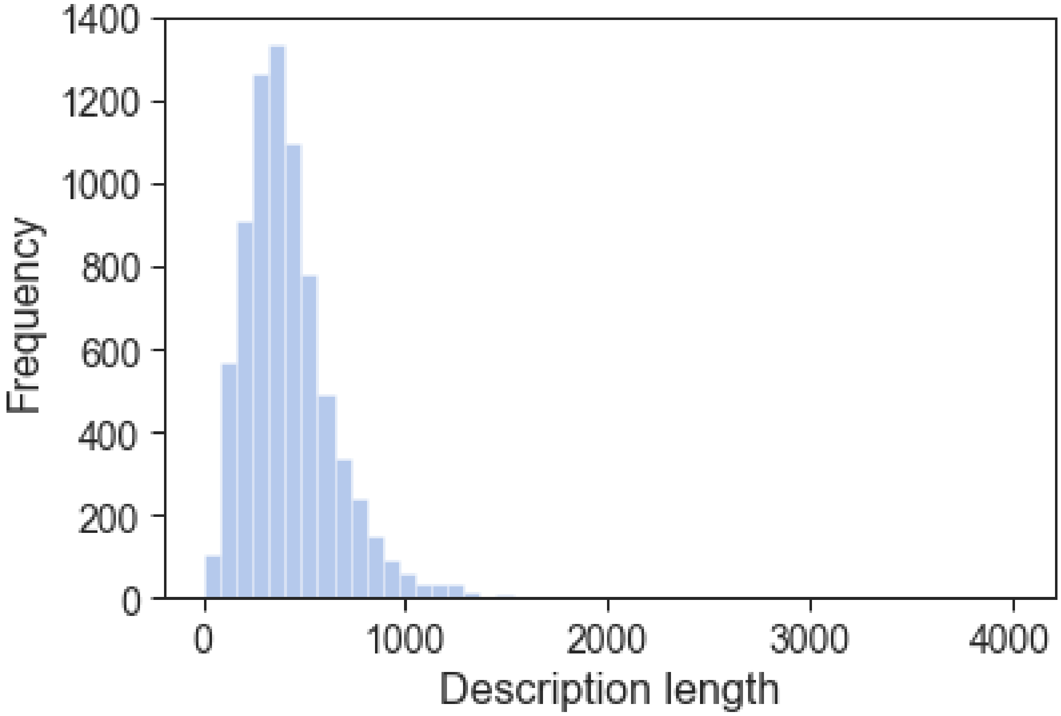
```
Out[25]: count    7575.000000
         mean      419.858218
         std       230.640717
         min         4.000000
         25%       266.000000
         50%       382.000000
         75%       522.000000
         max      4020.000000
         Name: description, dtype: float64
```

```
In [26]: desc_len_without_nouns.describe()
```

```
Out[26]: count    7573.000000
         mean      348.012809
         std       193.747781
         min         5.000000
         25%       218.000000
         50%       316.000000
         75%       434.000000
         max      3387.000000
         Name: description, dtype: float64
```

```
In [27]: ax = sns.distplot(desc_len_with_nouns.tolist(), kde=False)
         ax.set(xlabel='Description length', ylabel='Frequency')
```

```
Out[27]: [Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Description length')]
```



7 Notes

- N-grams should be considered in other methods, for example a very specific feature word pairing like `Hunger Games` is omitted in the result
- weird ending like for example `countri` instead of `country`. however this is not an issue because all words will be processed in the same way