cb-03-szymanskir-cb_validation_methods_analysis

January 19, 2019

1 Important

`make scores` has to be run before any cell in this notebook

2 Imports

```
In [1]: import pandas as pd
```

3 Load data

```
In [2]: results = pd.read_csv('../results/cb-results.csv')
```

4 Analysis

The goal of this notebook is to describe and evaluate recommendation models validation methods that were used during the project.

4.1 Precision and Recall measures

The first method was to using the classic precision and recall measures that are often used in classification problems. The idea consisted of comparing the set of books recommended by models and the set of books that were found in the `similar_books` tag in the xml data files. This method was used only for the purpose of comparing different models that use different representations of the items of interest and deciding which one is the best performing. The following results were obtained:

```
In [3]: results[['model', 'precision', 'recall']].sort_values(
            ['precision', 'recall'], ascending=False)
```

```
Out[3]:                                          model  precision  recall
        1                        tag-predictions.csv      0.033   0.241
        2   tf-idf-no-nouns-2grams-tags-predictions.csv      0.018   0.124
        7      tf-idf-nouns-2grams-tags-predictions.csv      0.018   0.124
       16             tf-idf-nouns-tags-predictions.csv      0.018   0.124
       13      tf-idf-nouns-3grams-tags-predictions.csv      0.017   0.121
       21   tf-idf-no-nouns-3grams-tags-predictions.csv      0.017   0.120
       23          tf-idf-no-nouns-tags-predictions.csv      0.017   0.120
```

18	tf-idf-nouns-2grams-predictions.csv	0.007	0.048
0	tf-idf-nouns-predictions.csv	0.007	0.047
8	tf-idf-nouns-3grams-predictions.csv	0.007	0.047
4	tf-idf-no-nouns-predictions.csv	0.006	0.037
12	tf-idf-no-nouns-3grams-predictions.csv	0.005	0.034
19	tf-idf-no-nouns-2grams-predictions.csv	0.005	0.034
20	count-nouns-2grams-tags-predictions.csv	0.005	0.031
22	count-nouns-tags-predictions.csv	0.005	0.031
5	count-nouns-3grams-tags-predictions.csv	0.004	0.030
9	count-nouns-2grams-predictions.csv	0.004	0.030
6	count-nouns-predictions.csv	0.004	0.029
14	count-nouns-3grams-predictions.csv	0.004	0.029
11	count-no-nouns-tags-predictions.csv	0.003	0.020
15	count-no-nouns-2grams-tags-predictions.csv	0.003	0.020
24	count-no-nouns-3grams-tags-predictions.csv	0.003	0.020
10	count-no-nouns-predictions.csv	0.003	0.019
17	count-no-nouns-2grams-predictions.csv	0.003	0.019
3	count-no-nouns-3grams-predictions.csv	0.003	0.018

The tf-idf models performed better than count models, tags features also usually resulted in better accuracy. However, for all models scores are rather low, but does it mean that all models are not working properly?

The recommendation problem is much more complex than the classification problem. The main difference is the subjective side of recommendations e.g. one book may be a good recommendation for one person but a bad one for another person. In case of classifications labels can usually be described objectively e.g. the picture presents the 1, 2, 3 digits.

Another issue is that if a book recommended by the implemented model was not in the test set, it does not necessarily mean that it is not a good recommendation.

While collecting ground truth data on similar books several phenomenons can occur. Let's consider the following example: the goal is to collect data about books that are similar to A, there are two books B and C that are similar to A. However B is more popular and C is more similar to A. The problem is that B will appear more frequently in the test data just because more people have read this book and will consequently be considered as the more similar one even though C is the more similar one.

The whole definition of `similar books` is very ambiguous. One might consider books to be similar because the main characters behave in a similar way, but the stories have a significantly different setting. The style of writing is also a factor that also determines whether books are similar.

Due to the reasons described above the precision and recall metrics do not represent the overall performance of models, but it can be a way of comparing models.

4.2 Modified precision and recall measures

As described in the previous section the precision and recall measures penalize the model when it's recommendations are not the present in the test set. The idea was to remove that property and only consider the positive feedback.

As all recommendation models recommend the same amount of books, the idea was to compare the amount of recommendations that were also present in the test set. The goal was to find a

measurable difference between the models in order to define some characteristics.

```
In [4]: results[['model', 'correct_hits']].sort_values(
            'correct_hits', ascending=False)
```

```
Out[4]:                                            model   correct_hits
        1                              tag-predictions.csv           4503
        16                tf-idf-nouns-tags-predictions.csv           1878
        7          tf-idf-nouns-2grams-tags-predictions.csv           1877
        2       tf-idf-no-nouns-2grams-tags-predictions.csv           1845
        13         tf-idf-nouns-3grams-tags-predictions.csv           1828
        21      tf-idf-no-nouns-3grams-tags-predictions.csv           1807
        23             tf-idf-no-nouns-tags-predictions.csv           1801
        0                     tf-idf-nouns-predictions.csv            693
        8               tf-idf-nouns-3grams-predictions.csv            688
        18              tf-idf-nouns-2grams-predictions.csv            688
        4                  tf-idf-no-nouns-predictions.csv            577
        19           tf-idf-no-nouns-2grams-predictions.csv            529
        12           tf-idf-no-nouns-3grams-predictions.csv            521
        20          count-nouns-2grams-tags-predictions.csv            472
        22                count-nouns-tags-predictions.csv            471
        5           count-nouns-3grams-tags-predictions.csv            461
        9                count-nouns-2grams-predictions.csv            460
        6                      count-nouns-predictions.csv            450
        14               count-nouns-3grams-predictions.csv            444
        11             count-no-nouns-tags-predictions.csv            330
        15       count-no-nouns-2grams-tags-predictions.csv            319
        10                count-no-nouns-predictions.csv            309
        24       count-no-nouns-3grams-tags-predictions.csv            306
        17          count-no-nouns-2grams-predictions.csv            297
        3           count-no-nouns-3grams-predictions.csv            287
```

5   Conclusions

The tags features have majorely enhanced the precision and recall scores of the models. However due to the complexity of the recommendation problem those measures are not sufficient to determine the overall performance of models. Further online evaluation methods are needed in order to confirm if the determined business goals are being achieved and subsequent models should be designed in order to maximize the online quality measure.