Practical session week 6

Data Science course

This week we continue working on assignment 2.

End goal for assignment 2

Deadline: April 17.

Task: Predict the relevance of search results on homedepot.com

- Information on the task: https://www.kaggle.com/c/home-depot-product-search-relevance/
- Information on the data: https://www.kaggle.com/c/home-depot-product-search-relevance/data

End product: a report containing:

1. (week 5) The task definition
2. (week 5) Data description: results of the data exploration
3. (week 5) Description of the baseline method
4. (week 6, 7) Description of the features you defined and the hyperparameter optimization
5. (week 6, 7) Results: a table with (a) Baseline results (replication of existing method); (b) Results for multiple regression models; (c) Results of the hyperparameter optimization, and (c) the results for different feature representations you experimented with
6. (week 8) A brief conclusion: which feature representation worked the best and why

This is a group assignment. I advise you to write your report in Overleaf. More specifications for the report for assignment 2 will be provided in the following weeks.

This week, you will work on part 4 and 5 of the report.

Goals for week 6

- Learn to evaluate the output of a predictive model in scikit-learn
- Learn how to pre-process semi-structured data to be used in a supervised learning task
- Learn to compare different feature sets for the same task

Preliminaries

Your code and completed tasks from week 5.

Tasks

1. Evaluation

In the code by Yao-Jen Chang, predictions for the test set are written to y_pred. We cannot evaluate these predictions, because the test set is unlabelled. Therefore, we evaluate our methods on a subset of the training set.

1. Make a 80-20 split of the training set, using 80% for training and 20% for testing using the train_test_split function in sklearn.
2. Evaluate the predictions on the test set in terms of Root Mean Squared Error (RMSE). Verify that your result is close to 0.48.

The obtained result is your baseline result. Make sure that you use the same train-test split in every run. Be aware that lower RMSE scores are better.

3. Evaluate the matching without stemming for search terms, product titles, and product descriptions.

2. Improving the matching

Add features for the query-product matching and evaluate the efficacy of each feature. A few suggestions are:

- Add features for matching query terms to the information in attributes.csv
- Use the structure of the attribute-value pairs to make better informed features
- Replace the simple term count matching functions with other overlap weights. You might consider using the function TfidfVectorizer in sklearn or the text similarity function in the spacy package.

Be creative: use any information from the queries and products that might improve the matching.

A description of the features you implemented, and an evaluation of the feature sets are part of the report for assignment 2.