Practical session week 5

Data Science course

This week we start 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 1, 2 and 3 of the report.

Goals for week 5

- Get acquainted with dataframes in Python
- Get to know the Home Depot data set through data exploration
- Learn to replicate an existing baseline

Preliminaries

Make sure you have installed Python 3 and the following packages: pandas, numpy, scikit-learn, matplotlib, nltk.

I advise you to work in a Python IDE such as Anaconda or Pycharm, or use jupyter notebooks. You can also use Google colab.

Tasks

1. Preparation

1. Download the data from https://www.kaggle.com/c/home-depot-product-search-relevance/data and unzip all files. You now have a directory with four csv files and one docx file.
2. Import the csv files in Python as separate Pandas dataframes.
3. Read the information on the task and the data. Provide a task definition in your report for assignment 2.

2. Data exploration

Answer the following questions about the data:

1. What is the total number of product-query pairs in the training data?
2. What is the number of unique products in the training data?
3. What are the two most occurring products in the training data and how often do they occur?
4. Give the descriptive statistics for the relevance values (mean, median, standard deviation) in the training data.
5. Show a histogram or boxplot of the distribution of relevance values in the training data.
6. What are the top-5 most occurring brand names in the product attributes?

The answers to these questions are part of the report for assignment 2.

3. Replication of baseline method

1. Download the code by Yao-Jen Chang, available here: https://www.kaggle.com/wenxuanchen/sklearn-random-forest, adapt the paths to the data files, make sure you have all required packages, and run the script.
2. Describe the method implemented by Yao-Jen Chang, including the data pre-processing (it is likely that you will need to look up the documentation for some functions). Make sure that you understand each step of the method.

This description is part of the report for assignment 2.