同濟大學
TONGJI UNIVERSITY

同濟大學軟件學院
SCHOOL OF SOFTWARE ENGINEERING

School of Software Engineering

Data Analysis and Data Mining

Assignment

Semester 2 2018/2019

30 April, 2019

1nd SSE Challenge on Data Analysis and Data Mining

We are delighted to invite you to participate the first SSE Challenge on Data Analysis and Data Mining. (-:

Instructions

You can work as at most three-person groups. You can select one of the following topics as your target.

Topic1: Averaging the GPS trajectories

The objective is to find methods to average a given set of trajectories so that they would match ground truth segments obtained from Open Street Map. The difficult point is the definition of the distance between two trajectories. Or, you can consider the problem as a trajectory clustering problem, where you need to find a representative trajectory to represent each cluster.

Training data has been given. Submitted methods should work on all training data but expected to generalize to other similar data beyond the training set. Each submission must contain:

Source code

Method description

Citation (if method is existing)

All submissions will be evaluated on a different test dataset, which will be similar to the training set but larger. The resulting segment averages will be compared with ground truth extracted from OpenStreetMap. The main criteria in the competition will be quality. The evaluation will be done by visual inspection and an objective measure. Trajectory similarity will also be calculated using other measures (DTW, Hausdorff, and Frechet) to gain more insight.

For datasets and more details:

http://cs.uef.fi/sipu/segments/training.html

Topic2: Product size fitting prediction

Product size recommendation and fit prediction are critical in order to improve customers' shopping experiences and to reduce product return rates. However, modeling customers' fit feedback is challenging due to its subtle semantics, arising from the subjective evaluation of products and imbalanced label distribution (most of the feedbacks are "Fit"). These datasets, which are the only fit related datasets available publically at this time, collected from ModCloth and RentTheRunWay could be used to address these challenges to improve the recommendation process.

Following type of information is available in the datasets:

ratings and reviews

fit feedback (small/fit/large)

customer/product measurements

category information

These datasets are highly sparse, with most products and customers having only a single transaction. Note that, here a 'product' refers to a specific size of a product, as our goal is to predict fitness for associated catalogue sizes. Also, since different clothing products use different sizing conventions, the sizes should be standardized into a single numerical scale preserving the order. Please read the paper for further details.

Rishabh Misra, Mengting Wan, Julian McAuley "Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces". RecSys, 2018.

The baseline of the task is: accuracy as 0.65 for RentTheRunWay and 0.61 for ModCloth.

Each submission must contain:

Source code

Method description and experimental results

Citation (if method is existing)

ASSESSMENT

50% of the scores are given from the performance of the methods on the topic. The performance is considered mainly from the effectiveness, and the efficiency as a minor criterion. 20% are from the documentation/ reports. The rest 30% are from the presentation. Each presentation should be in 8 minutes. As a team work, you should specify the contributions of each team member in the documentations and presentations.

No plagiarism from other teams!

Deadline for submission: 20th June, 2019
Presentation date:
Venue: Room 434 in SSE building
Date : 21st June, 2019