Sunku Bhanu Kedhaar Nath
Z1974769

CRITIQUE PAPER

REFERENCE :

Cong Yan and Yeye He. Auto-Suggest: Learning-to-Recommend Data Preparation Steps Using Data Science Notebooks.

- The first and foremost is that it is still in the theoretical based would need tons of resources and big team to develop the Auto-suggest.
- The paper talks a lot about processing GitHub notebooks. But paper also says there are 4 million notebooks present in the Github. It's gonna be huge data of notebooks to parse. There it needs lot of computational power and robust natural language processing models to understand the huge data of notebooks. Running each notebook takes computationally a lot of time. One better way is to understand the code in each cell and ignore the cells that take huge time such as visualizations. Although it is difficult to say whether the particular cell is manipulating the data that is used in the future cells. But it would be great addition if figured out a way.
- There is no conditional joins in pandas as many data scientists use mostly pandas for data analysis. I believe there will be less data for join suggestion as any join between table are proffered to have conditions and not just using Natural join by default.
- The paper talked about handling missing data files and missing packages but what about versions of the it for example an notebook is using sklearn package which was still sklearn at the time of creation of the notebook but now it has been depreciated (or not supported anymore) and now it is called as scikit learn and many functions inside them have changed their names of their keyword arguments.
- The model can use data structure to keep track of cells only for particular session because GitHub notebooks can be updated by the GitHub user any time they wants.
- Most of the operations that pandas have such as Group by, merge, pivot, stack can be implemented in excel. Mostly the paper suggest users which columns joined would be best fit or pivoting by certain columns would be better fit. What if the suggested columns are not better fit those columns might be of same domain but

may not be relevant the paper states this problem but the solution towards it isn’t clear in my opinion.