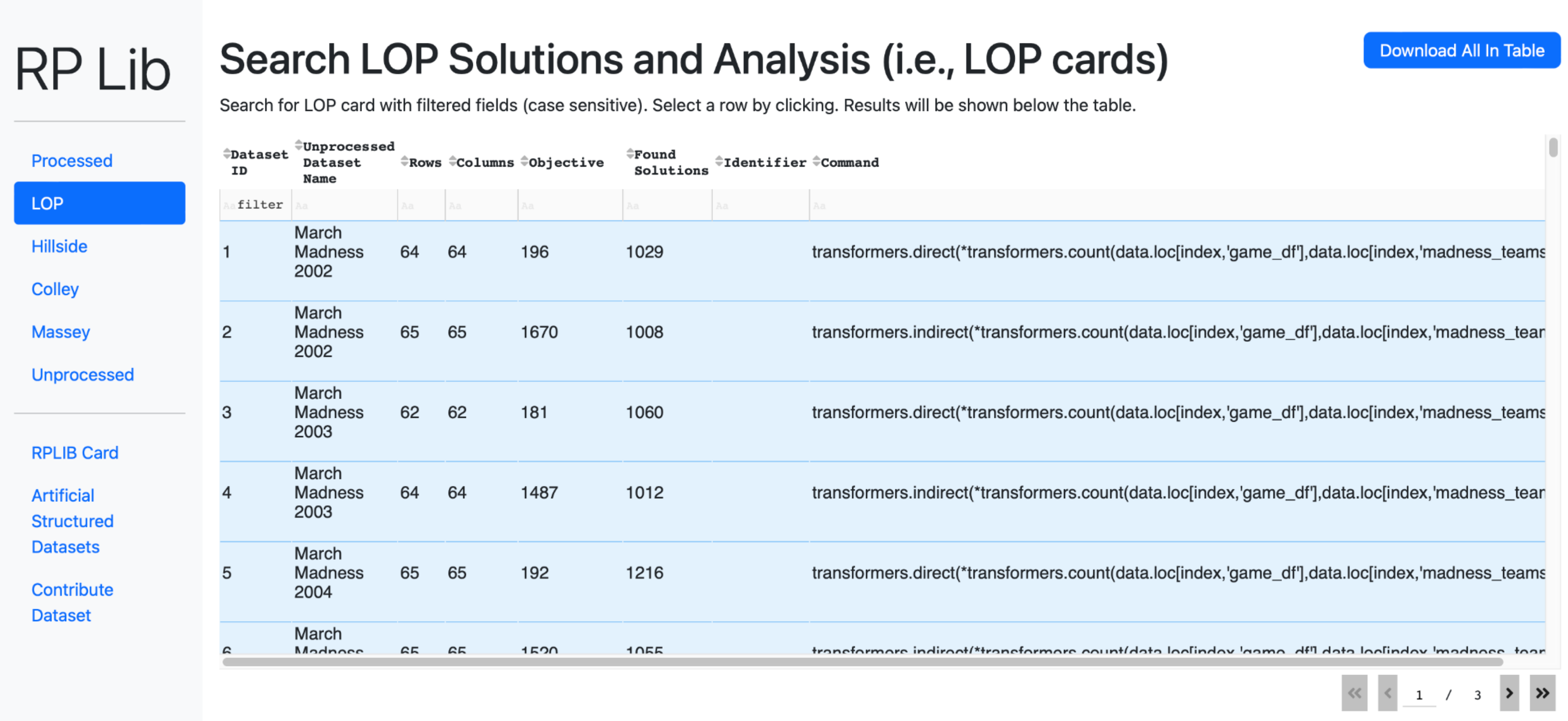
RPLIB and ‘Rankability’

Advisor: Dr. Paul Anderson
Students: Brandon Tat and Charlie Ward
Department of Computer Science and Software Engineering
California Polytechnic State University



RP Lib

- Processed
- LOP
- Hillside
- Colley
- Massey
- Unprocessed
- RPLIB Card
- Artificial Structured Datasets
- Contribute Dataset

Search LOP Solutions and Analysis (i.e., LOP cards)

Download All In Table

Search for LOP card with filtered fields (case sensitive). Select a row by clicking. Results will be shown below the table.

Dataset ID	Unprocessed Dataset Name	Rows	Columns	Objective	Found Solutions	Identifier	Command
filter							
1	March Madness 2002	64	64	196	1029		transformers.direct(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_teams
2	March Madness 2002	65	65	1670	1008		transformers.indirect(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_tear
3	March Madness 2003	62	62	181	1060		transformers.direct(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_teams
4	March Madness 2003	64	64	1487	1012		transformers.indirect(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_tear
5	March Madness 2004	65	65	192	1216		transformers.direct(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_teams
6	March Madness	65	65	1520	1055		transformers.indirect(*transformers.count(data.loc[index,'game_df'],data.loc[index,'madness_tear

Figure 1. A page of RPLIB website

Abstract

We present an improved library for the ranking problem called RPLIB. RPLIB includes the following data and features. (1) Real and artificial datasets of both pairwise data (i.e., information about the ranking of pairs of items) and feature data (i.e., a vector of features about each item to be ranked). These datasets range in size, application, and source. (2) RPLIB contains code for the most common ranking algorithms such as the linear ordering optimization method and the Massey method. (3) RPLIB also has the ability for users to contribute their own data, code, and algorithms. Each RPLIB dataset has an associated .JSON model card of additional information such as the number and set of optimal rankings, the optimal objective value, and corresponding figures.

Results

RPLIB

A web application in support of the Ranking Problem research community. Functionality includes uploading datasets, processing data to a proprietary standard format, running an automated analysis and statistical summary of the data for three different ranking algorithms, downloading datasets uploaded by other users, and searching for datasets. In tandem with this application we developed tools for researchers to run ranking algorithms on a given dataset, perform complex data transformations, and automate much of the data research process.

Beta Coefficient

A statistical metric that provides a quantitative approach to analyzing the fairness of using a particular ranking algorithm to rank a dataset. This is derived from the ranking inversions in the set of ranking solutions. A larger value indicates that the ranking being done for higher ranked data is unfair whilst a lower value indicates that there is likely less error in the highly ranked data.

X_b^*= a binary matrix where $[X_b^*]_{i,j}= 1$ if $X_{i,j}^*$ is fractional and 0 otherwise.

B_n = an $n \times n$ banded matrix. Then, $[B_n]_{i,j}= |j-i|$.

$$\text{For example, } B_4 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix}.$$

W_n = an $n \times n$ weighted matrix where each entry in row i is $\frac{1}{i}$.

$$\bar{\beta} = \sum_{i=1}^{n} \sum_{j=1}^{n} [X_b^*]_{i,j} * [B_n]_{i,j} * [W_n]_{i,j}.$$

$$\hat{\beta} = \frac{\bar{\beta}}{\beta_{max}}, \text{ where } \beta_{max} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|j-i|}{i} \text{ (maximum } \bar{\beta} \text{ given n).}$$

$$\beta = \frac{\hat{\beta}}{\#\text{ of fractional entries in } X^*}$$

A

Figure 2. Formal Definition of the Beta Coefficient

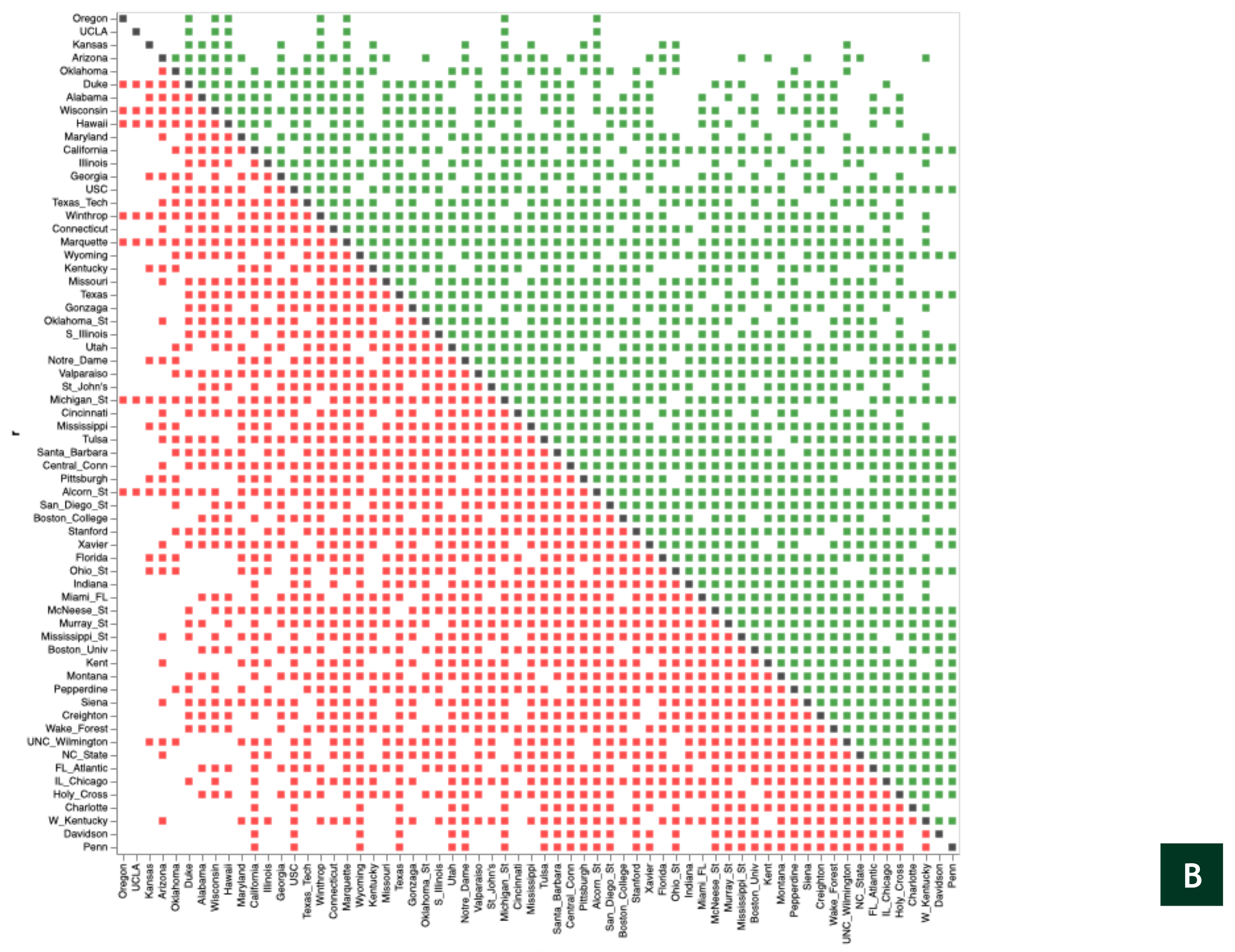


Figure 3. An X matrix generated from all possible solutions of the Linear Ordering Problem (LOP) on March Madness data*

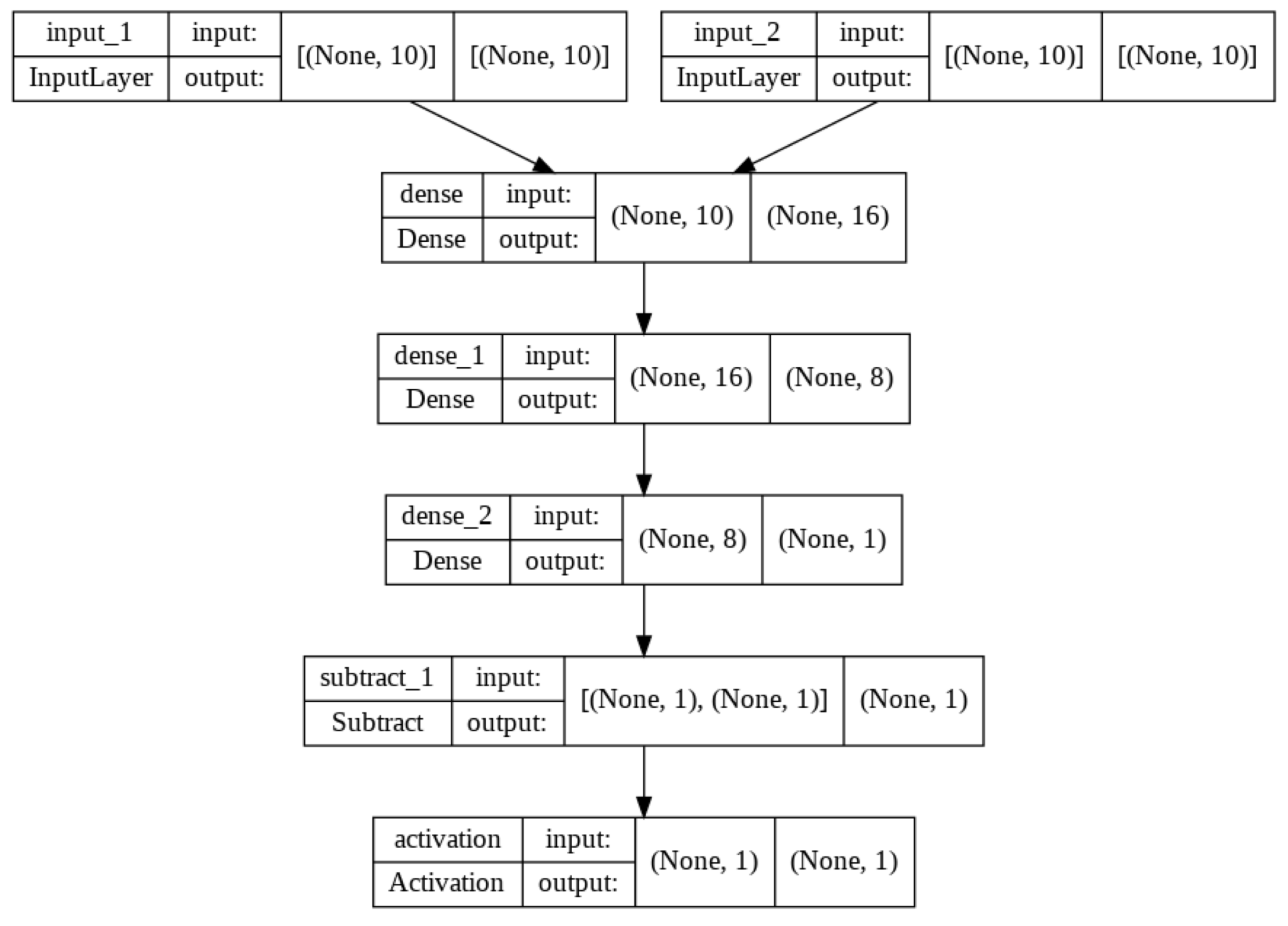


Figure 4. RankNet, a learning-to-rank algorithm and its architecture

Reflection on CCDSF

Brandon

Despite not being part of the data science minor offered at Cal Poly, I was still able to enroll into fellowship which opened many opportunities for learning. I was able to meet like-minded individuals who were passionate about data science and connect with faculty who were working on cutting edge problems relating to data science. During the fellowship, I was able to participate in paper reviews where we discussed seminal papers in the fields of statistics, databases, and machine learning. We discussed the history, present, and future of data science. On top of that, the fellowship brought in experienced individuals from industry who talked to us about their experiences and journey. I was able to work with an amazing team where me and Charlie were able to co-author a paper along with them.

Charlie

Through this fellowship I’ve had the chance to gain research experience, learn about the past, current, and future of data science, connect with faculty, and gain experience in teaching. Through Brandon and I’s research, I co-authored a paper and learned about the publication process as well as best practices in technical authorship. Throughout the fellowship’s social events and seminar, I’ve had time to interact more closely with Cal Poly’s faculty and become more informed as to what I want to learn more about. Through the outreach activities, I’ve gained experience as a TA and as a teacher through holding office hours and leading a high school Computer Science lesson. This fellowship really has been an invaluable experience and has greatly informed the path I hope to pursue professionally.