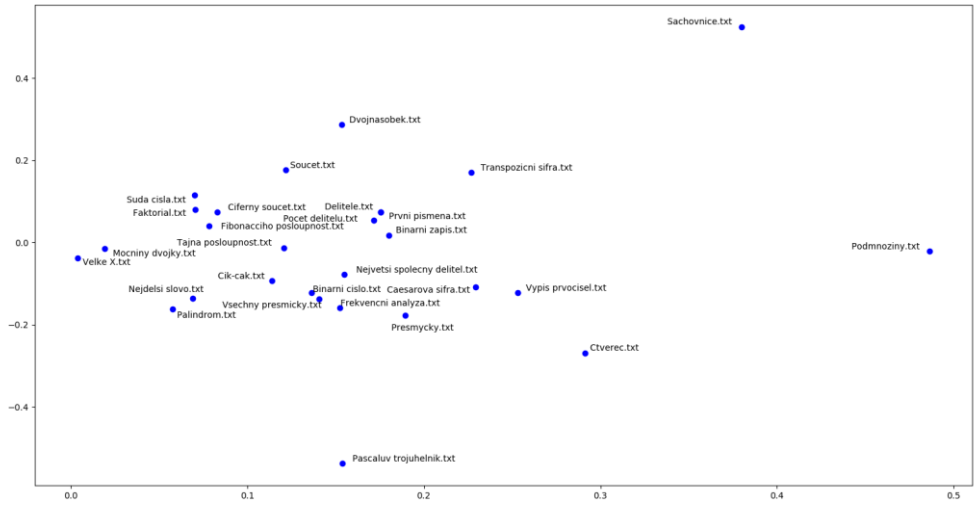
Item similarity

Vojtěch Sassmann

1. Parsing user's solutions into AST and analyzing the node types and calculating feature matrices

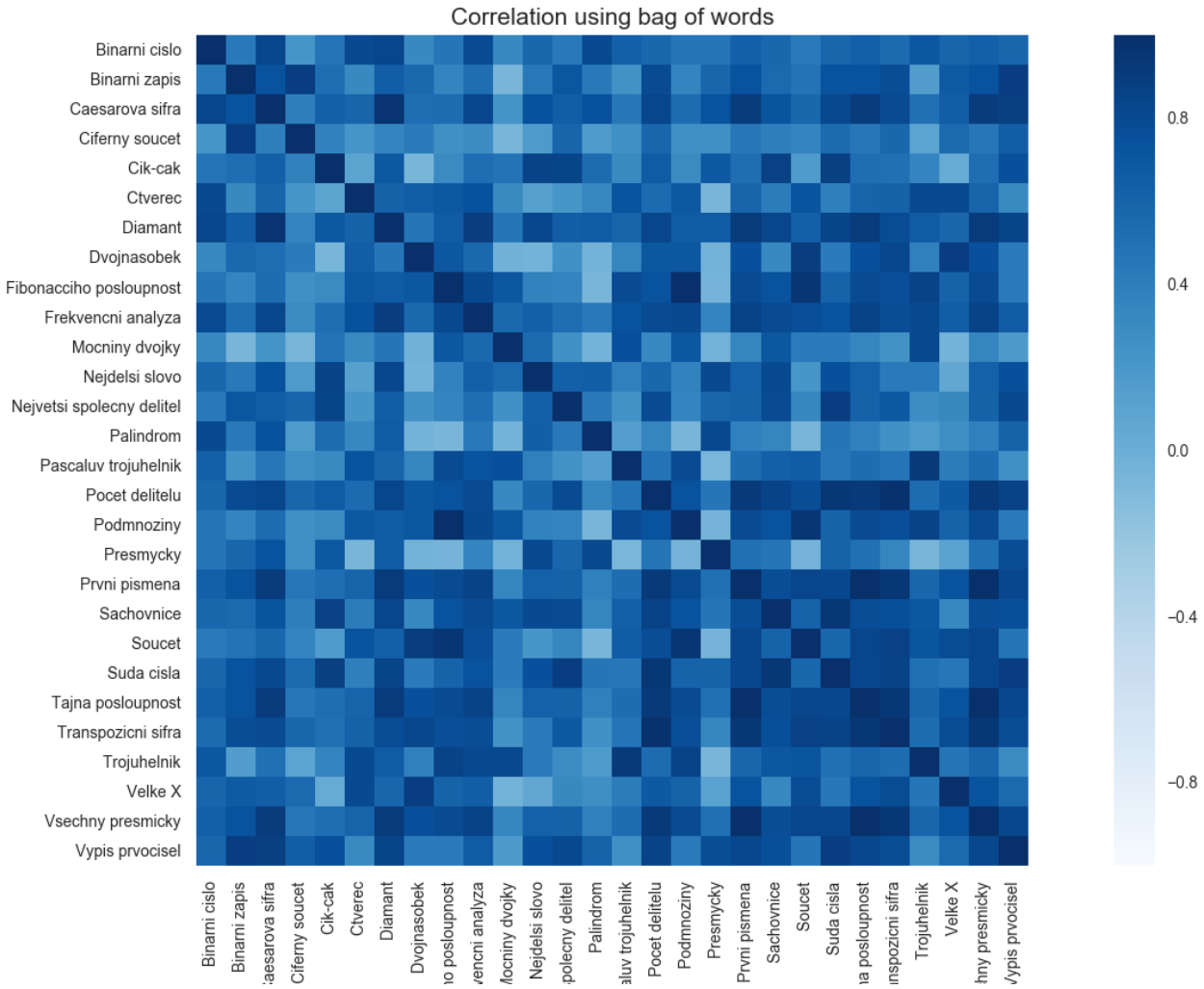
1.1 Created naive PCA projection



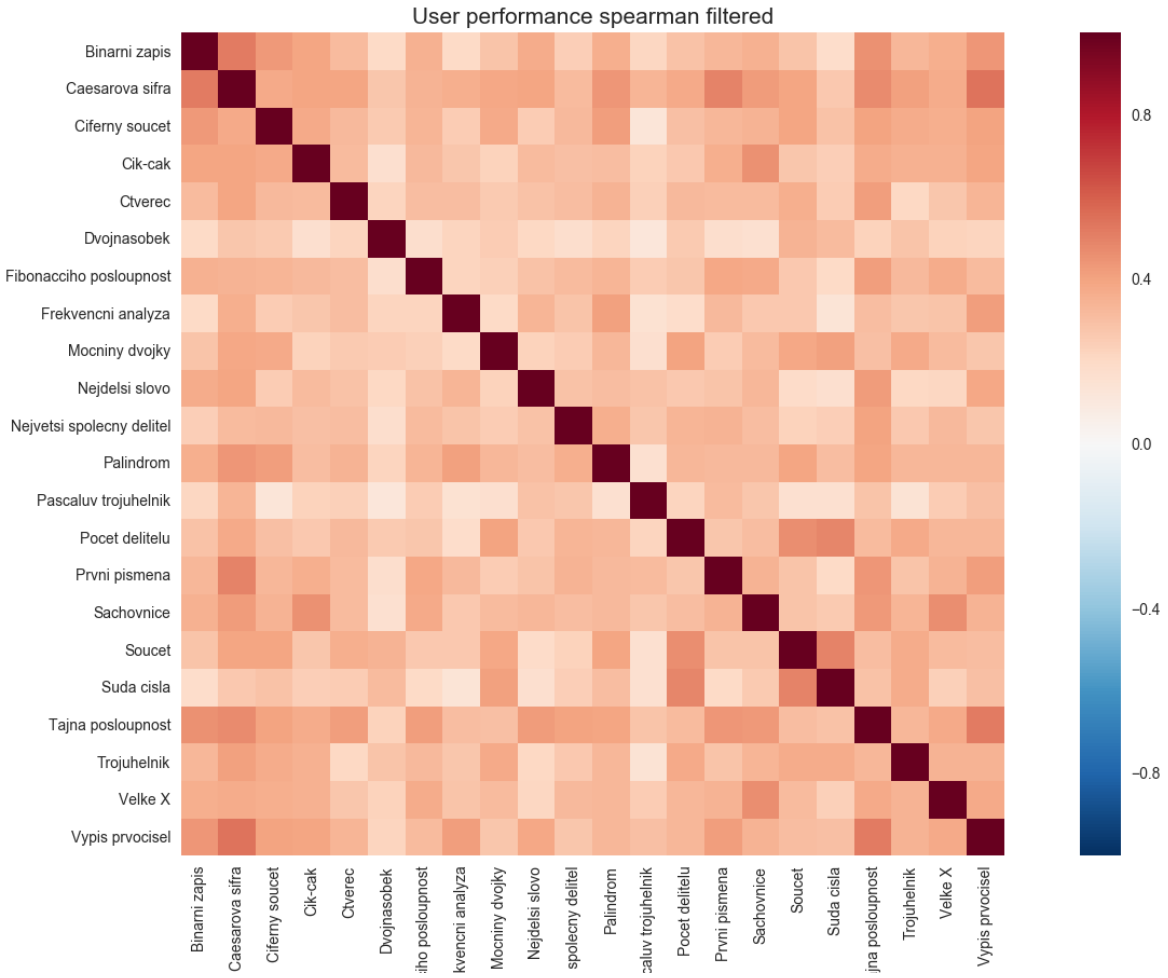
1.2 Discovered problem with python's AST -> decides to use simple BagOfWords method to calculate feature matrices.

2. BagOfWords analysis – Python solutions -> feature matrices -> item similarity matrices

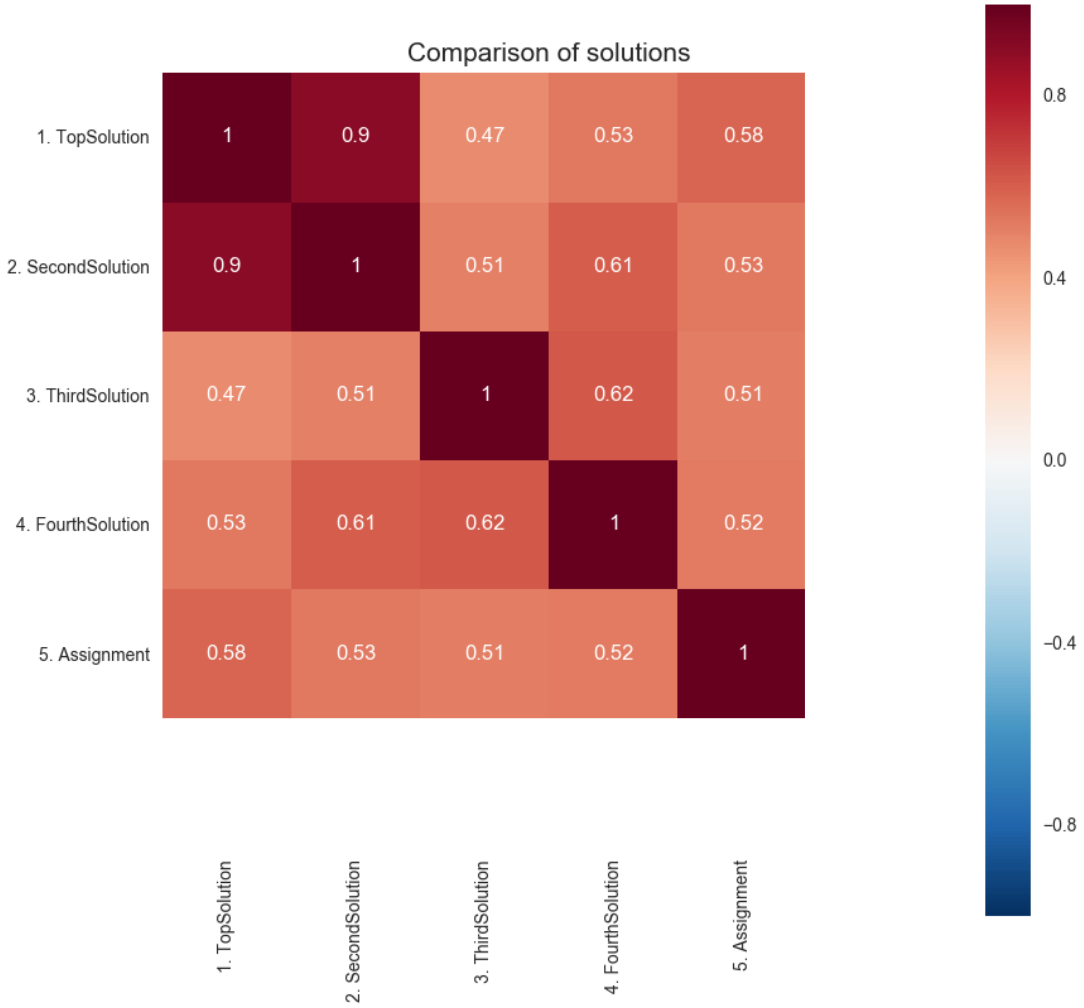
2.1 Created first pearson correlation of calculated features



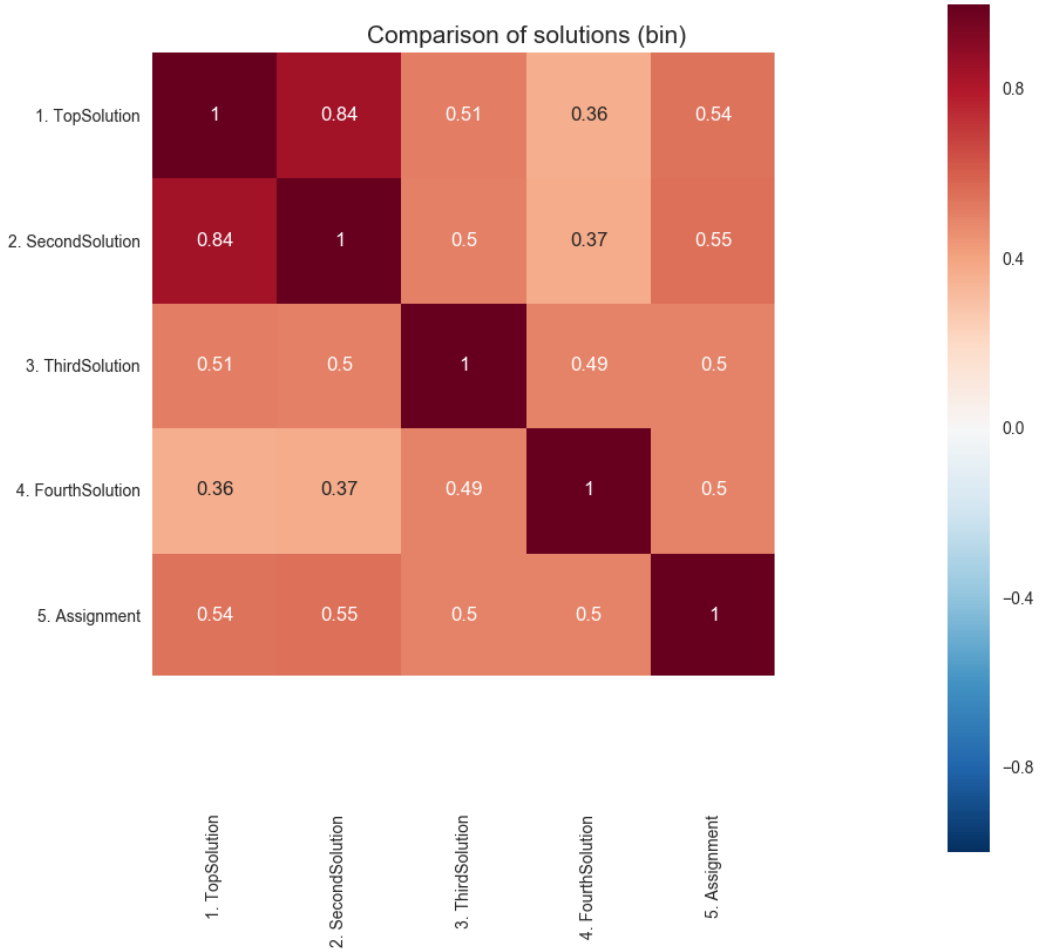
2.2 Trying to get rid of the noise in data -> filtering items with less then 300 solutions and using spearman correlation.



2.3 Trying to discover, how much the results are affected by choosing the best, second best, etc. solution. Created correlations of features with those solutions, then flattened and correlated again. Using two versions of feature matrices – with count of feature occurrence and with discrete occurrence (bin)



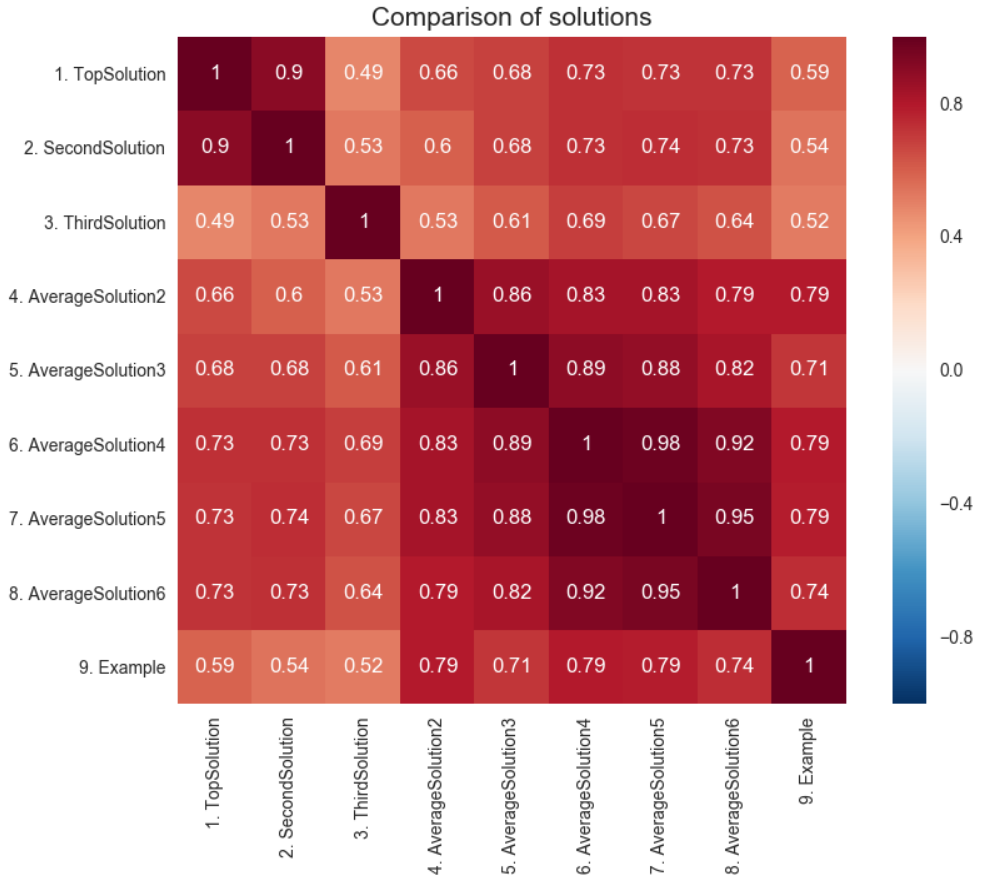
- Note: assignment is bad label, it should be “Sample solution”



Very low correlation -> depends a lot. So I have decided to choose a different approach.

2.4 Created new method when the average solution is created from best solutions.

- Note: AvarageSolution2 means, that were used two best solutions to calculate the average solution feature matrix, etc…



- We can see that when just two best solutions were used to calculate the feature matrix, the correlation with example solutions has increased -> that's what we wanted. It also shows, that it does not matter that much how many of top solutions are used so I decided to go with 4.