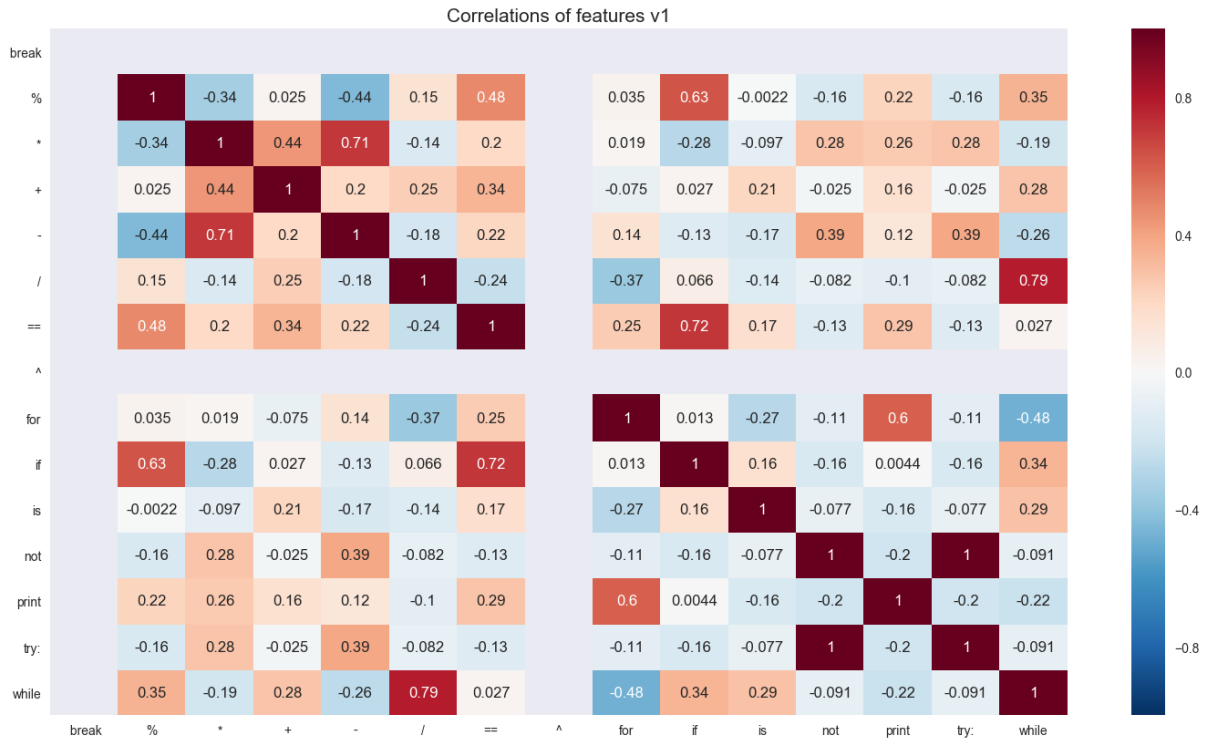
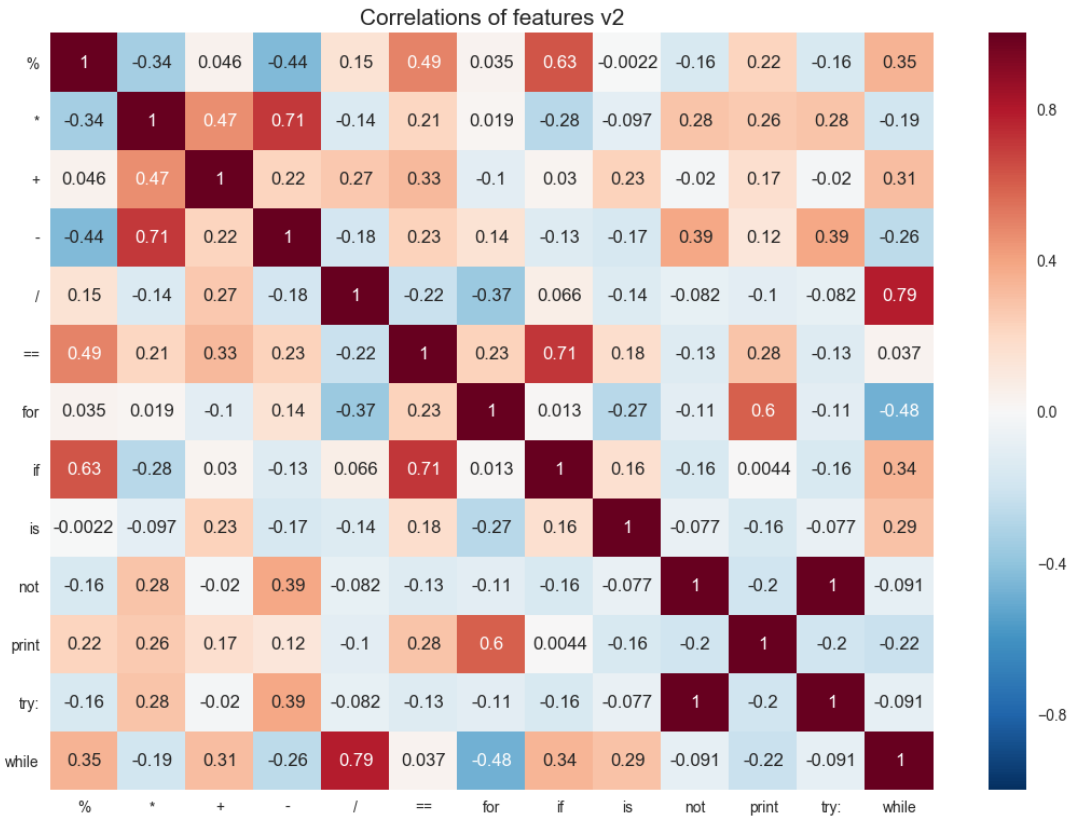
Decision of what features to use

First try:



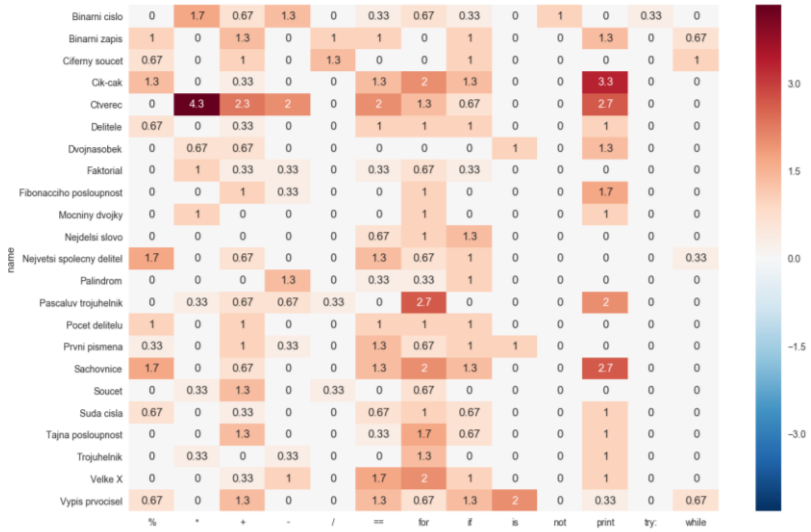
Results: the ‘break’ and ‘^’ features were not found at all so I decided to remove them.



Next try with removed ‘break’ and ‘^’ features:

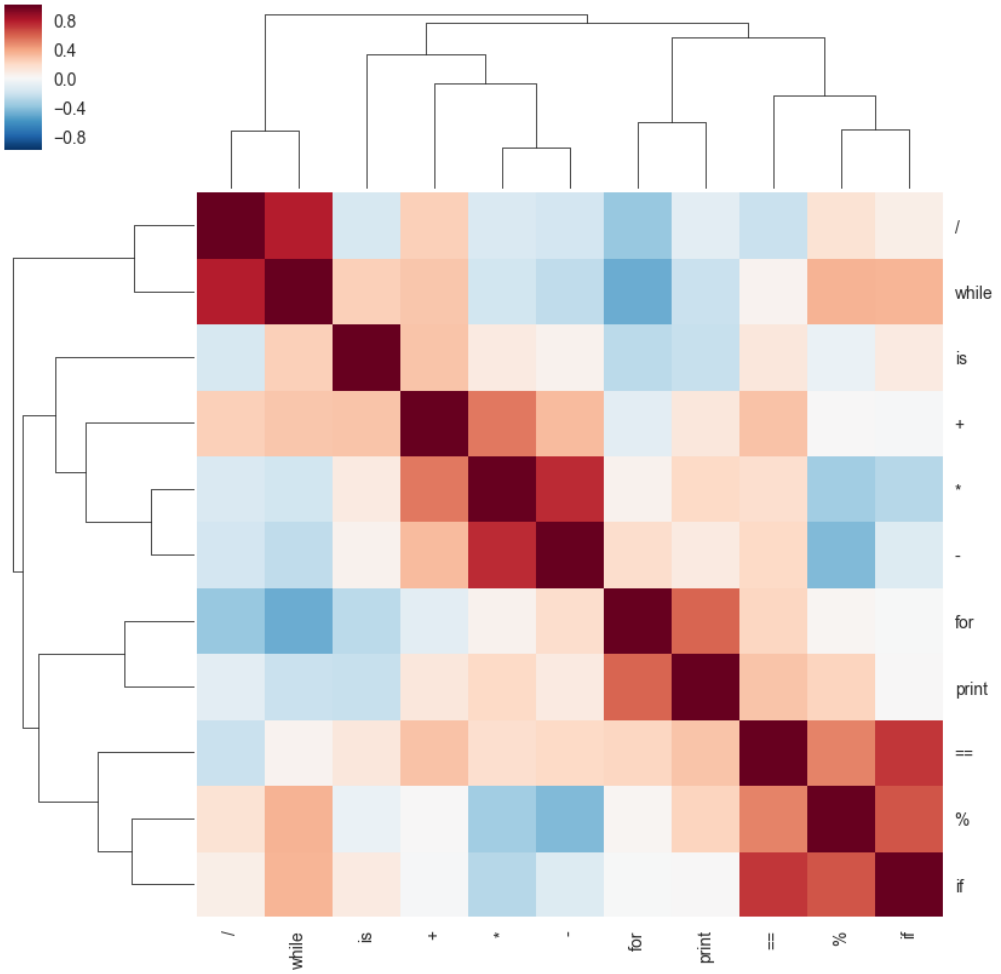
I had noticed that the correlation of ‘try’ and ‘not’ is too high so I decided to take a closer look at it.

This is the feature matrix used in the comparison above:



I have noticed that the ‘try’ and ‘not’ features have been used only in one item so they have a high correlation. I have decided to remove them as well.

The next correlation looks like this:



Results: We can see that some of the features has a high correlations but I think it is not that high that it must be removed. I think this features are the ones, that I would use.

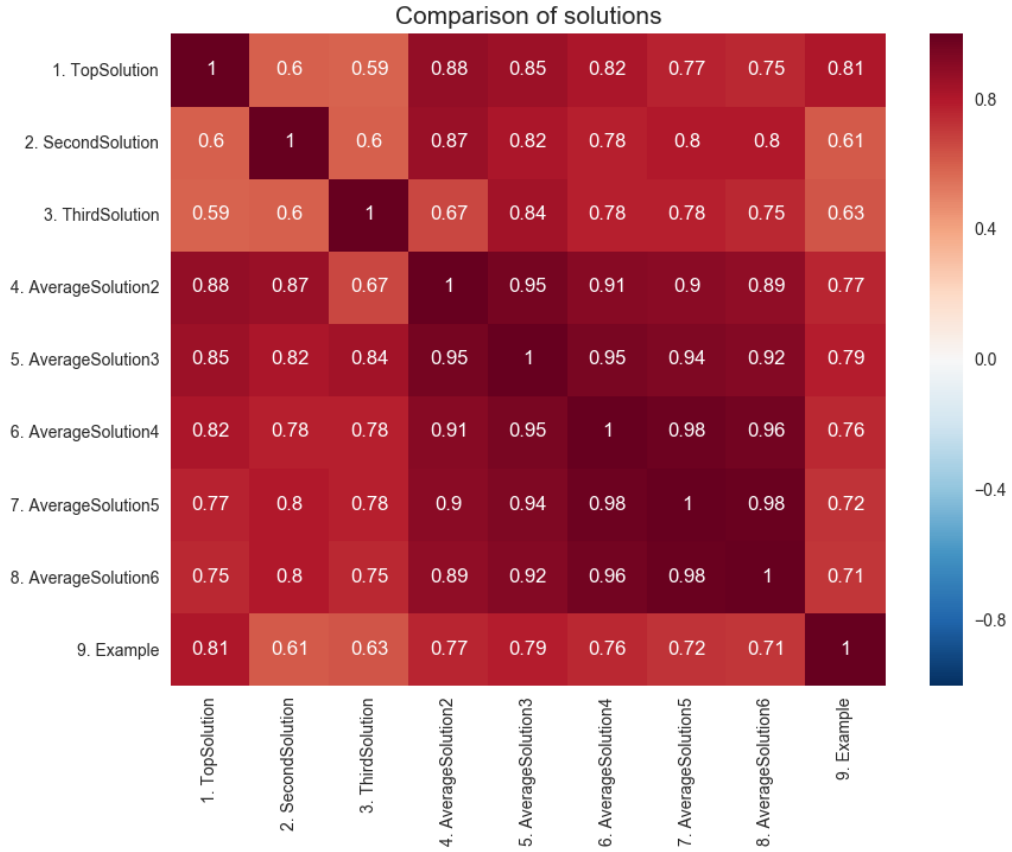
Decision how to process given solutions

I have come up with two methods of using the user's solutions.

- Use one of the most frequently submitted solutions
- Do average of those solutions.

Using similarities with all other items

I have created correlations for each of this method and then compared those methods how they

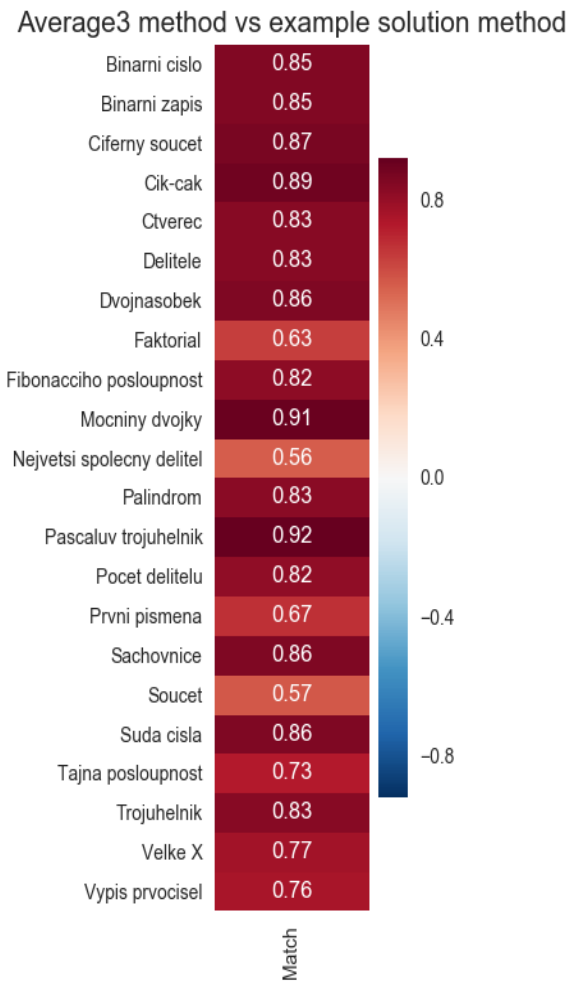


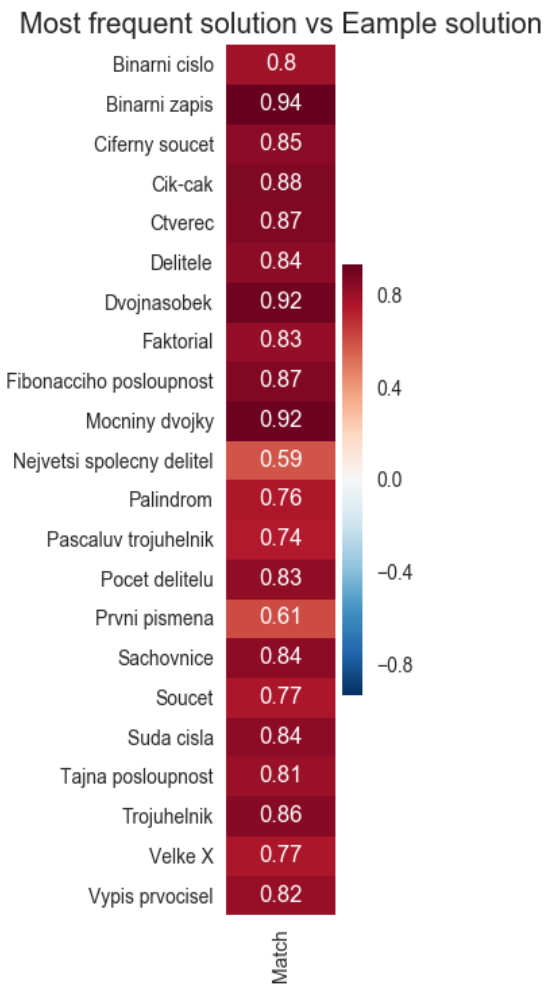
correlate. I have also included the 'Example' which is technique using bag of words on example solutions given by tutor.

From this heatmap we can see that the method using the most frequent solution 'TopSolution' and the method using the average of the three most frequent solutions 'AverafeSolution3' are the ones that have the highest correlations with the example solution.

Now I have to decide which of those two methods is better.

I have compared those methods with the example method by each item:





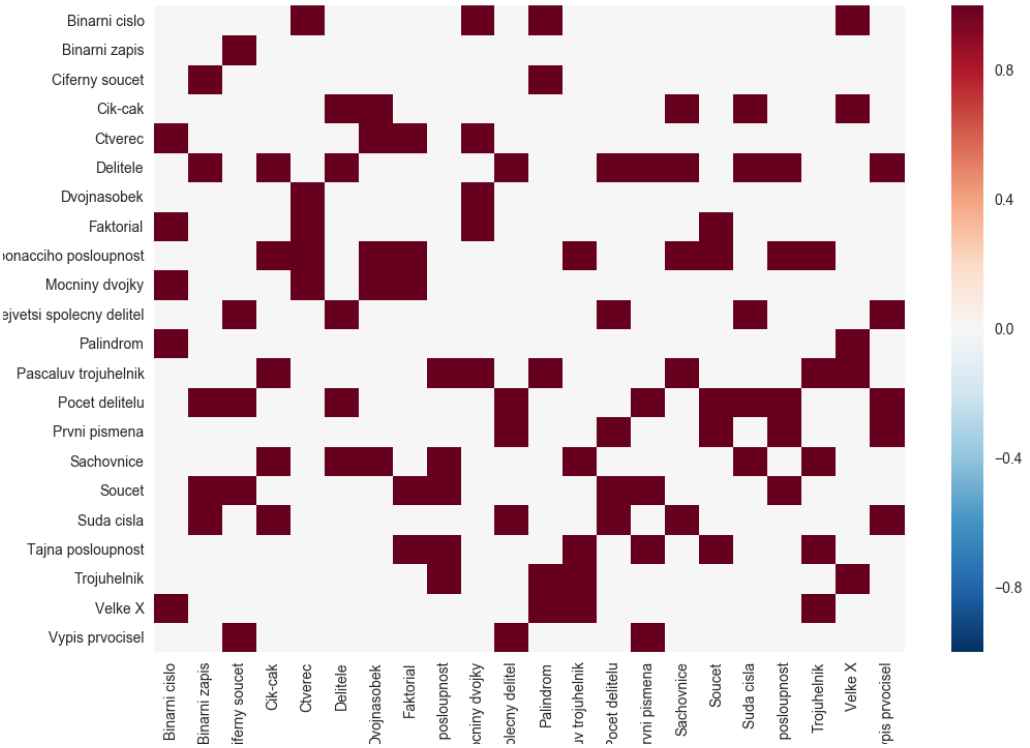
We can see that the method using the most frequent solution has some higher correlations for some of the items but that's only because the most frequent solution was very similar to the example solution. The method using the average has lower average correlation with the example solution method but I think that it is a good think. It means that this method might include some solutions that are different than the example solution.

Decision: I would prefer to use the average solution method because it can do a better job comparing items which might have more than one correct solution.

Using top 5 most similar items

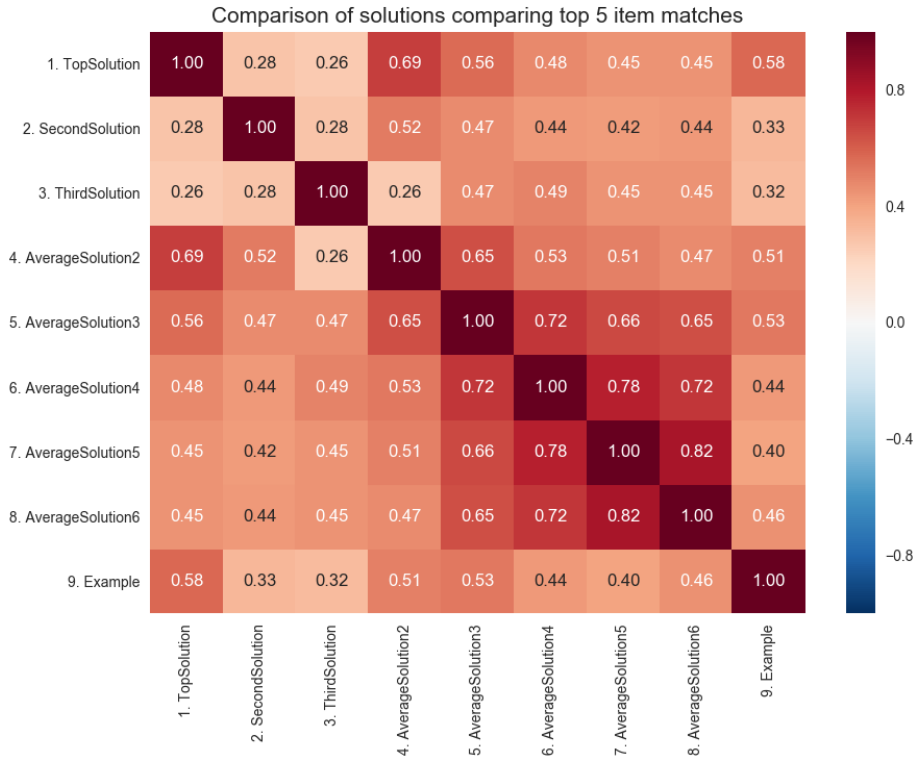
Now let's take a look what would happen if we compare the solutions methods using only the best 5 item similarities.

Example of such correlation with only top 5 matches (This one is for TopSolution):

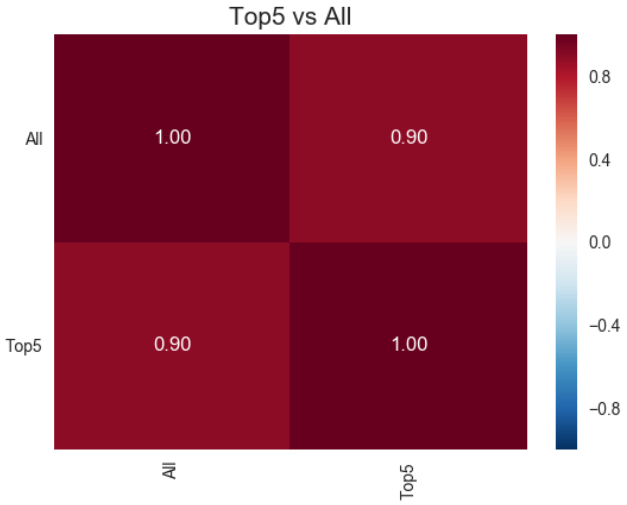


We can see that for each column there are only 5 values.

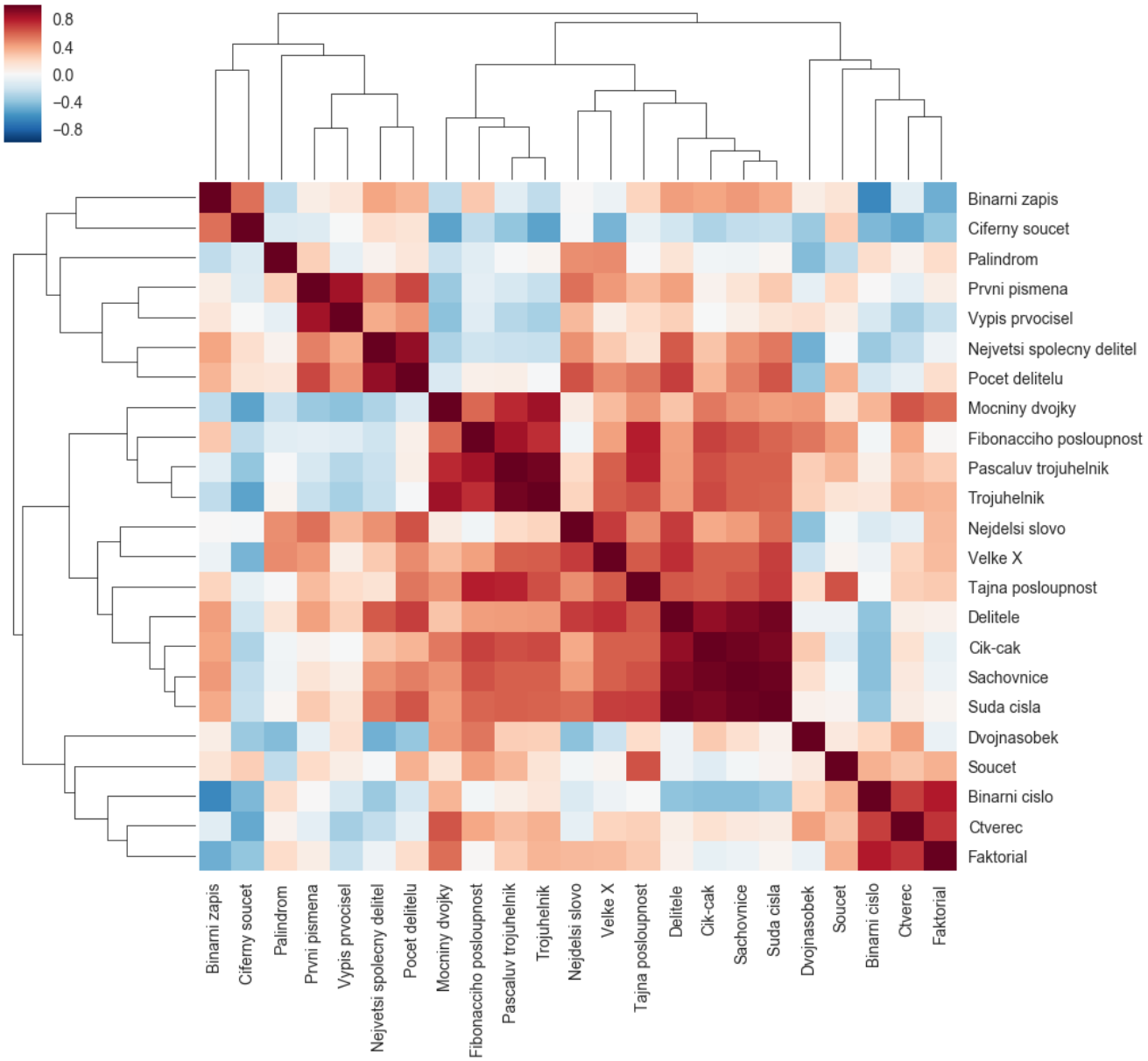
Now lets do correlation of these metrics:



And now we compare this with the previous technique using all similar items:



We can see that the correlation of these two techniques is really high so I think we do not need to explore this technique any more.



Here is show the comparison of the items using the method I have chosen in the text above:

This comparison is using the Average3 method and using only those features I have decided to use in the text above.