 [Translated from Russian to English - www.onlinedoctranslator.com](https://www.onlinedoctranslator.com/en/?utm_source=onlinedoctranslator&utm_medium=docx&utm_campaign=attribution)

National Research University Higher School of Economics

Graduate School of Business

Master's Program “Business Analytics and Big Data Systems”

**Project report**

\_\_\_\_\_\_\_\_\_\_\_\_

**Measuring the effectiveness of various**

**aggregation procedures**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Project title)

Performed by student

*Shumeiko Lyubov Anatolievna*

*(student's full name)\_*

**Project supervisor:**

*Veselova Julia Alexandrovna,*

*Associate Professor*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(position, full name)*

*\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_*

*(date) (signature)*

**Moscow 2024**

**Project's report structure**

1. Title page………………………………………………………………………………
2. Table of content………………………………………………………..
3. General description of the Project:…………………………………….
4. Initiator, head of the Project………………………………………………………………
5. Type of the Project……………………………………………………..
6. The main part……………………………………………………………..
7. Description of the implementation of the Project's work……………………..
8. Description of the Project's results……………………………………………………………
9. Description of the used methods and technologies used………………………
10. Description of challenges and difficulties encountered during the Project……
11. Conclusion …………………………………………………………………………………..
12. Project's results………………………………………………………………………………
13. Appendix………………………………………………………………………………..

General description of the Project.

This research is devoted to optimizing the mechanism for aggregating individual preferences into collective ones using the Kemeny distance, while maintaining the computational simplicity of simple aggregation methods.

The work is expected to conduct a thorough analysis of existing aggregation methods, identify their advantages and disadvantages, and develop improved algorithms that will more effectively take into account the Kemeny distance. It is expected that the results of the study will contribute to a more accurate and rapid integration of individual preferences, which is especially important in the context of collective decision making

Initiator, head of the Project and Structural Components

This work is an individual research project in the format of coursework.

Structural unit/other

Relevance of the research topic and practical value of the study: the mechanism of a technically simple but individually effective method of preference aggregation can find wide application in recommendation systems focused on processing large volumes of data in real time. This approach is particularly suitable for solutions that require immediate decision making based on large levels of incoming data and do not require long-term storage. A high focus on practical issues determines a fairly high level of theoretical attention to the issues of optimizing algorithms, and in particular methods for ranking preferences over many years.

Goal of the work: to optimize the mechanism for aggregating individual preferences into collective ones by Kemeny distance, while maintaining the computational simplicity of simple aggregation methods.

The challenge is to:

● Obtaining and comparing results using different aggregation methods

● Evaluation of the results of application of preference aggregation methods based on audience satisfaction using Kemeny distance

● Finding better optimization of existing simple aggregation methods by combining aggregation methods.

Description of the report structure: in the first sections of the work structure, a general description of the project is given, an explanation of its relevance of goals and objectives, the type of project is determined, the main part provides a detailed description of the work, including the stages of the study, the methods used, intermediate results and the main steps taken to achieve the goals of the project, a description of the results obtained and their significance in the study, the theoretical and methodological basis used to achieve the goals is given, the choice made is justified in terms of methods and means used to achieve the goals

The main part:

**Description of the implementation of the Project's work**

A detailed description of the work performed, including the stages of the study, the methods used, intermediate results and the main steps taken to achieve the project goal

The stages of this work were the first definition of the main issues and the construction of work goals, organizational issues (the purpose of the organizational stage was to determine the main characteristics of the work and the formation of the necessary basis for conducting the research), the next next stage was theoretical research, literature review and the final stage of the study was practical implementation, analysis the obtained results and their visualization. In accordance with the basic data, logical stages will be described and a more detailed description of the stages of the study will be made.

**Theoretical stages of research**

As part of the theoretical stage of the study, several important areas were covered and considered (differing from each other in the purpose of conducting theoretical analysis, within the framework of which various range of issues were considered:

Obtaining general information about recommender systems, decision-making methods (gaining context: areas of implementation and application of the research object)

Research on ways to determine both individual preferences and aggregation of individual preferences into collective ones (the goal is to understand the mechanisms of operation of aggregation methods and issues generated in the scientific community)

Identification of characteristics obtained as a result of literature analysis and selection of certain methods, their lexicographic description and formation of universal formulas reflecting the research model (formation of one’s own knowledge and its qualitative design in accordance with the objectives of the study)

Analysis of the vector of work already carried out and comparison with upcoming work (comparison of methodological and scientific foundations in order to confirm and adjust one’s own chosen methods and questions, reflected in more detail in the Literature overview as part of the course work itself)

Analysis of the practical part (code implementation) of previously carried out work (the goal is a practical understanding of the progress of the experiment, assessment of the implementation of the necessary tools in one’s own work and their effectiveness)

check it out

Description of the results obtained as a result of theoretical analysis.

(more details about each stage and the results)

**Description of the project progress**

1. \*\*Preparatory stage:\*\*

- Code was written and a list of necessary materials was generated to study the features of aggregation.

- A database of different sizes was generated, including data sets with 10 alternatives and 100 different options, integrated 10 times.

2. \*\*Practical comparison of aggregation methods:\*\*

- A practical comparison of aggregation methods in terms of time and efficiency, as well as an independent assessment of these parameters, was carried out.

- The work of individual aggregation methods was studied with an assessment of time costs relative to all parts of the function.

3. \*\*Identified problems and their solutions:\*\*

- At the stage of applying the Leximin Lexmax rule, a defect was identified: despite the execution of the code function to determine candidates, the rest of the data continued to be read, which technically unnecessarily complicated the operation of the method.

4. \*\*Comparative study of ranking methods:\*\*

- A comparative study of the effectiveness of ranking methods relative to the Kemeny distance was conducted.

- Standard deviation analysis

(The standard deviation of the performance of aggregation methods relative to their indicators has been studied)

6. \*\*Additional study of abrupt changes in Kemeny distance:\*\*

7. \*\*Building efficiency charts:\*\*

- Graphs were constructed reflecting the effectiveness of the aggregation method in terms of operating time and Kemeny distance.

8. \*\*Application of a ready-made function and determining the winner:\*\*

- A ready-made function was applied to the received data, the winner was determined in accordance with the Pareto principle.

9. \*\*Creation and evaluation of combined aggregation methods:\*\*

- Models with STV showed negative results in terms of speed and quality, so they were not used for further combinations.

- An assessment of combined methods showed that the combination of Copeland’s methods turned out to be the most optimal, despite assumptions about the effectiveness of the Bard or Ksenin methods.

10. \*\*Conclusion and conclusions:\*\*

Description and analysis of the results obtained.

**Description of the Project's results**

Description of the results obtained in general and significance:

A drop in the level of pleasure according to Kemeny from 22-24 (on average) to 18 (on average) showed a deterioration in results, therefore, despite the completion of all the tasks set for this study regarding the main goal, this method of combination showed ineffectiveness. However, it led to relatively new (if not for scientific literature, then for personal understanding) conclusions, the emergence of new questions and hypotheses. In particular (a description of the particular results obtained at each stage of the project will be described below):

Their analysis

The results obtained showed the following picture: with respect to individual preferences, the general (relative to some groups) combination of collective ranking methods showed an almost random result. Combining aggregation methods on some date sets gave the best result, but on others the opposite.

A detailed chronological description of the results obtained, their analysis and significance in the context of the project goal.

1. At the stage of assessing the speed of the resulting code, the following results were obtained: the speed of the Couplen 7.5 method, STV and Bard is close to two. 5 seconds. All other methods work up to a second even with a fairly heavy load.

Significance in the context of the project's goals: This procedure, firstly, helped to correctly evaluate the method of code operation and correct the error, and secondly, to empirically feel the speed of the functions.

2. At the stage of characterizing the evaluation of pure aggregation methods, it turned out that the Copeland method and the Bard method have both the same efficiency in terms of performance and an approximate response to changes in the dataset. differ but have approximately the same results and similar behavior Leximax, although with a greater deviation from each other, and a greater spread in the variability of behavior according to Kemeny. The STV model performs the worst and on various data sets it reacts completely differently from other aggregation methods. At the same time, the standard deviation of the STV model is significantly higher than that of other models: 56 models with STV, 50 lyxemax, 49 coolend, 46 borad , 43 leximin.

Analysis and significance in the context of the project's goals: Evaluating the effectiveness of the topic on different data sets helped to form an understanding of the response of different game geating methods to changing individual preferences. and it turned out that despite the absolute difference in the methodology for calculating preferences, absolute and relative indicators practically do not differ from each other. it turns out that on average each participant is individually satisfied in 43% of the best cases, which is a low indicator that aggregation methods take into account individual opinion, however, regardless of the method, they do not significantly help to increase individual satisfaction as such, depending on the method, rather chaotically depending from a specific dataset. This also helped to form an understanding of the random significance of datasets and the formation of new requirements for further correlation and tracking of the internal components of datasets, which was not taken into account previously.

3. When analyzing Pareto efficiency taking into account the parameters of work time and efficiency on the topic, the copland method showed the best efficiency. Analysis and significance in the context of the project goals: this conclusion served as the basis for using the method as the main one when searching for a combination of aggregation methods.

4. The result of the speed of combination of several methods showed the following results: the best combination was the use of the Leximin Leximax method and the Copeland method with a speed of up to 0.001 when the combination of other methods showed results from 0.006.

In terms of Pareto efficiency, the combinations showed the worst results in both relative and absolute terms relative to pure methods. Although the combination of Copeland Bard and Leximin turned out to be more effective, their effectiveness in terms of numbers decreased to 16 on average, which significantly reduced both speed and efficiency.

Analysis and significance in the context of the project's objectives: The results obtained from this study are not clear to me from a logical point of view. Although it allowed us to form an understanding of the actual characteristics of the work of aggregation methods, and the formation of further questions from which it would be worth starting: For example, it raised the question of the need to use a method for solving the problem of kemen in practice, since the search for optimal algorithms based on the distance of the name depending on different data sets in depends largely on chance, which with a large sample and averaging shows relatively identical results, explained rather by statistics and probability, which, in principle, explained the increased interest in the literature regarding this topic to large practical aspects, either regarding specific data in a particular area or to optimization issues and search for patterns, as well as the manipulability of various methods depending on the datasets.

**Description of the methods and technologies used;**

Theoretical and methodological basis necessary to achieve the goals. Justification of your choices in terms of the methods and means used to achieve your goals

Within the theoretical framework on which the research was conducted, use knowledge (questions) generated in various fields:

Methods of voting theory (social choice), mathematical statistics, combinatorics, theory of measurement and aggregation of preferences were used. Methods of optimization, data analysis and machine learning methods are partially covered.

The main theoretical basis on which the study is generally built in isolation from scientific knowledge is the methods of voting theory, the theory of measuring the aggregation of preferences, since they were tested and used in the study.

Optimization techniques have been used to try to combine aggregation techniques to improve code efficiency.

Machine learning methods were used to visualize as part of the imported libraries. mathematical statistics and combinatorics were used to evaluate the conclusions and results obtained as a result of applying methods for working with data on the studied samples.

However, it is worth saying that many methods not indicated in the course work were also studied in the report (for example, as part of the analysis of other studies related to the problems of the work), but since they were not directly applied and used in the work, only taken into account, we consider it necessary to indicate them availability. For example, in their research, many relied on methods of network and cluster data analysis, pattern analysis theories, and graph theory. Although these examples were not used, thanks to paying attention to these areas, an attempt was made to update the work.

For practical implementation, the following methods and technologies were used:

| For visualization | Matplotlib | From Github |
| --- | --- | --- |
| For data repurchase | NumPy  Pandas  SciPy  itertools  collections  and many others | [[1]](#footnote-0)<https://github.com/PrefLib>  [[2]](#footnote-1)<https://github.com/erelsgl/PrefLib-Tools>  [[3]](#footnote-2)<https://github.com/erelsgl/PrefLib-Tools/blob/master/preflibtools/notebooks/MSS_16_Tutorial.ipynb>  [[4]](#footnote-3)<https://github.com/mikedillion/PrefLib-Tools>  [[5]](#footnote-4)<https://github.getafreenode.com/django/django>  [[6]](#footnote-5)<https://hturner.github.io/PlackettLuce/reference/preflib.html>  [[7]](#footnote-6) <https://github.com/logc/borda>  [[8]](#footnote-7)https://github.com/johnh865/election\_sim |

From the above libraries, the following methods were directly applied: the method of generating data, both the library itself and the application technology, the time calculation function was used, an attempt was made to directly use a library that would include aggregation methods. However, as a result of the difficulties that arose, described below, there were functions are applied, not as imported, but as embedded.

Libraries for visualization and statistics have also been fully imported and applied.

**Description of challenges and difficulties encountered during the Project**

The biggest challenge and complexity in this project was its practical implementation due to the fact that the original goal of using ready-made libraries with a good solution as code functionality was impossible to use due to a change in the principles or version of Python, including When implementing a studied library, when Flip selected from a ready-made solution, an error was thrown about the impossibility of importing the downloaded library. This problem is related to the purpose of the fact that Depending on the compilation of the functions that count the aggregation method, its complexity and speed are determined; incorrect writing of such functions would lead to incorrect practical results.

Therefore, in addition to spending effort on the work of libraries, the next challenge was checking and rewriting functions manually, rather than implementing them as a ready-made solution, so organizing your own comparison code and optimizing inside the code, searching for solutions to problems that arose in already open libraries, then further organizing all received functions into a single structure took an unreasonably long amount of time.

**Conclusion**

In the course of the work carried out, the following classes of competencies and individual results were formed/eveloped:

1. Analytical thinking:

Ability to analyze and compare various data aggregation methods, including assessing their effectiveness and time costs.

Improving problem solving skills (problematic thinking)

(Identification and elimination of defects during execution)

Improve statistical analysis and estimation skills. Analyze standard deviations and other statistical measures to evaluate the stability and effectiveness of aggregation methods.

2. Improving programming skills and reading libraries, implementing other people's code. Development and implementation of code to perform tasks of data aggregation and analysis. Skill in evaluating and analyzing the resulting code and its work in the system.

3. Quite a bit - the ability to work with data. Create and manage databases of various sizes for conducting experiments.

4.Graphical presentation of data and presentation skills. Construct and interpret graphs to visualize results and identify anomalies in data. Prepare reports and presentations explaining research findings and proposing solutions to identified problems.

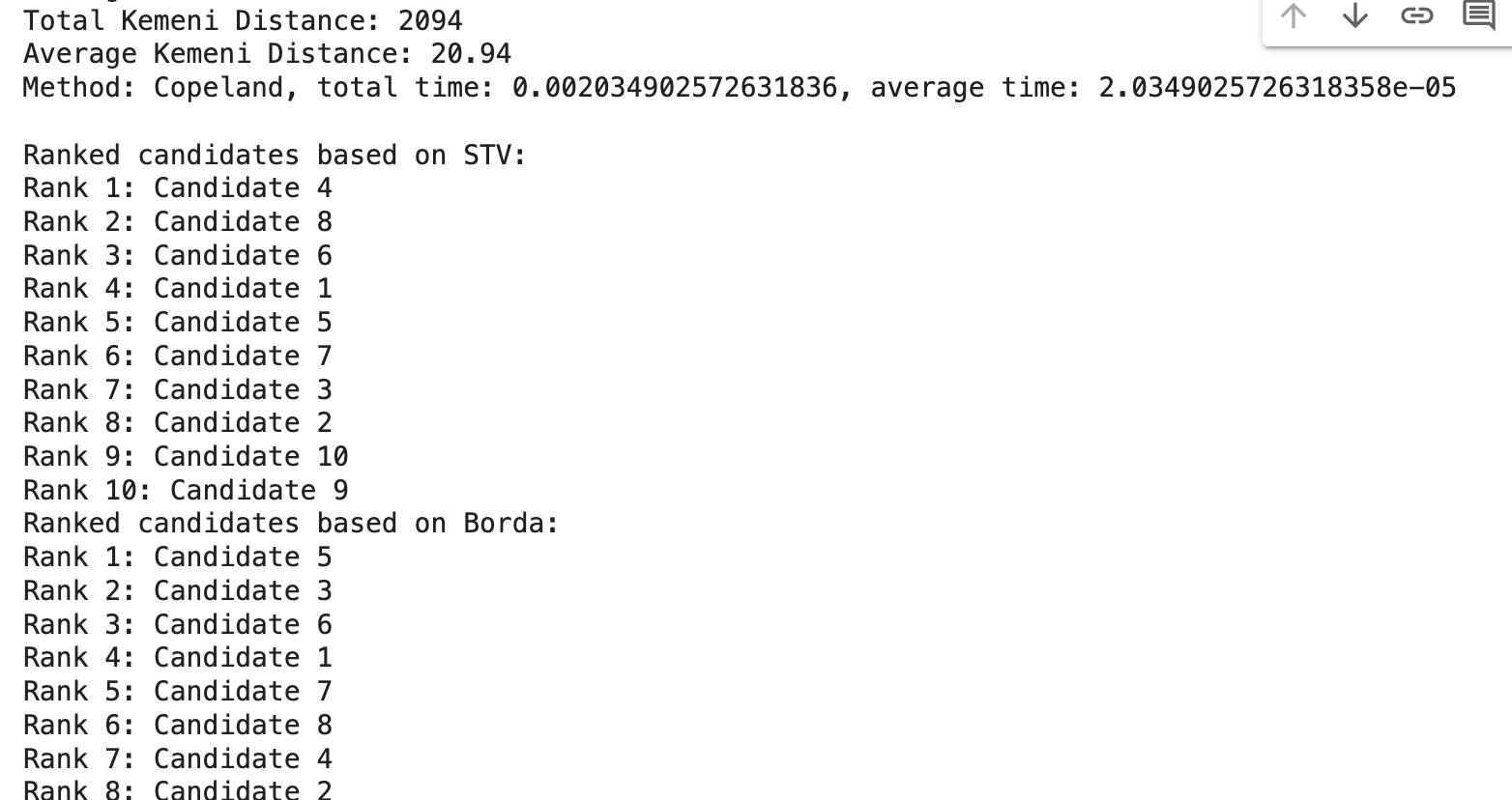
10. Project management and improving research skills

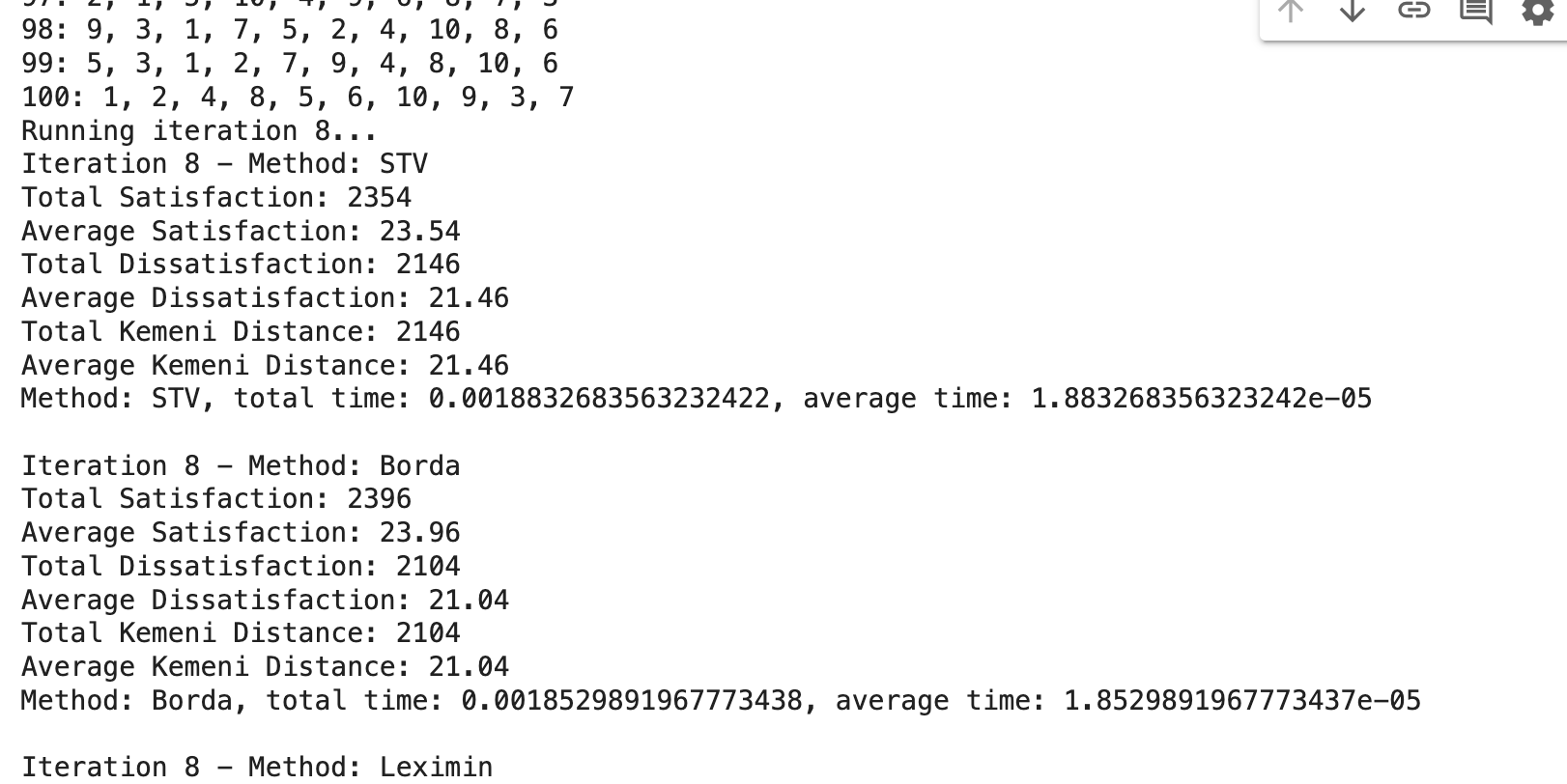
Planning and organizing the stages of work on the project, monitoring the completion of tasks and achieving goals within the established time frame. Conducting comparative studies and analyzing the results obtained to identify the effectiveness of various methods.

**Project's results**

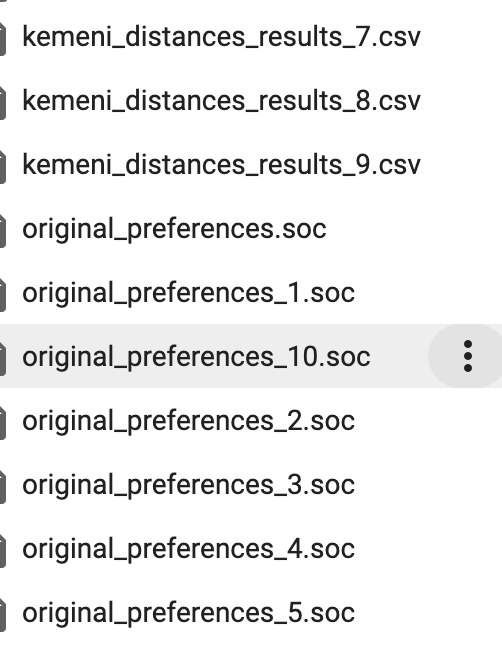
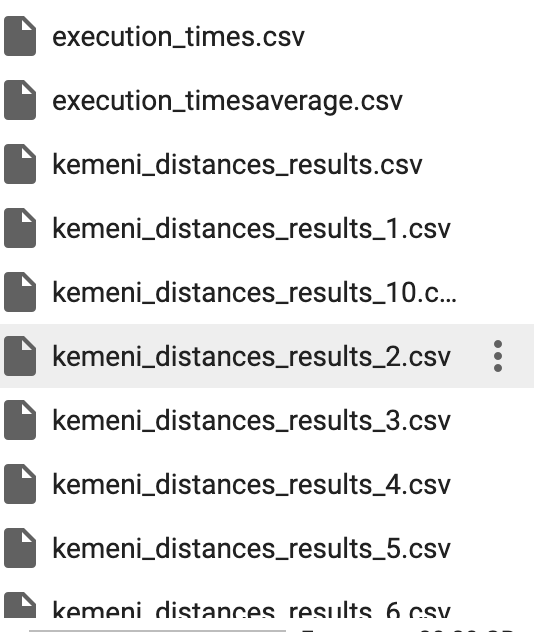
The presented results are expressed in the format of the presented results of the code (presented below in the form of screenshots), graphs (duplicated in the presentation), as well as code (attachment link to Google collab with the source file)

Visualization of the internal processes of the code. General calculation of distance by whom name For various integrations, calculation of distance kameni time for specific interactions of specific aggregation methods, visualization of rows of individual preferences, dates generated for applying functions.

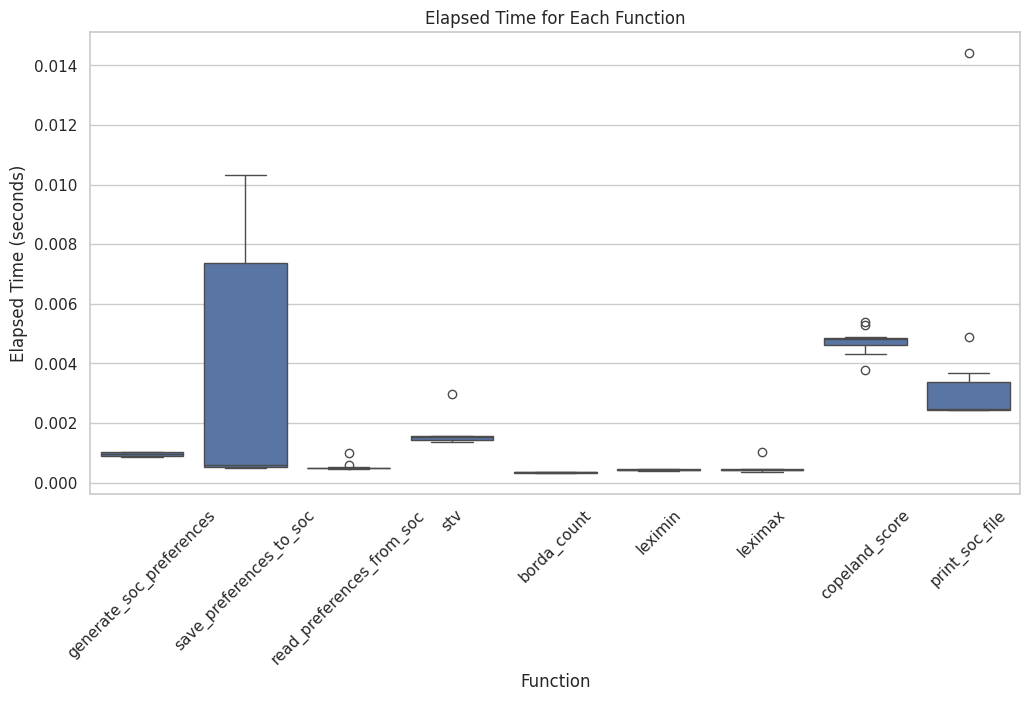


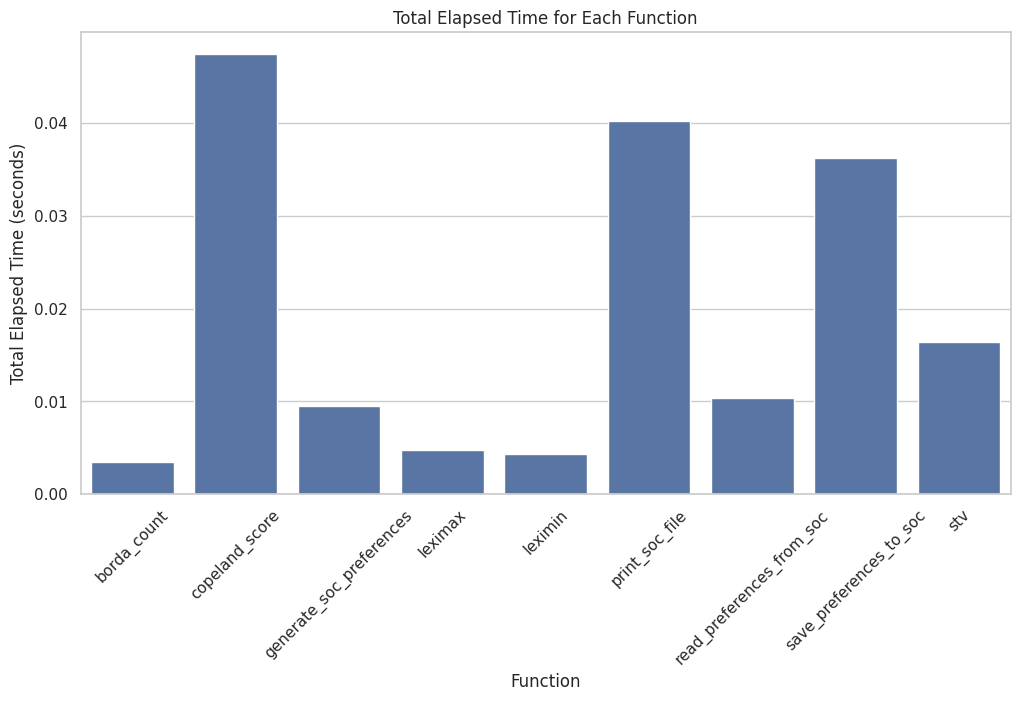


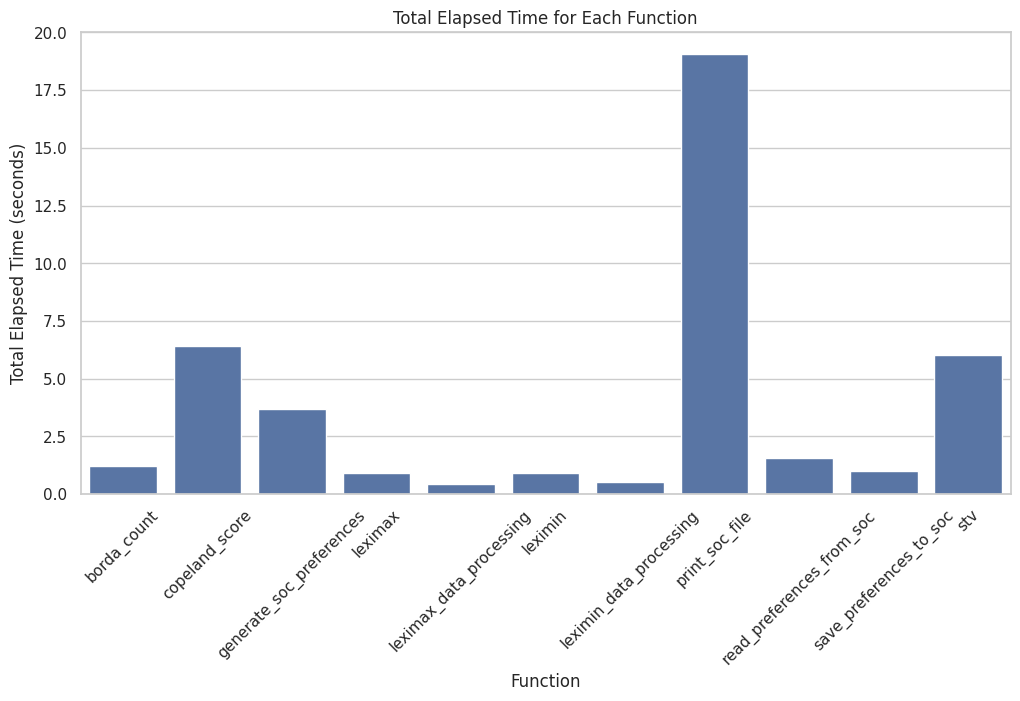
Screenshot with a partial list of files in which the received data is located (if it is necessary to download and use it)

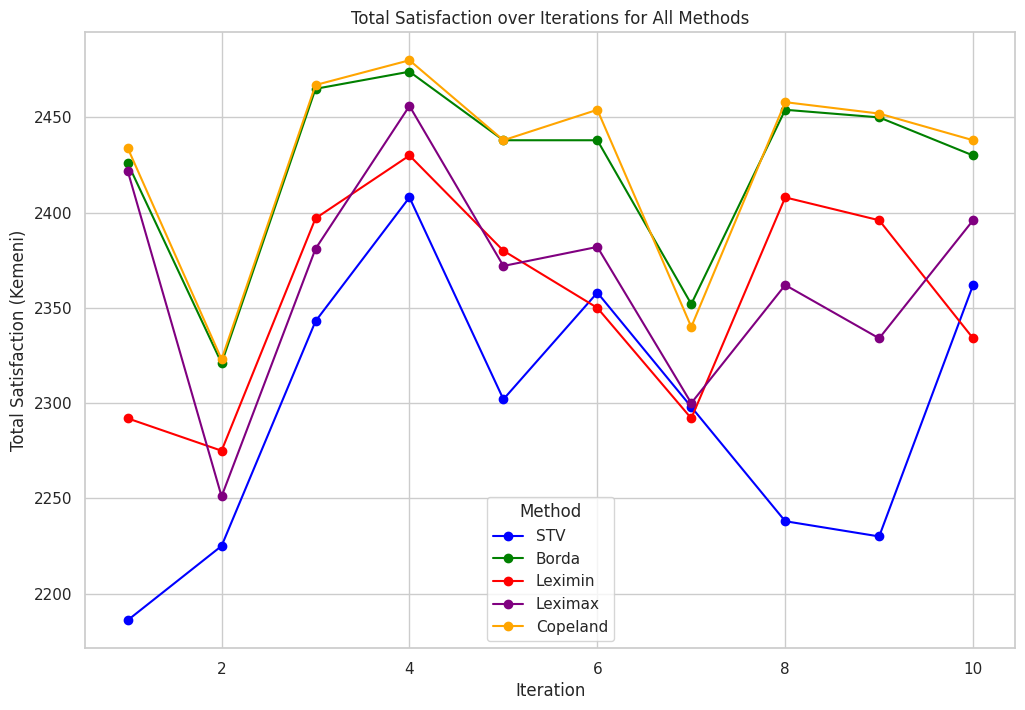


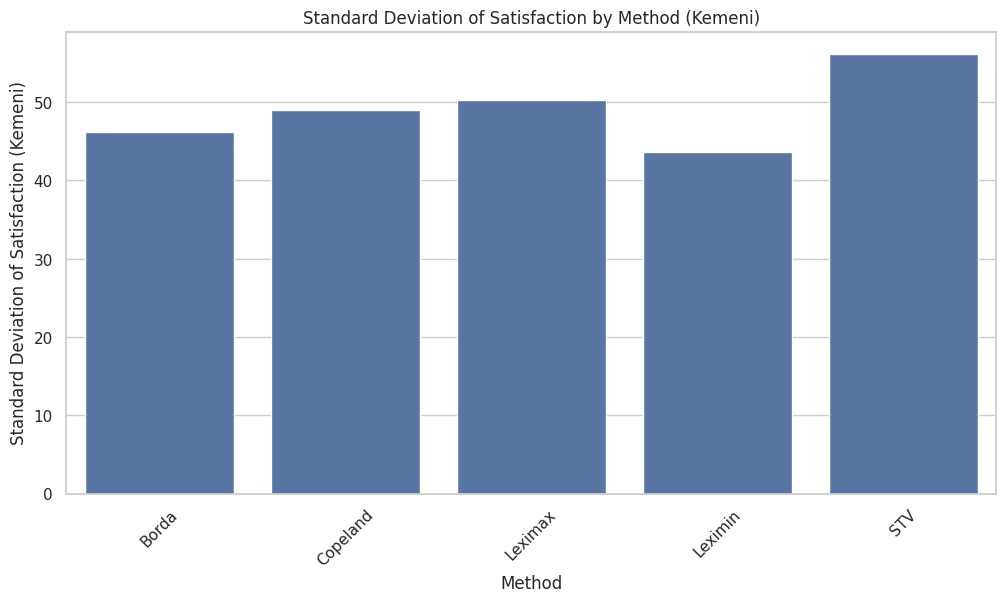
Visualization of the results obtained

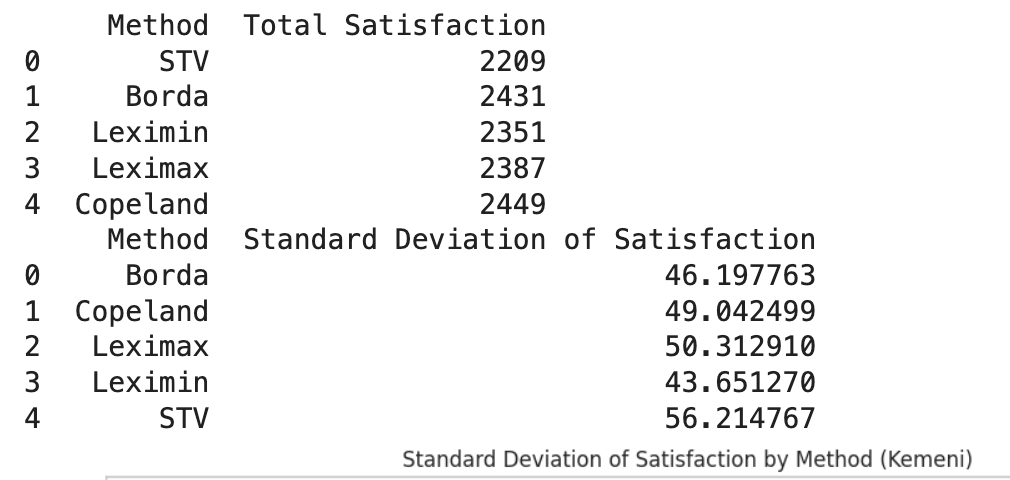
* 

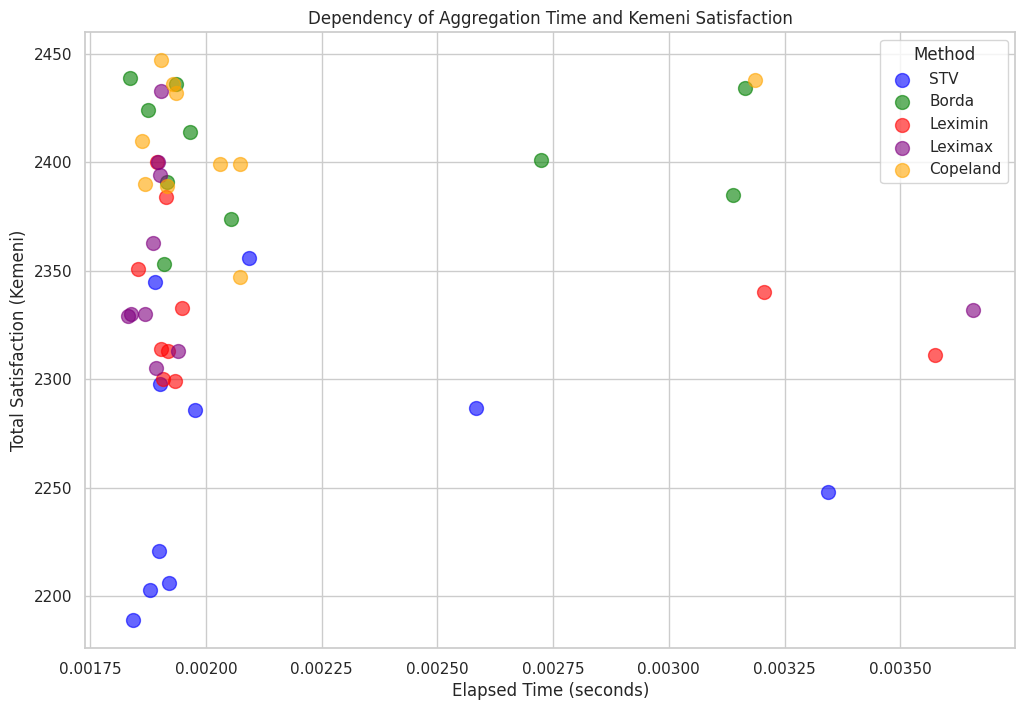


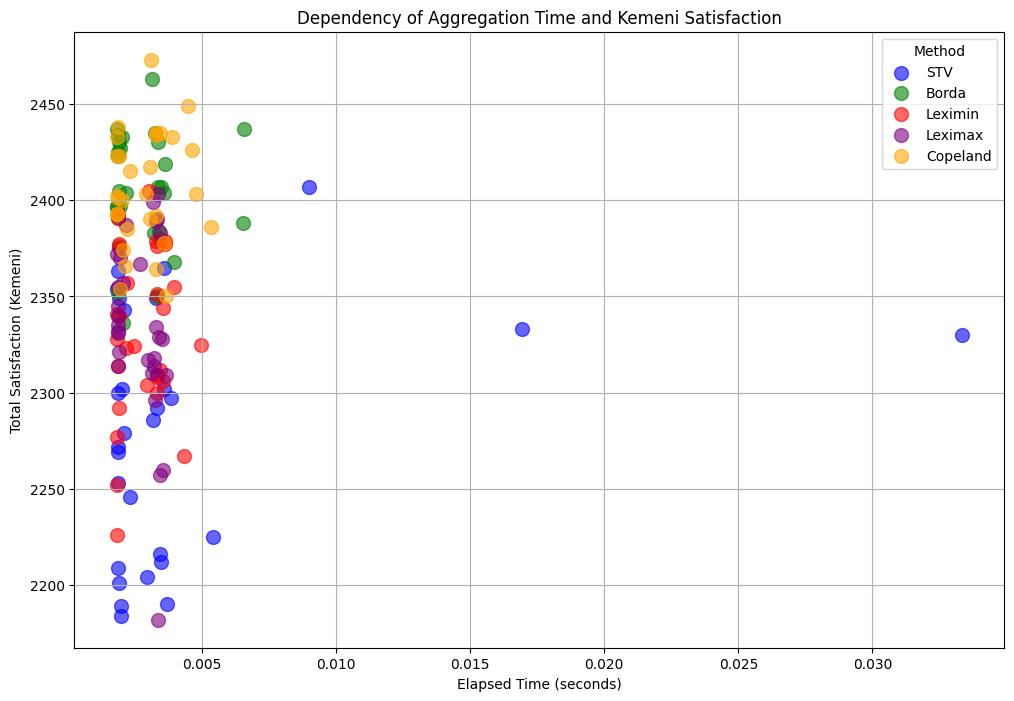


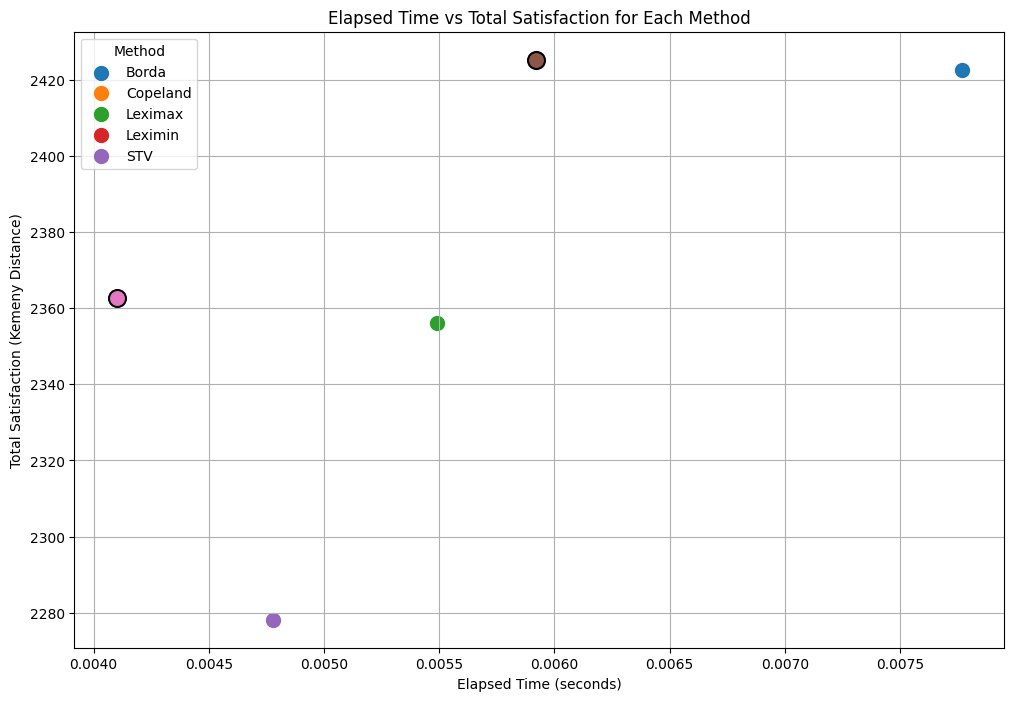


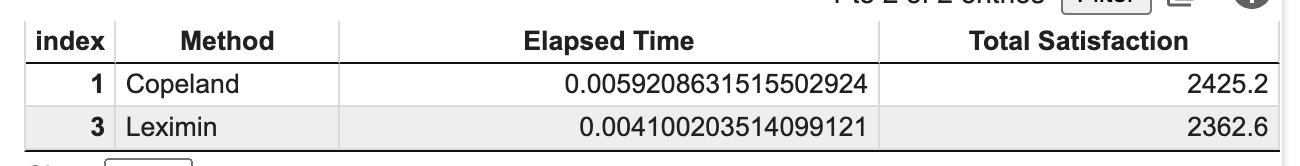


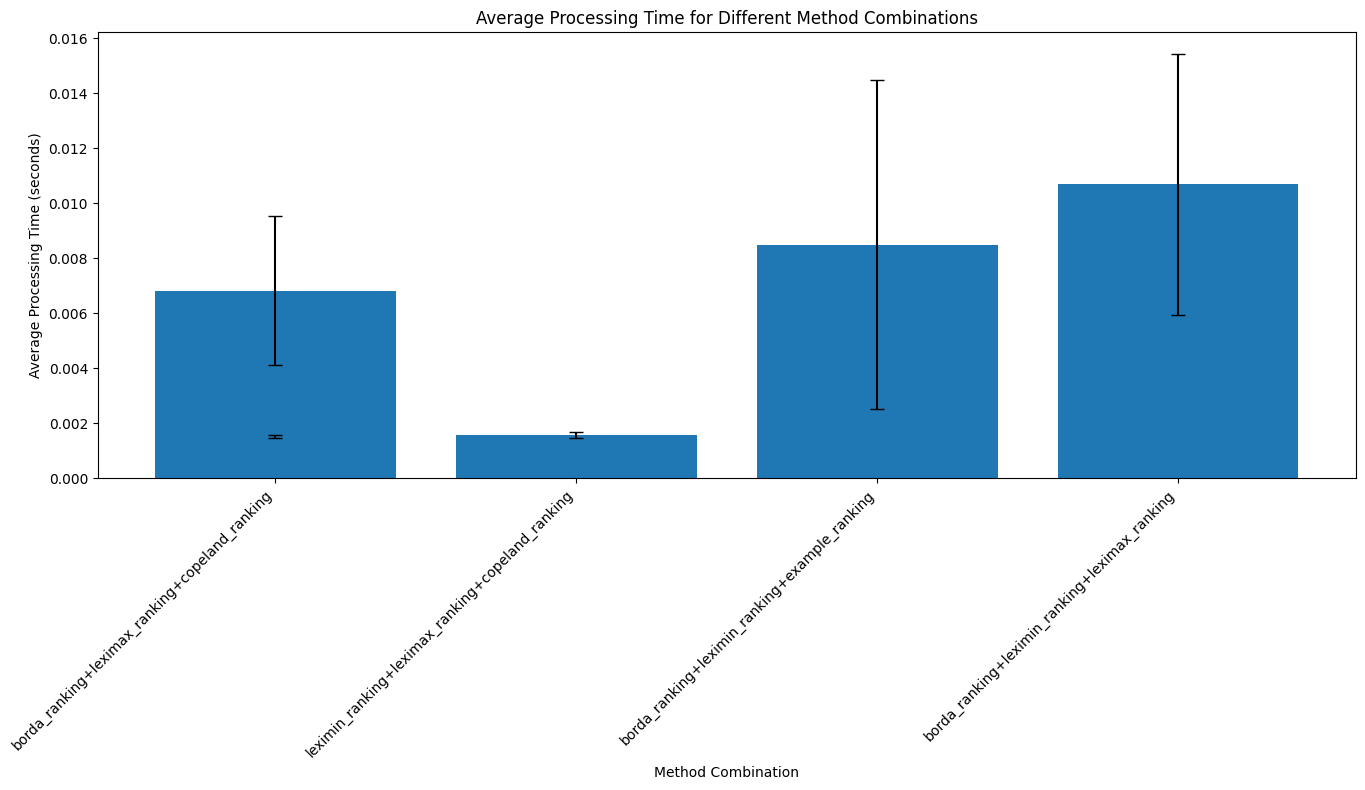


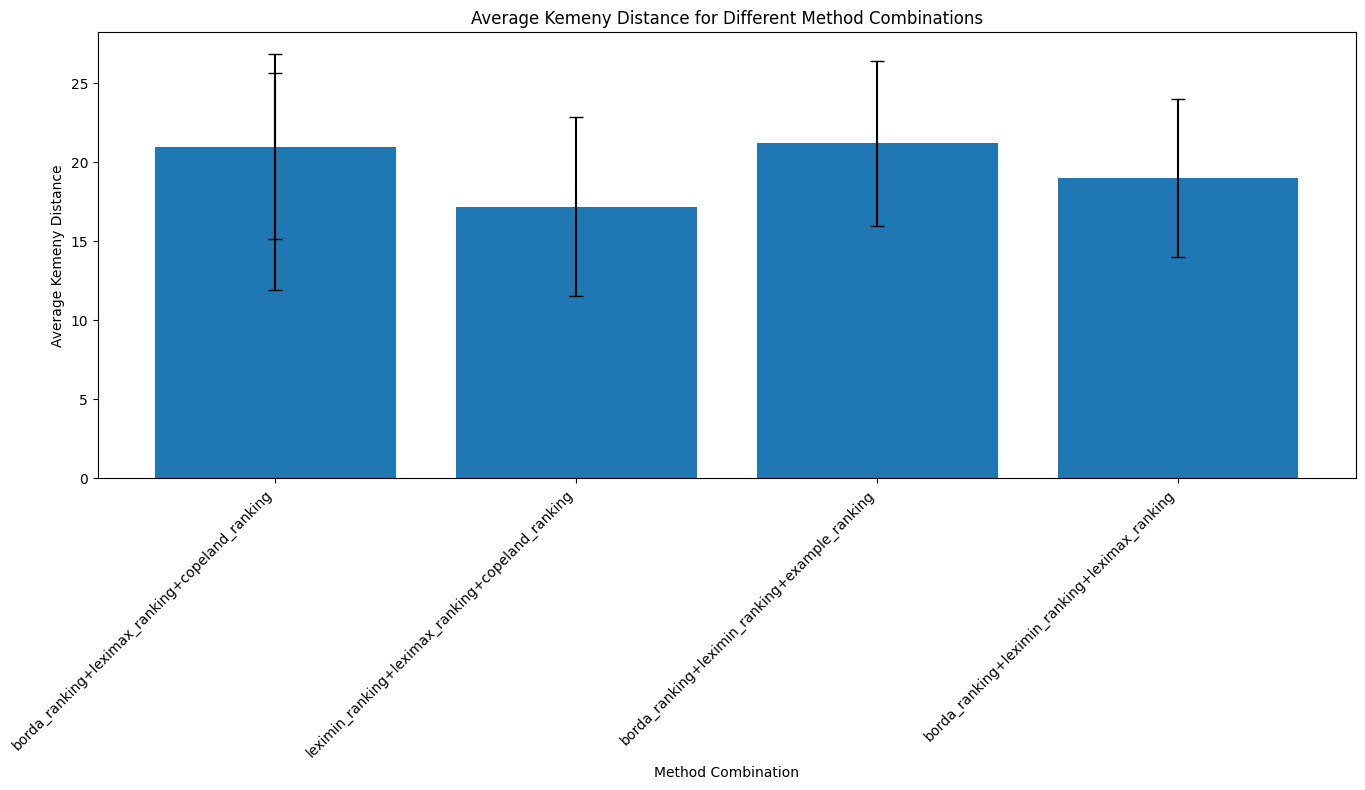


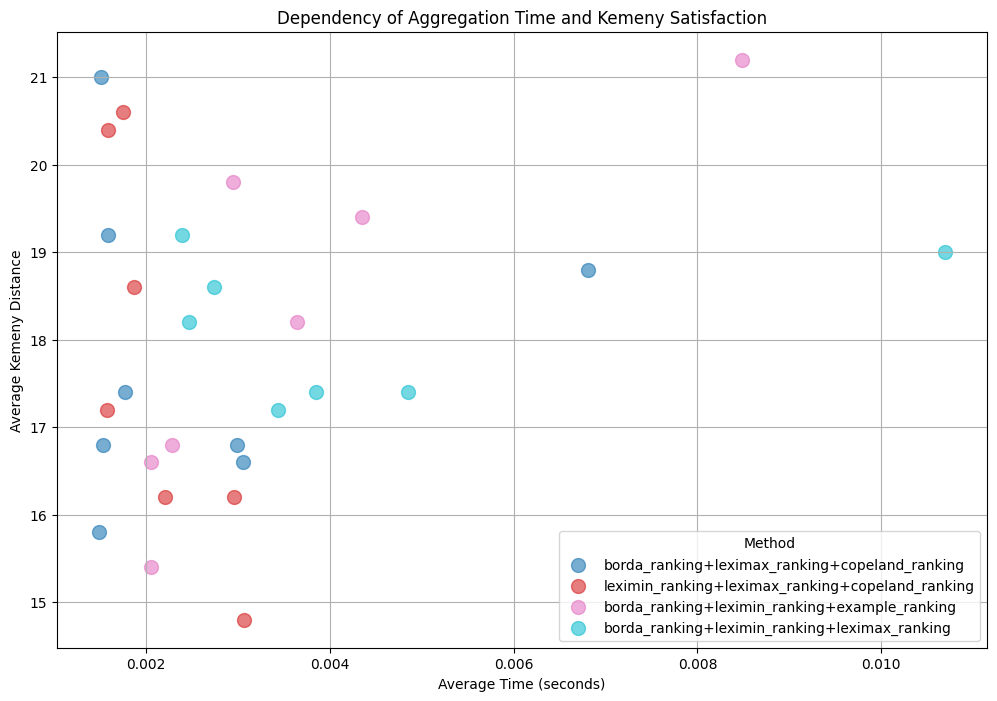


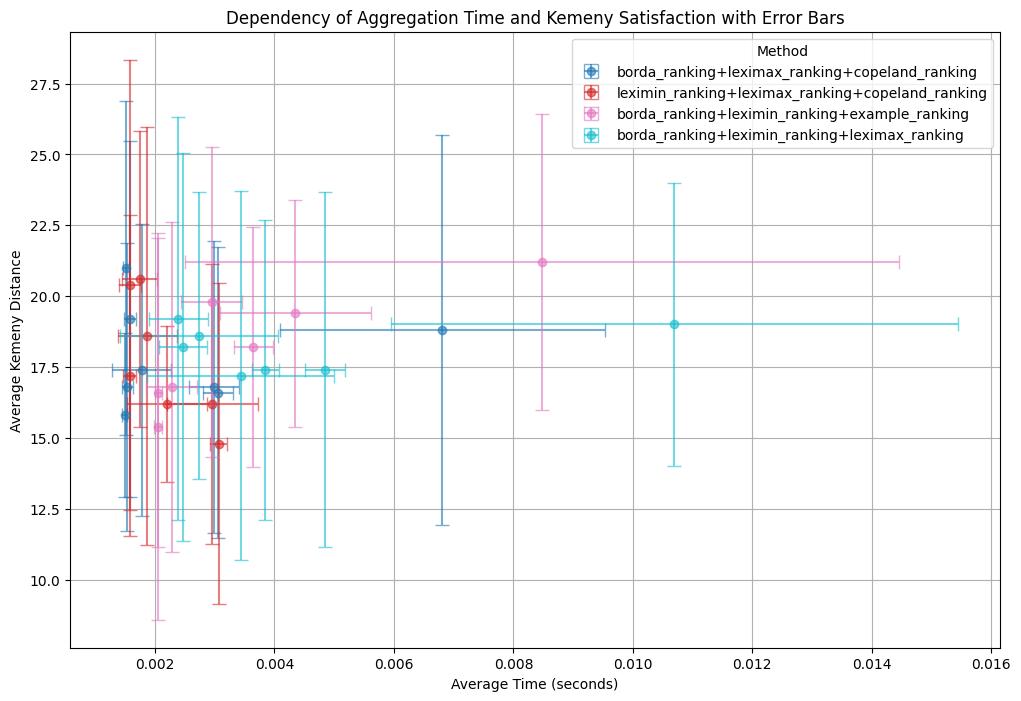


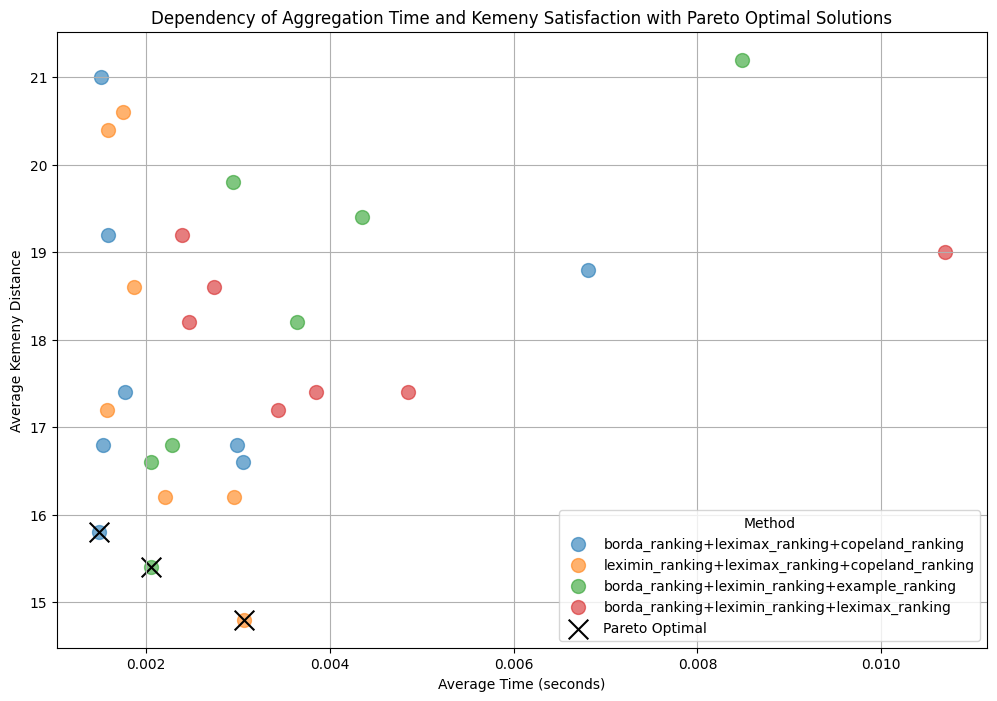












Appendix

| Link to main code containing main functions |
| --- |
| https://colab.research.google.com/drive/1z71flZf1cGxUCKwScss7QSA0aczNshZ1?usp=sharing |
| Link to supporting files confirming the gradual implementation of various functions in the code |
| https://colab.research.google.com/drive/115PSBeuMcaFajfFC3XC2bbPjxW02AEdS?usp=sharing |
| https://colab.research.google.com/drive/11RCNK7H\_8DLUqrLe7i1iOOS\_6tN2IKL1?usp=sharing |
|  |

1. https://github.com/PrefLib [↑](#footnote-ref-0)
2. https://github.com/erelsgl/PrefLib-Tools [↑](#footnote-ref-1)
3. https://github.com/erelsgl/PrefLib-Tools/blob/master/preflibtools/notebooks/MSS\_16\_Tutorial.ipynb [↑](#footnote-ref-2)
4. #https://github.com/mikedillion/PrefLib-Tools [↑](#footnote-ref-3)
5. https://github.getafreenode.com/django/django [↑](#footnote-ref-4)
6. https://hturner.github.io/PlackettLuce/reference/preflib.html [↑](#footnote-ref-5)
7. <https://github.com/logc/borda> [↑](#footnote-ref-6)
8. https://github.com/johnh865/election\_sim [↑](#footnote-ref-7)