Step 10: Putting it all together – design and perform an experiment

In this step you will design your own experiment. For this experiment, you will need to also acquire your own data. For instance, the analysis can be based on your favorite music. Examples are:

- Profiling similarities between different musicians.
- Comparing two musicians.
- Profiling the progression of the music of a musician.
- Analyzing similarities of specific songs to musicians.
- Analyzing similarities between albums.
- Unsupervised learning of musicians, songs or albums.
- Any other idea that you might come up with

These are obviously just a few examples.

Each project will be different, based on the different datasets and different goals. In general, each project should include either classification or regression. Statistical inference (e.g., p values) and analysis/visualization of the results (e.g., confusion matrix, graphs) are also required. Other tools can also be used, including outlier detection, clustering, etc.
Please choose the tools that best fit your needs. For instance, if you apply a regression analysis you do not need to use a confusion matrix to show your results. Instead, you need to show your P values and a graph that shows your data.

We covered a lot of tools during the semester. In case you are not sure how to use a tool that you need, feel free to contact me, preferably in class so that other students are also reminded.

Tools that you can use include tools we used in the course (Weka, Udat, scikit-learn, etc), but if you prefer to use other tools not discussed in the course you should feel free to do that.

After you finish the experiment, you will need to describe your work in a short research paper.

The deliveries of your project include:

- Paper (pdf, as well as Latex source code)
- Code
- Data files (preferably csv format, but that can change during the project)

For convenience, a different entry in Canvas for each delivery will be created, but only the paper entry will be gradable. All of these deliveries need to be submitted, unless agreed otherwise.

Music feature extraction (if you choose to use music data)

In the case you prefer to work on music, you will need to also complete the task of feature extraction. Feature extraction is a common step in real-world data science. You will use your own music files to create your experiment. If you do not own music files, you can use free services that convert YouTube videos to MP3 files. You can also download from other music repositories available on line.

To compute numerical image content descriptor from each music file, several publicly available tools can be used such as jAudio, Essentia, and more.

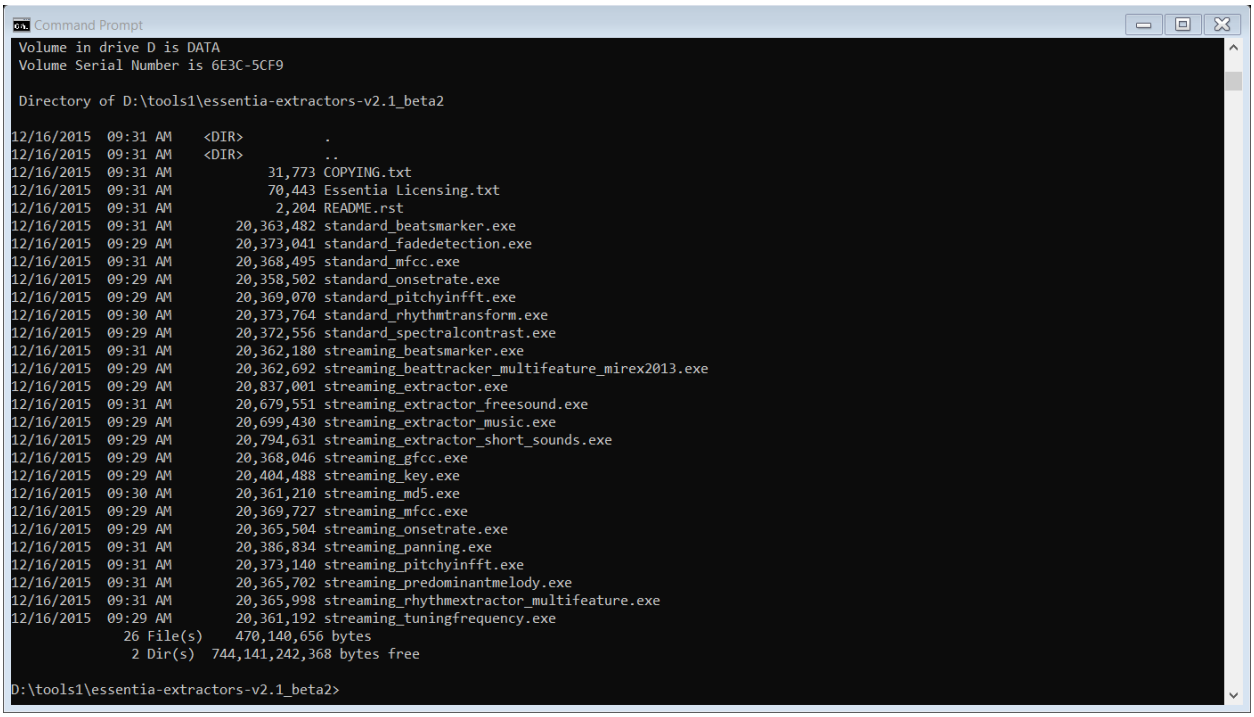
Perhaps the easiest tool to use is Essentia (https://essentia.upf.edu/).

Executables for common platforms are available at:
https://essentia.upf.edu/extractors/

To use Essentia, you need to download the executables of your platform and extract the files in your system.

Essentia can be used in various forms. One of the simpler forms is a command line tool that receives a music file as input, and generates as an output a json text file that contains the feature values.

If you use windows, extracting the file would provide a folder that looks like this:



To extract features from a music file, you can use the command “streaming_extractor_music.exe”.

```
streaming_extractor_music.exe  music_file.mp3  music_file.txt
```

That command will extract features from the file “music_file.mp3” and create the file “music_file.txt” with the values of the values and names of the features inside it. You can open the output file with a text editor to see its content. Obviously, the file “music_file.mp3” should exist, and should have a valid music content in it. You might have some warning in some of the files, but for our purpose these warnings can be ignored.

Since you have more than one file, you will need to write a script (in the programming language of your choice) that processes all your files automatically. The script will just run the streaming_extractor_music.exe for all of your music files.

After you created the json files of all music files, you need to convert them to a CSV file that you can open and analyze with tools such as Weka or Scikit. For that, you can use services provided by Essentia (https://essentia.upf.edu/FAQ.html), or you

can write your own script that reads the json files and convert them into CSV. Once you converted all files into a single CSV file with the features and classes, you can start analyzing your data.

You will see in the files some non-numerical values such as “major, “minor”, or notes. These values should be replaced with numerical values as we discussed in class. For instance, “major” can be replaced with “0”, and “minor” can be replaced with “1”. Notes (e.g., “C”, “C#”, “Bb”, etc) should also be replaced by numbers.

Like in real-world data science, the feature extraction is a fairly long process that requires substantial efforts.

Length of the paper: 2000-5000 words.

Writing the paper

Instructions for writing the paper will be provided separately, and will be discussed in class. The paper needs to be submitted in pdf format, but prepared in Latex. Latex source files should also be submitted.

The paper should summarize your project goals, describe and justify the experiments that you performed, software you developed, and present your results, your analysis of the results, and your conclusions. The paper should be between 2000 to 5000 words, and include figures, tables and bibliography as needed.

Please note that the paper should not describe what you **did** in your project, but it should describe what you **discovered**, or what you produced that is of value to the potential reader of your paper. Therefore, the paper is not a chronological description of everything that you did, but a summary of the highlights and the important parts of your project. However, the paper should provide sufficient details that allow a reader to replicate the results.

The paper should include several sections, which are the abstract, introduction, methods, results, conclusions, and bibliography.

Abstract

The abstract is a single paragraph of 150-300 words, which summarizes your project, and can be read independently of the rest of the paper. The abstract should define the goals of the research project, the methods (very briefly) and then the results and conclusions.

Introduction

The introduction section should include the following:

a. Description of the problem that is being addressed, and an explanation of why the problem is difficult and why it is important to solve it.

b. A summary of previous work that was done for solving the problem at hand. This should include references to papers published previously that described solutions to the same or similar problems. These papers should also be listed in the *Bibliography* section.

Methods

This section should describe in details the methods that you developed and/or used. A method can be a certain existing software tool that you used, a formula, an algorithm, etc'. If you used existing methods, a reference to the paper, book, or web site of the method that you used should be specified.

Additionally, the sections should also describe the data that you used. Important details include the source of the data, the number of files, average size of files, file format, the typical content of the files, dates in which the data was created, etc.

Results

Here you present the results of your research, which can be the performance of your methods or the findings of your research. You can describe your results in words, but can also use tables and graphs when needed.

Conclusion

This section is the analysis of the results presented in the *Results* section, and should describe the conclusion of your work. It can describe, for instance, under which condition the method works better, or if the results support the hypothesis. It can also include a discussion about the usability and advantages of the method and propose ideas about related future work.

Acknowledgments

List here any person who assisted you in your work. Receiving help from others is basically allowed in this course, but please consult with me before asking for someone else's help.

References

The bibliography should list sources (papers, books, web sites, etc') of previous work related to your project, or work that was done by others and used by you in your project. For most references, specify the name(s) of the authors, the title of the work, and the source (book, journal, web site, conference) in which it was published. Ideally, the references should be prepared in bibtex, which is the most convenient way to handle references.

Figures and tables

In most cases you will want to include figures and tables in your report. Each figure and table should be numbered, and should be referenced in the text. Also, each figure or table should have a caption that briefly describes the figure or the table. Figures and tables should be placed in the paper where they are discussed, not in the end of the paper. Latex normally assigns figures and tables with numbers automatically.

Paper grading policy:

Clearly explaining and expressing the goals of the research	20%
Include all required sections of the paper	5%
Results are comprehensive and explained clearly. Detailed results are included in the paper, including	20%

statistical analysis when needed.	
References	5%
Conclusions are supported by the results	10%
Paper has the required figures and/or tables (e.g., confusion matrix)	10%
The paper contains all details required to replicate the results	20%
The paper contains just what the reader needs to know about the experiment. I.e., the paper is not about what you did, but about what the reader needs to know.	10%