The Dataset.

For the dataset we retrieved information from the National library of Netherlands. We found a newspaper catalog which included information of the name of the newspaper, the production year, the year the production stopped and the number of copies that existed. In order for the dataset to work better with what we wanted the algorithm to do, we removed the end-of-production year and the number of copies and replaced them with an identification number column and an author column, with the last one including made up information.

We thought that the information was easy to work with. And each member could work with it and understand it without problems, because we did not need to understand what the name of the articles was, just what the names was.