Why do we need to do all of this?

Firstly, it turns out that many MEPs have very inconsistently reported the attendees of their meetings. For example, a firm may have been reported in Hungarian, Italian, English or with an abbreviation. This leads to differences in the spellings of each name, even making some names unrecognisable. This makes it difficult for us to uncover patterns behind MEP meetings. Therefore, we have to standardise the names of each attendee if we want to properly analyse their lobbying activity.

Secondly, it is quite difficult to automate the classification process. Thus, we have to do most of it manually if we want to have a more meaningful analysis.

Steps for the classification of attendees in the files

1. Split the complete data into the largest files, unnest tokens, group by attendees and summarise.

2. Download file as xls.x into data folder

3. Create 4 columns next to the attendees column:
   - Fixed_names: standardised/fixed spelling of attendees
   - Class: NGO, business, government, academia or other
   - Structure: single actor or umbrella group
   - Comment: remarks for groupmates or future notes

4. Copy the attendees entrees into the fixed_names column.
   Before you can begin correcting the spellings of the entrees, read the following notes:

   - **DO NOT** change entrees in the attendees column. We need this column to stay intact if we want to successfully merge the correct spellings into the larger data later on. You can standardise the spelling by changing the entrees in the fixed_name column.
   - Some entries are very similar. Standardising their spelling works best if you replace the varying names in the fixed_name column by copy+pasting the most suitable name into the other rows.
   - It is preferable to have attendee names that include a transparency register number, if listed.
   - Be aware that some attendees have different names in different languages. Thus, you will have to do some research to discover which organisation they are part of. The same goes for attendee entrees that solely consist of a number.
   - The process works the fastest if you also fill in the other columns (class, structure, comment) as you fix the spelling of each attendee.

5. In order to overcome language differences and discover the class/structure of attendees, you will need to use a variety of tools. Here's a list with some suggestions:
   - Google translate.
   - Transparency Register: [Zoeken in het register - Europese Unie (europa.eu)](#)
   - Lobbyfacts: [Search | lobbyfacts](#)
   - Websites of the attendees.
   - ChatGPT or AI tools can be useful if you are really unsure about the class or structure of a given attendee. However, make sure it identifies the same attendee as the one in the data.

Below is an example of what a finished table may look like:

	A	B	C	D	E
1	attendees	fixed_names	class	structure	comment
2	Amnesty International Limited	Amnesty International Limited	NGO	Single_actor	
3	APEAL	APEAL - Transparency Register Number: 7003690	Business	Umbrella_group	
4	Apeal	APEAL - Transparency Register Number: 7003690	Business	Umbrella_group	
5	APEAL - Transparency Register Number: 70036906501-10	APEAL - Transparency Register Number: 7003690	Business	Umbrella_group	
6	Human Rights Watch	Human Rights Watch	NGO	Single_actor	
7	International Commission of Jurists	International Commission of Jurists	NGO	Single_actor	
8	McDonald's	McDonald's	Business	Single_actor	
9	US Ambassador	US Ambassador	Government	Single_actor	
10	Vinces Consulting	Vinces Consulting	Other	Single_actor	Consultant
11					

6. **Do not forget** to repeatedly save the file. Also be sure to give the modified version of the file a different name, otherwise your progress will be lost if you re-run the R code.

7. After the table has been filled in, we can then import it back into R and join it with our file data.