**Matching Twitter handles of Companies and Brands to ThreadAuthors and Authors on Twitter**

The objective of the task is to create a match between the author list and company/brand Twitter handles. We are tracking Twitter conversations about a particular topic (e.g., #goodfood) and have collected all the mentions of this hashtag in the last 5 years. We would like to find out if companies or brands have started a conversation about #goodfood or have responded to such conversation started by other users. On Twitter, the authors who starts a conversation is called a Threadauthor while the author who responds to the conversation is called an Author.

Thus, we have collected mentions by the two types of authors:

1. **Threadauthors** (366.590 of them): these are the authors who started a particular Twitter thread.
2. **Authors** (1.3 million of them): these are the authors who responded to a thread.

We need to match these **threadauthors** and **authors** to two types of entities that have Twitter handles:

1. **Companies** (453 of them): Publicly listed companies (e.g., Coca Cola) that have their own Twitter handle
2. **Brands** (1016 of them): These are brands (e.g., Fanta) that are owned by the 453 publicly listed companies and also might have their own Twitter handle.

We have gathered all the Twitter handles of 453 companies and 1016 brands.

Thus, for the task, we have 2 types of authors (threadauthor/author) and 2 types of entities (company/brand). As the author file was delivered in Excel (which handles up to 1 million rows) it is split into two parts author\_1 and author\_2. Accordingly, we end up with 3 files for authors (threadauthors, authors\_1, authors\_2) and two files for entities (companies, brands).

Each author file has to be matched **separately** with the company and brand file. Thus, for convenience, the following file names have been created:

1. “threadauthors\_company.xlsx”,
2. “threadauthors\_brand.xlsx”.
3. “authors\_1\_company.xlsx”
4. “authors\_2\_company.xlsx”.
5. “authors\_1\_brand.xlsx”
6. “authors\_2\_brand.xlsx”.

In each of these files in column A you can find the author names (either threadauthor or the author).

In addition, in each of these files you will find either the **cik** (Column B) and **conm** (Column C) in the **\_company** files OR **ugov\_id** (Column B) and **brand\_name\_tax** (Column C) in the \_brand files. You also will find the Twitterhandle and Twitterhandle2.

These are the fields that need to be filled with information from the following files:

1. “company\_twitter.xlsx” (453 companies)
2. “brand\_twitter.xlsx”.

As an illustration, consider the following example for the steps:

1. Open the “threadauthors\_company.xlsx” file (for example)
2. Open the “company\_twitter.xlsx”.
3. Take a Twitter handle of a company (e.g., from company Adobe take @Adobe and @AdobeNews)

* Note that sometimes companies can have more than 1 Twitter handle.
* When looking for a Twitter handle, please disregard @ as this character is not contained in the authors and threadauthors files. Easiest is to create a new column(s) without @.

1. Search for “Adobe” and the “AdobeNews” in the “threadauthors\_company.xlsx” file.
2. If a match if found -> copy the cik and conm and the Twitterhandles in the “threadauthors\_company.xlsx” file
3. If not match is found -> proceed to the next company.

**For step 5**: Ideally you would find an efficient algorithm that automatically searches for the exact Twitter handles but also allows for less than exact match.

* Please use an automated algorithm that search for the exact Twitter handle but also looks around the words. One idea is to use smth like fuzzywuzzy, or similar fuzzy string matching in python, to generate similarity scores, and then pick the highest scoring items as matches
* Report a match score in column D of the files.

To illustrate how this task should be done we have created an example. In the file, “threadauthors\_company.xlsx” we have found Adobe matches. Similalry we did so in the “threadauthors\_brand.xlsx” and reported the ugov\_id and brand\_name\_tax.

Note that we also matched AdobeCreate and AdobeXD which are NOT exact matches. We did this manually, but an automatic approach is preferred.

Please also see an illustration in the “authors\_1\_brand.xlsx” and “authors\_1\_company.xlsx” files in which we matched the brand Citibank and the company Citigroup INC with the Twitter handles @Citibank and @CITI, respectively with Author “Citgroup” “Citibank” and are not exact matches. This is the **challenging** part.