**Business problem**

Create a model that predicts if a review of a restaurant is positive or negative.

**Explaining the model**

1. Import the dataset, which is in a tsv format. Regarding text analysis, it is important to have a file with tab separated values, as we might have some commas on the values of the dataset, and it makes it hard for the machine learning model to analyze it. The first argument is the file, the second allows us to ignore the quotes (putting it blank means the algorithm will ignore them) and the third will prevent the model from identifying the reviews as factors (because we want to analyze the content of our reviews).
2. Now, we need to create some independent variables, which will be some words that we can identify as “positive” or “negative” (each word will have a single column). But first, we need to clean the text.
3. We start by creating a corpus with the text that we want to clean (the review column). Then, we put all the reviews on lower case, remove numbers, punctuations, irrelevant words, stem the document (transform the words to its ‘roots’ – loved to love, for example) and we also remove the white spaces that were created by applying the previous steps (to optimize our document).
4. Now, we create our table with one column per each word. The function DocumentTermMatrix creates a matrix with all the features. Then, we filter the non-frequent words (we chose to keep 99.9% of the most frequent words).
5. After all this, we apply a classification model (randomForest, for example).