# Using duplicated\_info\_finder.py to determine if two words should be the same node in the graph

This script is used to treat the output of the Machine Learning model designed to extract information from an article. Since this output may not be consistent between the different extracted results, we need a way to know if two different words have the same meaning so we don’t create duplicated nodes.

First, we transform all word separations to spaces and transform everything to lowercase using the all\_to\_spaces function. We do this to not differentiate between results such as “pipe-break” and “pipe break”. Then, we transform all multiple spaces into single spaces to not differentiate between results like “243 cases” and “243 cases” with the remove\_double\_spaces function. Then we transform all numbers in the results to numbers so we can affirm that “4 deaths” and “four deaths” are the same using the transform\_to\_numbers function, which makes use of the w2n script published on the Internet. Finally, we want to transform complex words to simpler synonyms, so we don’t consider different words that mean the same, such as “Escherichia Coli” and “E. coli” or “ecoli”. Right now, this is implemented using a csv file and the transform\_similar\_words function, but since this file is prone to getting very heavy using a lot of words, we will probably be working with a machine learning model or a database.

When all transformations have been applied, we calculate the Levenshtein distance between the two words and, if this distance is under a certain number, we can affirm that the words mean the same, such as “pipe break” and “pipe broke”.

Here we see an example comparing “pipe-broke four deaths” and “Pipe break 4 deaths”.

# Comparing how similar are emergencies from the API to documents in the graph

For now, the from\_db\_to\_emergency and the compare functions in emergency\_comparator.py allow us to fetch documents from the graph by their title and compare them to see how similar they are with a score from 0 to 100.

The emergencies are compared by their sources, causes, locations, contaminants, and symptoms.

The sources and the causes use the functions defined in the previous section to compare how similar they are. They are penalized if they have a different number, because an emergency with 2 causes is different than one with 5 causes.

The locations are compared by range and by the locations themselves, boosting those that have the same type of affected area, and then incrementing the score if the affected areas are the same. For example, two emergencies that affected a whole country would be more similar than one that affected a whole country and another that only affected a street, and two emergencies in Spain have more in common than one in Spain and one in France.

Finally, contaminants are compared to the other contaminants by the same means of the sources and causes, but the symptoms are only compared between similar contaminants.