**Project Design for Data Science Project:**

**Natural Language Inference Model**

Name: Limor Nunu

Email: [NunuLimor@gmail.com](mailto:NunuLimor@gmail.com)

Project goal:

Creating Natural Language Inference model for determining how pairs of English sentences are related (neutral connection / contradiction / entailment).

Example:

Sentence A: “The kids are frowning”

Sentence B: “Children are smiling and waving at the camera”

Nature: contradiction

In my project I’ll follow this process:

1. Identify the question / problem I want to solve
2. Collect data
3. Explore the data
4. Prepare and process the data
5. Model building
6. Evaluate the results
7. Consider whether there’s a need to go back to earlier stage if not, continue to 8
8. Deployment

So let’s elaborate on each stage:

1. The problem I want to solve or the goal of my project is determining how pairs of English sentences are related.

To do so I’ll create Natural Language Inference model.

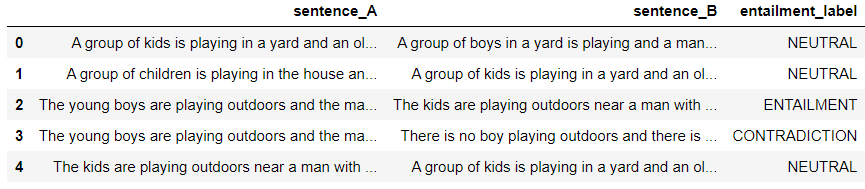
1. For this task I use dataset from the following source: <https://zenodo.org/record/2787612#.X76d0mgzZPY>
2. In the data exploration stage I’ll look for missing values, weird values and take care of them if needed.

Moreover, I’ll look at the distribution of the data.

1. Preparing the data for using it in a model – models don’t understand text but they understand numbers. So, I’ll create vectors for the model use (embedding).
2. Building the model using Python.
3. Check the accuracy metric and see how the model performs.
4. Investigate what led to the results and decide whether to continue or to gather more data/reprocess the data/fix the model.
5. After I’ll be satisfied by the results – it will be the time to deploy the model!

If the weights of the model are small, I’ll be able to upload the project to Github and deploy it using Heroku and Flask. Otherwise, it will run locally, because I don’t have the resources to run it on cloud for example.

Dataset structure:



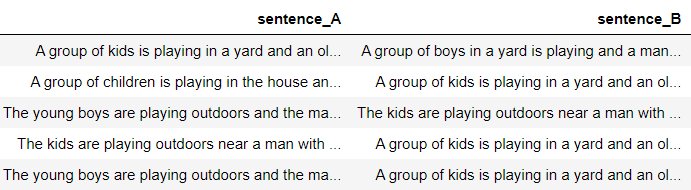
The dataset contains two sentences and a label (the type of connection between them).

I’ll split it into three groups: test, train and trial (validation).

Running the model:

Everyone who wants to run the model will need to install the requirement file (requirements.txt) in their Python environment.

Afterwards, they will need to upload CSV file in that structure:



Two sentences without their label.

The output:

The same table with another column with the label.

Extending the project:

I’ll try to extend the output to represent the connection between key words in the sentences, which determine the nature of the sentences.

For example, I’ll show that the word ‘smiling’ has the opposite meaning of the word ‘frowning’, and ‘kids’ has similar meaning to ‘children’.