**DS Team Home Project**

**The problem:**

Here at ------, data arrives in the form of invoices. One of the first questions we face in the data’s flow is whether or not the document at hand is relevant for tax purposes. Even though that’s usually the case, our domain requires us to be thorough, and it’s essential to filter out cases where the document is not valid.

In this project, you’ll be asked to tackle the issue yourself, using the (randomly sampled) partial data you’ve been provided. For privacy reasons the original text won’t be available to you, only features that have been extracted from it.

**The scope:**

The whole project should take 4-6 hours. The idea is for you to “80/20” it - do the first 80% of the work that takes 20% of the time. Our goal is to see how you approach a DS problem - not to see how far you can get in limited time.

We’d like to see a full “data science pipeline”, so make sure you include all your work in the script and add as many explanatory comments as needed. If there is anything different/more that you’d do if you had a week or a month - write it down along with a short explanation.

**The data:**

The pickle file you’ve received contains 4 objects, 2 project datasets and 2 holdout datasets. The holdout data is meant to emulate the future, and therefore does not contain the label. Here’s a description for each object:

* *project\_df*: a pandas DataFrame with data for the training task, including a **rel\_doc** column which holds the labels.
* *project\_tf\_idf\_mat:* a CSR sparse matrix with tf-idf scores for the training data. Rows in this matrix correspond to rows in the *project\_df* DataFrame (i.e. the nth row in *project\_df* corresponds to the same invoice as the nth row in *project\_tf\_idf\_mat*).
* *holdout\_df:* a pandas DataFrame for invoices needing prediction. The columns here are the same as the ones in the *project\_df* DataFrame - except for **rel\_doc**, which is what you’ll be predicting.
* *holdout\_tf\_idf\_mat:* a CSR sparse matrix with tf-idf scores for the to-be-predicted data. Rows in this matrix correspond to rows in the *holdout\_df* DataFrame (i.e. the nth row in *holdout\_df* corresponds to the same invoice as the nth row in *holdout\_tf\_idf\_mat*).

**The goal:**

The goals of the project are:

1. Building a model that predicts the **rel\_doc** field.
2. Providing predictions for the attached holdout dataset.
3. Provide estimates for the model’s performance on the holdout data (or any future data). It’s up to you to choose which metric/metrics are most relevant.

**The output:**

Please submit your results as a zipped folder including the following:

* A .py file with the project’s script.
* A pandas DataFrame saved as a parquet file, with predictions for the holdout set and their corresponding datapoint\_id.
* A pkl file holding the model used for those predictions.
* A word file containing estimates for the model’s performance on future data, along with any notes/comments that came up during the project.

**Good Luck!**