V2024-04-23

Text Mining – 2nd Semester 2023/2024

Project Assignment Handout

This handout details the rules of the first practical project delivery for the Text Mining class, to be handed in as per delivery rules below (delivery date is 23h59 of the last 31st of May).

Table of Contents

1. Project Summary

The goal of this project is to use Natural Language Processing (NLP) models to predict whether a property listed on Airbnb will be unlisted in the next quarter[1]. For this, you will resort to real Airbnb property descriptions, Airbnb host descriptions, and comments from previous guests.

In summary, with the NLP techniques you will learn during the Text Mining course, you must implement an NLP classification model able to predict, for each property, if it was unlisted (1) or is still listed (0).

The project should be developed using Python and libraries such as NLTK, Scikit Learn, Keras or PyTorch. Also, the project can be solved in various ways, which means there is no exact correct solution. And, of course, groups should not use code from each other!

[1] http://insideairbnb.com/

2. Group Rules

The project should be done in a **group of two to four (2-4) students**; we consider 2-4 students to be ideal size, but allow individually sized groups, to facilitate things.

Should you constitute groups of size other than 2 to 4 members without explicit permission from the teachers, you will be subjected to the following penalties:

- For each member in excess, or below, the stated size: cumulative penalty of 1,5 points (out of 20 total points) in final project grade.

The only allowed exceptions that will be made regarding group size will be for situations beyond the control of the group members (for example, one of the students dropping out of the course) and will be evaluated on a case-by-case basis.

We will provide a Moodle group section and a Project Forum that will allow you to check and engage with different class members, to help you find a group; however, this is the sole responsibility of each student, and the only formal group that counts is the one listed in the cover page of your Report, when you submit your project.

Note, also, that members “gone missing” from Group (or not contributing to desired level), are not the teacher’s problem; as such, this will not be an acceptable excuse for deadline extension requests or for asking for a project grade review/allowance.

3. Project Starting Point – The Corpora

The corpora for the project is divided in following sets:

- Train (train.xlsx) (6,248 lines): Contains the Airbnb and host descriptions ("description" and "host_about" columns), as well as the information regarding the property listing status ("unlisted" column). A property is considered unlisted (1) if it got removed from the quarterly Airbnb list and it is considered listed (0) if it remains on that same list.
- Train Reviews (train_reviews.xlsx) (361,281 lines): This file has all the guests' comments made to each Airbnb property. Note that there can be more than one comment per property, not all properties have comments, and comments can appear in many languages.
- Test (test.xlsx) (695 lines): The structure of this dataset is the same as the train set, except that it does not contain the "unlisted" column. The teaching team is keeping this information secret! You are expected to provide the predicted status (0 or 1) for each Airbnb in this set. Once the projects are delivered, we will compare your predictions with the actual (true) labels.
- Test Reviews (test_reviews.xlsx) (41,866 lines): The structure of this dataset is the same as the train reviews set, but the comments correspond to the properties present on the test set.

Note that you are not required to use all the textual fields from the provided corpora as input for your model. For example, your best solution may just use the "description" column as input.

4. Expected Deliverables

In terms of the solutions developed, you must deliver:

- One **.pynb file (notebook)**, named NLP_XX (XX stands for the group number), containing the techniques you experimented and, by the end of the notebook, your ready-to-run final solution.
- A **.csv file**, named "Predictions_XX", with the column "id" from the test set and your predicted labels for the test set in a column named "predicted".

Additionally, you must submit a **PDF report** named "Report_XX", documenting your work, as outlined in section 6 "Report Structure".

5. Detailed Evaluation Criteria

Your solution should present the following points:

- Data Exploration (1.5 points): Here you should analyze the corpora and provide some conclusions and visual information (bar charts, word clouds, etc.) that contextualize the data.
- Data Preprocessing (2 points): You must apply a method to split your training corpus into train/validation sets to evaluate the performance of your model (you can also resort to K-Fold cross validation, or other methods). Moreover, you must correctly implement and experiment at least four (4) of the data preprocessing techniques shown in class (stop words, regular expressions, lemmatization, stemming, etc.). You should apply more data exploration after the preprocessing.
- Feature Engineering (5 points): You must correctly implement and experiment with two (2) of the feature engineering techniques seen in class (TF-IDF, GloVe embeddings, etc.).
- Classification Models (4.5 points): You must correctly implement and test three (3) of the classification algorithms seen in class (KNN, LR, MLP, LSTM, etc.).
- Evaluation (1.5 points): You must evaluate your models resorting, at least, to Recall, Precision, Accuracy and F1-Score.

Moreover, the development of extra work (more techniques than the minimum required in the previous points and/or techniques not shown in class) is highly recommended and will account for a maximum of 4.5 points divided as follows:

- Data Preprocessing – 0.25 points for each extra method (unseen in class) used (maximum of 2 extra methods).
- Feature Engineering – 1 point for each extra method using Transformed-based embeddings (maximum of 2 extra methods)
- Classification Models – 1 point for each extra model using Transformers or other advanced models (maximum of 2 extra methods).

6. Report Structure

As already mentioned in previous sections – and as repeatedly pointed out in the Lectures and Lab classes – your Project Report should ONLY address the specific project decisions, work and results made by your Group; we don't need – nor do we want – the Report to include generic and largely "theoretical" text about Text Mining or any of the subjects covered in this semester.

In other words, the text that should appear in your Report is text that ONLY your Group could have written – because it is specifically about your unique context and situation. Do not place "filler" text (general opinions and considerations) in your project: just get right to the point and tell us what is really important, which is "**what did you do and why did you do it that way?**".

The structure of the project report should follow, as much as possible, this one:

i. **Data Exploration** – data presentation and explanation of the main findings from the exploratory analysis (accounts for 50% of criteria 4.1).
ii. **Data Preprocessing** – explanation of the different preprocessing methods developed (accounts for 25% of criteria 4.2).
iii. **Feature Engineering** – description and discussion of the methods implemented including parameters used (accounts for 25% of criteria 4.3)
iv. **Classification Models** – description and discussion of the models implemented including parameters used (accounts for 25% of criteria 4.4)
v. **Evaluation and Results** – description of the performance of the models and main conclusions (accounts for 50% of criteria 4.5)

The PDF report should have a maximum of 20 pages (without the cover, index and references pages, but the annex will count for the 20 pages) describing the previous points. Exceeding this number will incur a 0.5-point penalty for each extra page.

Any extra work developed must be clearly defined as such in the PDF report, or else it will not be considered for evaluation as extra work! You should add, in each section, a paragraph pointing out the extra work developed.

7. Delivery Guide

The delivery of the project will be done through a Moodle project submission section. All files should be saved in a folder named “Group_XX”. This folder (zip it if you need) must be submitted through the project submission section, until 23h:59 of the 31st of May (Friday).

Detailed instructions on the Project Delivery will be provided within the Moodle closer to the submission date (and this handout may be updated, if deemed necessary).

Failure to deliver on time will incur a 0.5 point penalty for each late day (for example, 4 late days will accrue a 2.0 point penalty).

Failure to comply with the delivery guide (no proper naming of objects, duplicated or unclear files, improper folder configuration, etc.) will meet with a 0.5 point penalty.

8. Project Competition

We will compare your predictions with the actual Label from the test set (“test.csv”). The three (3) groups with the highest performance in F1-Score will receive points as follows:

- 1 point for the group with the best model
- 0.75 points for the group with the 2nd best model
- 0.5 points for the group with the 3rd best model

Questions and Clarifications

Should it be necessary, we will provide further clarifications and answers to questions from students made through the Project and General Help forum in Moodle.

Good luck with your project!

END OF HANDOUT