Group Project
COMP9417 Machine Learning and Data Mining
T1, 2020

Introduction

This is a group project that will be done by a team of 3-5 students and the aim is to apply machine learning techniques to predict some specific outputs in a dataset.

The first step is to go to the course Moodle page and in the Moodle>Homework & Assignment>Group_Project - Member_Selection create your groups.

Group project contributes to 30% of the total mark (**30 marks**). The deadline to submit your report is **Monday 27 April, 5:00 pm**.

Submission will be via the Moodle page.

Recall the guidance regarding plagiarism in the course introduction: this applies to this report as well and if evidence of plagiarism is detected it may result in penalties ranging from loss of marks to suspension.

Objective

In this project, you will use text classification to predict the most relevant news articles for each of the 10 topics. The practical application of this project can be described as follows:

There are 10 users who like to read news articles. Each user, however, is interested in a different, single topic. The topics are:

1. *ARTS CULTURE ENTERTAINMENT*
2. *BIOGRAPHIES PERSONALITIES PEOPLE*
3. *DEFENCE*
4. *DOMESTIC MARKETS*
5. *FOREX MARKETS（外汇市场）*
6. *HEALTH*
7. *MONEY MARKETS*
8. *SCIENCE AND TECHNOLOGY*
9. *SHARE LISTINGS*
10. *SPORTS*

*As a Data Scientist, you are tasked to help these users find the most interesting articles according to their preferred topics. You have a **training dataset** containing about 9500 news articles, each assigned to one of the above topics. In addition, (as in real life situation) the dataset contains about **48% of irrelevant articles** (marked as* IRRELEVANT*) that do not belong to any of the topics; hence the users are not interested in them. The distribution of articles over topics is not uniform. There are some topics with large number of articles, and some with very small number.*

*One day, 500 new articles have been published. This is your **test set** that has similar article distribution over topics to the training set. **Your task is to suggest up to 10 of the most relevant articles from this set of 500 to each user**. The number of suggestions is limited to 10, because, presumably, the users do not want to read more suggestions. It is possible, however, that some topics within this test set have less than 10 articles. You also do not want to suggest 10 articles if they are unlikely to be relevant, because you are concerned that the users may get discouraged and stop using your application altogether. Therefore you need to take a balanced approach, paying attention to not suggesting too many articles for rare topics.*

Datasets

You have two datasets: **training.csv** and **test.csv**, each with a header: "article_number, article_words, topic". Text pre-processing can be very complex, therefore for this project each article has already been pre-processed and contains all article words separated by comma (","). Some words occur in the same article more than once. Each distinct word should be treated as a feature.
"article_number" should only be used for reference in the report, and **not to be used as a feature**.

Project implementation

Each group has to implement a minimum of two classification methods and select the best method for the **final task**, which is **suggesting up to 10 news articles per topic from the test set**. The suggestions should be based on a score for each article, obtained from your method. You are free to **select the features and tune the methods** for the best performance. However, you are allowed to do this **using only the training set**, **without even looking at the test set before the final test** (except the test set size). For this purpose, you may want to use cross-validation for your **feature selection, method selection and tuning**. You can choose your method for classification even if the method has not been covered in the course. You can use any open-source library you need for your implementation.
You also free to **choose any performance metric or metrics** that are best for your method selection task, based on the data analysis, e.g. topic distribution. However, **you need to justify your choice** in the report. Dividing the training set into training and development sets (e.g. 9000 instances in training and 500 instances in the development set) can also be useful to estimate the class distribution in the final test set.

You can use all the provided features or a subset of features; however you are expected to give a justification for your choice. You may run some exploratory analysis or some feature selection techniques to select your features. There is no restriction on how you choose your features as long as you can justify it. In your justification of selecting methods, parameters and features you may refer to published results of similar experiments.

Report

Each group has to submit one report which contains introduction, dataset exploratory analysis, methods and evaluation, results, discussion and conclusion. The report is expected to be 10-12 pages (with single column, 1.5 line spacing) and **easy to read**. The body of the report should only contain main presentation, e.g. tables, charts, with the rest of them (if required) included in appendices.

Here are guidelines for the report:

- **Title page**: title of the project, name of the group and group members
- **Introduction**: a brief explanation of the problem, the aim of the project, the main issues that need to be addressed and intended methods for those issues.
- **Exploratory data analysis**: class distribution, feature statistics and anything else that can help to solve the problem.
- **Methodology**: A detailed explanation and justification 调整 of methods developed, method selection, feature selection, hyper-parameter tuning, evaluation metrics, design choices, etc. State which method has been selected for the final test and its hyper-parameters, including the number of features to be used.
- **Results**: The following results are required:
  - Cross-validation results on the training set for selected metrics, feature sets and implemented methods.
  - Final results for each class calculated on the whole test set using the final selected method with its hyper-parameters.

    These results should be in the following format (in alphabetical order of topics):

Topic name	Precision	Recall	F1
ARTS CULTURE ENTERTAINMENT			
BIOGRAPHIES PERSONALITIES PEOPLE			
DEFENCE			
DOMESTIC MARKETS			
FOREX MARKETS			
HEALTH			
MONEY MARKETS			
SCIENCE AND TECHNOLOGY			
SHARE LISTINGS			
SPORTS			

  - Final article recommendations using the final selected method with its hyper-parameters. In addition, please calculate Precision, Recall and F1 for these recommendations.

    These results should be in the following format (in alphabetical order of topics and in numerical order of article numbers):

Topic name	Suggested articles	Precision	Recall	F1
ARTS CULTURE ENTERTAINMENT	e.g. 9511, 9512, 9513			
BIOGRAPHIES PERSONALITIES PEOPLE	e.g. 9501, 9502, 9503, 9504, 9505, 9506, 9507, 9508, 9509, 9510			
DEFENCE				
DOMESTIC MARKETS				
FOREX MARKETS				
HEALTH				
MONEY MARKETS				
SCIENCE AND TECHNOLOGY				
SHARE LISTINGS				
SPORTS				

Note: to calculate the metrics in the above two tables, you can use actual classes from the test set.

- **Discussion**:
  - Compare different methods, their features and their performance. State any general observations.
  - Discuss the metrics in the above two tables. Which metric(s) is/are more appropriate and why?
  - If you continue this project, how would you improve it, e.g. using of other methods and parameters that have a potential to be useful but not tried yet?
- **Conclusion:** Give a brief summary of the project and the findings, what have you discovered and learned from this project (if anything).
- **Reference:** list of all literature that you have used in your project if any.

Code submission

Code files should be submitted **as a separate file** along with the report.

Peer review

Individual contribution to the project will be assessed through a peer-review process which will be announced later, after the reports are submitted. This will be used to scale marks based on contribution.

Anyone who does not complete the peer review by the **Thursday of Week 12 (7 May)** will be deemed to have not contributed to the assignment. Peer review is a confidential process and group members are not allowed to disclose their review to their peers. 机密的

Project help

Consult Python package online documentation for using methods, metrics and scores.
There are many other resources on the Internet and in literature related to text classification.
Some introductory reading can be found in:

https://towardsdatascience.com/text-classification-in-python-dd95d264c802

"Introduction to Machine Learning with Python: A Guide for Data Scientists",
Book by Andreas C. Müller and Sarah Guido, chapter on text classification.

When using these resources, please keep in mind **the guidance regarding plagiarism in the course introduction.**

General questions regarding group project should be posted in the Group project forum in the course Moodle page.