Natural Language

MP2

Question Classification

Alameda and Tagus

2021

With this project we will simulate an evaluation forum. As you should know by now, international evaluation forums are competitions in which participants test their systems in specific tasks and in the same conditions. Training/development sets are given in advance, and, on a certain predefined date, a test set is released. Then, participants have a short period of time to return the output of their systems, which is evaluated and straightfowardly compared with one another, resulting in a final ranking where the state-of-the-art system is acknowledged.

Tasks

(1) Build (at least) two models, in python, that classify questions according to the following labels:

GEOGRAPHY
MUSIC
LITERATURE
HISTORY
SCIENCE

As an example, given a file with a list of questions and respective answers (a simplified/altered version of the Jeopardy! dataset from Kaggle):

Question	Answer
"Any device that turns one kind of energy into another; a microphone is an example of one"	Transducer
Phoenix lies on a river named for this substance found in the name of another state capital	Salt
"Many consider his 1952 book ""Invisible Man"" the greatest post-war novel about black life in the U.S."	Ralph Ellison

Your system should return the “type” of the question:

SCIENCE
GEOGRAPHY
LITERATURE

(2) write a short scientific report of your work.

Your models

To build your models, you can use (alone or combined):
a) a rule-based approach (although by itself results will probably be very poor)
b) a model based on similarity/distances
c) a model based on n-grams (here, I mean a language model built with n-grams)
d) a machine learning algorithm such as Naïve Bayes, Support Vector Machines, etc. (**deep learning architectures, neural word embeddings or any kind of neural pre-trained models are not allowed**).

Details

Groups:
This project should be done in groups of 1 or 2. If you are looking for a colleague to create a group, please add your contact here.

Questions:
As usual, questions should be sent to meic-ln@disciplinas.tecnico.ulisboa.pt (subject: MP2). However, we might release FAQs about the project. Please, check them.

Input/output format:
As in an evaluation forum, you will be given a "training" set (train.txt) and a "development set" (dev.txt) in which each line has the format:

label question answer

The test set will be released after the release of your code and report. Each line will have the format:

question answer

In all the cases there will be a tab between the label/question/answer.

You should run your best model and return a file in which each line has the format

label

Notice that the line number in which the pair question answer appears in the test file should be the same line number of the corresponding label in the output file (the evaluation depends on this).

Testing your system:
In order to evaluate your system, we will run the following command:

python qc.py –test NAMEOFTESTFILE –train NAMEOFTHETRAINFILE > results.txt

Language and libraries:
You should implement your model in Python 3. You can take advantage of code already available (and I strongly advise you to do so), as long as you identify the source. You can use the following

libraries/software: NLTK, Spacy, NumPy and scikit-learn. If you really want to use another library, ask first, please.

Evaluation

(1) Report (in Portuguese or English) (10 points):

The report should be named NUM.pdf (NUM is the number of the group). It should have a maximum of **2 pages** containing the following sections (**mandatory**):

1. Group ID: The number of the group, and the number and name of each group member should be written the first two lines (you only have 2 pages, so please, don't spend one with a hysterical title page)
2. Models: a clear description of your models, including all the pre-processing done (if applicable)
3. Experimental Setup: a clear description of how you have developed your project (which evaluation measures were used, what was the baseline, etc.)
4. Results: in this section you should present results (general and by label), in a table, that should include:
   4.1. the baseline results
   4.2. your models' results, considering the mentioned evaluation metrics
   4.3. accuracy resulting from evaluating your final models in the development set
5. Error analysis: here, I expect you to show me that you have properly analysed the obtained results. So, try to explain the most common errors (examples are more than welcomed!)
6. Future work: if more time was given to you, explain what you would do to improve your system
7. Bibliography (if applicable)

How your report will be evaluated:

- 2 points for the general quality of your report (correct syntax, clearness, no typos, etc.)
- 2 points for the creativity of your approach
- Sections 2, 3, 4: 2 points
- Section 5: 3 points
- Section 6: 1 point

(2) Automatic Evaluation (10 points):

- 5 points will be given by evaluating the dev_results.txt file (Part 1, see below)
- 5 points will be given by evaluating the test_results.txt file (Part 2, see below)

How your results will be evaluated:

- A*ccuracy* will be the evaluation measure.
- If you beat a weak baseline (Jaccard + stemming) that results in an accuracy of 68% you will have 2.5 points
- If you beat a stronger baseline, based on a Support Vector Classifier and a CountVectorizer that results in an accuracy of 80% you will have 5 points.

Notice that:

- 3 values will be taken if any instruction is not followed, including the ones regarding the report.
- If the report has more than 2 pages, we will only evaluate the first two, even if the first one is a cover page with your numbers and names.

- We will randomly select a set of projects and we will run them in the dev and test sets. If any difference in results is found, the group will have a 0 in the project.

Submission

Part 1 – on November 10th, 2021, **before 1:PM (13h)**, you should deliver, via Fénix (MP2-Part1), a zip file (NOT a rar) containing the project, named after the group number **NUM** (ex: 3.zip).

- the zip file should contain:
  - the file **NUM.pdf** with the report (details above)
  - a file named **qc.py** with the project code
  - possible extra python files
  - a file named **dev_results.txt** with the results from the given development set, that is, a list of the labels returned by your **best model** when it was applied to the given development set (see below).

The test set, test.txt, will be released between 1:30PM (13h30) and 1:35PM (13h35).

Part 2 – on the same day, November 10th, **between 1:35 PM (13h35) and 11:59 PM (23h59)**, you should deliver, also via Fénix (MP2-Part2), a file named **test_results.txt** with the results from the given test set, that is, a list of the labels returned by your previously submitted **best model** when it was applied to the given test set (see below). ****No manual editions should be done to this file after running your model; no changes in your code should be done or in the obtained output****.

Comments/tips:

- This is not a B.Sc project; this is a M.Sc project: there is a clearly identified problem that you need to solve in the best possible way, but we do not tell you how to do it.
- Remember what you have learned during the class about methodology: try to do a systematic work. Evaluate your models every time you (try to) improve them.
- Pre-processing applied to the training set should also be applied to the dev/test set.
- When evaluating your project on the test set, you can add the dev set to the train set and use this bigger file (nice!). **HOWEVER, DO NOT ADD THE DEV SET TO THE TRAIN SET WHEN YOU ARE EVALUATING ON THE DEV SET AND DO NOT ADD THE TEST SET TO THE TRAINING SET WHEN YOU ARE EVALUATING ON THE TEST SET!!!**
- Attention to blindly removing stop words: your questions can be empty at the end.
- Understand that language is too complex to deal with each example individually; also remember that your model will need to be able to generalize.
- There is no 100% accuracy (this is a research problem).
- This is a "real" dataset. Datasets in NL have errors and are usually unbalanced (welcome to NLP!). Some categories might be poorly represented (or even nonexistent) (welcome to NLP!). You will also probably find many labels that you don't agree with (welcome to NLP!). You are probably right, but the dataset will not be changed. Write about these situations in your report.
- Look at results!!
- A kit-kat will be offered to the winning team.

Thank you! I really hope you enjoy the project and have a good learning experience with it! ♥