Data Processing Advanced: Group Assignment

Last updated: 2018-11-14

For this assignment you will participate, in groups of 3 students, in a competition. You will need to submit your solution to a text mining problem as well as a report describing your solution and the code which you used to generate it.

This assignment is worth 45% of your course grade. The assignment grade will be based on the quality of your work as reflected in the report and code (50%), as well as your ranking in the competition (50%).

Your report should be **2 pages maximum**, and should include the following:

- Description of your experimental setup, including:
  - preprocessing if any
  - representations you used for the data points
  - details of your method: for example choices of similarity/distance metric and how you tuned it
  - discussion of the performance of your solution
- Detailed specification of the work done by group members
- The name of the account under which you submitted your results to the competition on Codalab (see below).

You will also need to submit your Python code and the instructions on how to run it. Your code can be either a plain Python script, or an IPython notebook. Include all the code and data necessary to re-run your experiments. Put the report in **PDF format**, the code and the data in a single zip file named with your group number, e.g. Group17.zip, and submit it to the BlackBoard assignment.

In addition you will need to submit your solution file the competition server. The competition is hosted on https://competitions.codalab.org. One member of the team will need to get a codalab account, and will be responsible for submitting your solution. Indicate the name of this account in your report. See section **Submission to Codalab** for additional details.

IN SUMMARY: submission consists of two parts

1. Zip file with your report and your code (BLACKBOARD)
2. Submission of your solution (CODALAB)

Group work

Your report needs to contain a detailed description of who did what, so make sure to keep track of this information.

Note: it is **not acceptable** to just say *All members worked together and contributed equally.*

If there are any problems with collaboration, such as serious disagreements, a group member not contributing, or a group dissolving, make sure inform the course coordinator as soon as possible via email.

Code reuse rules

Remember this assignment is group work. You are **not allowed** to collaborate or share code with students outside your group. **Submissions will be checked for plagiarism.**

If you are found breaking the above rules you will be reported to the Board of Examiners for fraud.

You are, however, allowed to use code examples provided by the instructor during the course, or as part of the competition.

Data

The dataset consists of two files:

- Test data: test.csv
- Development data: dev.txt

The test data has the following format:

```
ID,TEXT
0,What are the hottest IT startup companies in Mumbai?
1,How often do you drink coffee (-based) drinks?
2,Which contries provide financial help to India?
3,What are some interesting facts about the NSG?
```

There are two columns:

- ID: the identifier of a datapoint
- TEXT: a string with a question asked on a particular online service

For each question in the dataset there is another one which expresses approximately the same meaning. Your goal is to find it.

The development data has a similar format, but in addition to the columns ID and TEXT, it also has the column PARA_ID (for paraphrase ID) which specifies which other question asks for the same information:

```
ID,TEXT,PARA_ID
10000,What is the value of pi?,16250
10001,How can I attract Hyderabad boys?,16551
...
16250,How do you find out the full number of Pi?,10000
...
16551,How do I attract Hyderabad boys?,10001
```

You can use this development data to tune or improve your method in any way you wish. You can also ignore it.

Warning

Please note that this is unexpurgated user-generated data originating from the web and that it contains language which you may find offensive, disturbing or otherwise objectionable.

Task

Your task is determine which of for each question, which other question is seeking the same information. Your solution file should look like this:

```
ID,TEXT,PARA_ID
0,What are the hottest IT startup companies in Mumbai?,12345
1,How often do you drink coffee (-based) drinks?,54321
```

That is, it should have a header, and for each datapoint in `test.csv` there should be a corresponding row in your solution. The first column should be the `ID` of the question. The second column can be anything as it will be ignored for the scoring, but you can use it to display the text of the question. The third column should give the `ID` of the other question which according to your method means the same as the current one. In this dataset there is always exactly one other questions with the same meaning.

Method

There is are two restriction on the method used:

- it is automatic, that is, by re-running your code it should be possible to re-create your solution file,
- it does not use any external dataset which contains the exact same questions as the provided files.

Some hints:

- represent sentences as vectors with counts (or other weights) of words, sequences of words, or sequences of characters
- use one of a number of similarity or distance metrics to choose the most similar sentences
- use the development data to decide which of the possible option is likely to work well
- you could even use the development data for supervised learning if you wish.

Evaluation

The evaluation metric used is the error rate (the lower the better).

Submission to Codalab

This is the link to the Codalab competition: Matching questions

You can submit your results in the **Participate** link. After uploading your file, make sure to *submit to leaderboard*.

Over the course of the competition you can make 7 submissions. The results from all the participating teams will be displayed in the **Results** tab.

The submission file should be a `.zip` file with a file named `result.csv` in it. (Make sure there are not additional subdirectories in the zip file.) Your solution file should look like this:

```
ID,TEXT,PARA_ID
0,What are the hottest IT startup companies in Mumbai?,12345
1,How often do you drink coffee (-based) drinks?,54321
```

The order of the lines doesn't matter: the evaluation is based on matching IDs. Your file needs to use a valid CSV format (be especially careful if you are including strings with commas in them). It is recommended to use the Python `csv` library to create the file.