DPA Group Project Guidelines

Spring 2019 Block 4

Overview

For this assignment you will participate in a competition as a member of a group. The groups will be assigned by the instructor. Your group will need to submit a solution to a text mining problem, a report describing your solution, and the code which you used to generate it. One solution and one report should be submitted by each group unless extenuating circumstances exist.

Evaluation

The assignment grade will be based on the quality of your work as judged by the report and code. In addition, you will receive a bonus proportional to 1/rank in the competition. For example if your report/code is graded as 8.0, and you place second in the competition, your assignment grade will be 8.0 + 1/2 = 8.5.

Deliverables

Your group must submit a report and code. Your report should be 3 pages maximum, and should include the following:

1. A description of your final solution including: data preprocessing (if any), the representations of the data, details of your method including choices of similarity/distance metric, clustering method, and how these tools were tuned (feel free to include pseudo code).
2. A description of all other solutions your group evaluated including: in what way each solution was different than your final solution, and what aspect of these solutions led to improvements in the final solution.
3. A discussion of the performance of your solution and how it could be improved.
4. A specification of the work done by the members of the group.

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 a Jupyter 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. GroupA.zip, and submit it to the BlackBoard assignment page.

In addition you will need to submit your solution file to 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.

IN SUMMARY: each group submission consists of two parts:

1. A zip file with your report and your code (submit to BLACKBOARD)

2. Submission of your solution (submit to CODALAB)

Code reuse

Remember that 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 this rule you will be reported to the Examination Board for fraud.

You are, however, allowed to use code examples provided by the instructor during the course.

Data description

The dataset consists of two files:
• Test data: competition_descriptions.txt
• Auxiliary data: coco_val.txt

The descriptions data file is structured such that each line is a description of an image of a person. There are four sentences describing each image in the dataset.

The auxiliary data is also a list of sentences describing an image. Each group of five consecutive sentences refer to the same image. You can use the auxiliary data to tune or improve your method in any way you wish. You can also ignore it.

Groups are welcome to include any additional resources they want to augment the data.

Task

Your task is determine which of the sentences describe the same scene. Your solution file should look like this:

0,196,42,700
1,16,99,490
2,3,390,10

That is, it should not have a header, and for each line in the description file there should be a corresponding row in your solution, in the same order. The first column should be the row number of the sentence, and the remaining three columns should give the row number of the sentences which, according to your method, describe the same image. There should always be exactly four columns in each row.

Methods

There is no restriction on the method you use, as long as it is automatic: that is, by re-running your code it should be possible to re-create your solution file.

Please restrict yourself to using python libraries that are installed on the uvt jupyter server. If your group wants to use an additional library beyond numpy, scipy, pandas, and nltk, please contact the instructor.

Some hints:

- Represent sentences as vectors with counts (or other weights) of words, sequences of words, sequences of characters, or combinations of those features.
- Look for specific words or phrases that are particularly informative.
- Use one or more similarity or distance metrics to choose the most similar descriptions and be creative about how descriptions are grouped together.
- Use the auxiliary data, which you know the correct answers to, to decide which methods are likely to work well on the competition dataset.
- You could even use the auxiliary data for supervised learning if you wish.

How submissions are evaluated

The evaluation metric for the competition is accuracy: the proportion of correct descriptions grouped together. Your score is not punished for incorrect guesses, thus randomly guessing will likely produce better scores than no grouping at all, and random guessing cannot perform worse than an empty response.

Making a submission

You should submit your solution to the CodaLab web page for our competition (see Blackboard for the link). You should submit the file detailing your results in the Participate page. After uploading your file, make sure to click the button to submit to the leaderboard. Over the course of the competition your group can make 100 submissions. The results from all the participating teams will be displayed in the Results tab.

Your solution file must be named `submission.txt`. However, to submit it to CodaLab, this file must be zipped into a .zip file and uploaded to their site as outlined above. The zip file can have any name but the contents must consist only of your file and cannot contain any additional files or folders. Please see the sample baseline submission on blackboard for an example.

I strongly suggest that early on in the process your group try to submit any results, even random guesses, to ensure the submission procedure works smoothly. Please contact the instructor or post on the forum if you have any questions about submissions or other aspects of this assignment.