Extra Credit Assignment: Community Question Answering (CQA)

Submission: December 1st 2017 (before the end of the day)

Advanced Natural Language Processing
COSC 7336 - Fall 2017

Community question answering forums are heavily used. Their popularity is due in large part to the fact that the crowdsourced nature of these platforms makes it very likely that good answers will come by in a short amount of time. However the popularity of these sites and the often relaxed user-moderated approach, has some drawbacks as well. One of the disadvantages is that users often post repeated questions. Instead of searching in the archive, a user prefers to post the question directly. In many cases, the answer to a recently posted question might be in the answer/comment to previously posted questions.

In this assignment the goal is to develop a system that for a newly posted question can help identify the most relevant answers/comments from the list of most related questions.

- **Prediction output**
  In order to run the evaluation scripts, your program output has to conform to this format:

  ```
  Q268    Q268_R13_C1    1301    0.000768639508070715    false
  Q268    Q268_R13_C2    1302    0.000768049155145929    true
  Q268    Q268_R13_C3    1303    0.000767459708365311    true
  Q268    Q268_R13_C4    1304    0.000766871165644172    false
  ```

  The first column identifies the original question, the second column identifies the relevant question and the comment, the third column is an id of the comment and relevant question, the fourth column is the predicted relevance score, and the last column is the binary prediction of relevance for the comments.

- **Evaluation of Q&A Systems**
  The systems will be evaluated using the Mean Average Precision (MAP) metric.

- **Download Data**
  The training data for this assignment is in this link train. The test data can be downloaded in this link test. Additionally, we make available a sample prediction file so you can see the format we are expecting. Download the sample prediction here: sample.

- **CodaLab Submissions:** We have set up this assignment as a shared task in CodaLab. Instructions for this are in the handout for assignment 1. Access to the shared task is through this link (or search for Question Answering in Coda Lab, the shared task has the UH logo). Please make sure you're regularly checking this link, as we may need to change it if we find that we have to make changes to the shared task set up.

  NOTE: CodaLab is an open source framework for running competitions. Your system submissions will be ranked according to accuracy of the system but the ranking will be public and thus it's super important that the username you choose for the submission is not disclosing your identity. In order to identify which student gets credit for which system submission, please note in your report, and in your source code, the name you used to identify your submission in CodaLAb (user name).

Deliverables

You will need to turn in the following:

1. System predictions in CodaLab
2. System submission in CodaLab
3. Technical Report describing the system and relevant experiments and analysis.
4. Zip file with source documents and report. Create a zip file with all the above items and submit using Dropbox file request. Name your zip file `cosc7336-extracredit-student-name.zip`.

Useful Resources

This assignment is based on the Semeval 2017 Task 3. You can read the overview paper in the following link: Task 3