Assignment 1: Neural Network Basics, Word embeddings, and Author Profiling

Submission: October 13th 2017 (before the end of the day)

Advanced Natural Language Processing
COSC 7336 - Fall 2017

---

1. **Neural Networks and Embeddings Practical:** Complete the Jupyter Notebook from this link.

2. **Author Profiling:** For the second part you will need to design and train a system to perform author profiling where one of your features comes from word2vec embeddings. Author profiling (AP) is the problem of learning to predict demographic information about the author of a given document. This AP task can be defined as a classification problem where the training data has been labeled with the demographic variables of interest and a machine learning algorithm can be trained to predict these variables with the help of a good feature engineering process.

   - **Data sets** For this assignment, the training and test datasets are labeled according to two target classes: *male* and *female*. The source of the data is social media, one subset of the data is from Twitter and the other one is from blogs. In Table 1 we show some statistics about the datasets for this assignment. Since the data was originally collected by different people, they are in different XML formats and thus you need to use different preprocessing steps for each one. In general, XML tags should be removed, but you are free to explore using any metadata available that you consider could help improve accuracy.

Dataset	**Males**	**Females**	**Total**
Blogs-Train	6,762	6,762	13,524
Blogs-Test	2,898	2,898	5,796
Twitter-Test	1,800	1,800	3,600

Table 1: Data sets available for the author profiling task

   - **Prediction output** In order to run the evaluation scripts, your program output has to conform to the format of having one single prediction per line. Each line should contain the name of the file and the prediction for it in the following format:

     <filename.txt>,<male|female>

     For example:

     1.txt,male
     2.txt,female

- **Evaluation script** You can download the evaluation script we will use for the system ranking from this dropbox link. We provide this so you can evaluate performance of your system with partitions of your training set. But you can also do multiple submissions in CodaLab, we will take the best submission received by the deadline.

  As we can see in the table above, there is only one training set. You should use that data set for training and tuning your model. Then you will need to make predictions for the test sets. The goal of having a Twitter test set is to evaluate the robustness of the model when the test data comes from a different domain. To download the datasets follow this link.

- **CodaLab Submissions:** We have set up this assignment as a shared task using CodaLab. You will need to create an account in CodaLab and register for our competition (COSC 7336 - Author Profiling) by accepting the Terms and Conditions under the Participate Tab. Access to the shared task is through this link. 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 that we have set up two phases, one per test data set. For successfully completing your system submission you need to submit a text file (submission.txt) with your system's predictions and your source code in a single zip file for each phase by the submission deadline. The submission.txt file should contain one instance prediction per line and it should begin with the file name, as discussed above. It's important that the zip file only contains the two files mentioned above, not a zip containing a folder with the contents in it; this will cause the evaluation script to fail.

  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. The Jupyter Notebook with the appropriate exercises completed via the file request in Dropbox.
2. System predictions in CodaLab
3. System submission in CodaLab
4. Technical Report describing the system and relevant experiments and analysis.
5. 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-assign1-student-name.zip`.