Week 10: A.4 - Fourth Research/Programming Assignment

Submit Assignment

* **Due** No Due Date

* **Points** 100

* **Submitting** a file upload

* **File Types** zip, pdf, py, txt, ipynb, and html

***Note that you are expected to post a summary of the fourth research/programming assignment to a***[***Week 10 discussion board***](https://canvas.northwestern.edu/courses/112790/discussion_topics/725824)***as well as posting to Assignments/Grades for grading. That way all students in the class will have a chance to learn from what you have done.***

For the fourth research/programming assignment, you define the management problem, corpus or corpora, and methods of natural language processing to be used with the understanding that **the methods must involve neural networks.**

You are welcome to build on the work you have been doing with your individual corpus, as developed in the First Research/Programming Assignment.  You may choose one of the corpora cited in the course or find an appropriate corpus online. When choosing a corpus for this assignment, ensure that the training set is sufficiently large (say, two thousand or more documents or sentences). And if you are working on a supervised learning problem, ensure that the hold-out test set is also sufficiently large (say, five hundred or more documents or sentences).  **The Reduced Reuters Corpus may not be used for this assignment** because extensive jump-start code is provided for that corpus.

**Management Problem.** Part of your job in this assignment is to **define a meaningful management problem or application**. The corpus you use should relate in some way to a business, organizational, or societal problem. Regardless of the neural network methods being employed, research results should provide guidance in addressing the management problem.

Covid-19 is a possible focus this term. See additional information at the end of this assignment write-up.

The Week 8 Module provides documentation about numerous natural language applications, including information retrieval and search, topic models, sentiment analysis, and question answering, to name a few. Any of these applications may set the stage for the management problem being addressed.

One possible application, well illustrated with examples in this course, is to build a conversational agent (chatbot). The Rasa provides a convenient framework from which to start, and Rasa satisfies the neural network requirement for this assignment because it relies on TensorFlow for deep learning.

**Neural Network Methods.** For this assignment, you must use a neural-network-based method such as a one-dimensional convolutional neural network (1D CNN), recurrent neural network (RNN), a long short-term memory (LSTM) model, or a gated recurrent unit (GRU). Generative language models are also good candidates for this assignment.

Word embeddings are an important part of contemporary natural language processing applications. Consider using word embeddings (perhaps comparing them with simple one-hot encodings) in one or another of the many NLP applications.

Generally speaking, an experiment comparing the performance of alternative neural network language models or techniques would be a good research topic for this assignment. **Consider using a factorial design with alternative methods** (dense, 1D CNN, LSTM, versus GRU) as one of the factors. A comparison of one-hot encoding versus word embeddings for word/term vectorization could be another factor. Comparisons across alternative settings for dropout regularization (none, 5 percent, . . . , 50 percent) may also be useful.

When making comparisons across models, **employ training best practices**, including a tripartite splitting of the corpus into training, validation (development/dev), and hold-out test sets and early stopping. Ensure that meaningful indices of model performance are selected, such as the area under the ROC curve for binary classification and F1 for multinomial classification. Additional criteria for comparing methods should include training time considerations, including the number of epochs to meet an early stopping rule, the training time per epoch, and the total training time.

**Methods should be fully documented** by showing the structure of each model and the number of parameters being estimated. Utilize visualizations of loss and accuracy across epochs for training and validation (dev) data. Provide summary tables showing training, validation (dev), and hold-out test set performance, as well as training time data.

Please draw on Python/TensorFlow/Keras and the rich domain of NLP tools (such as Scikit Learn, NLTK, gensim, spaCy) for your work. Note that Keras provides convenient utilities for reporting on model structure, early stopping, and monitoring the training process.

**Technical Resources.** In addition to the textbooks for the course, extensive jump-start code is available for this assignment. See the zip archive  [tf-nlp-v01.zip](https://canvas.northwestern.edu/courses/112790/files/8435233/download?wrap=1)

*Note that there are two general types of neural-network-derived word embeddings: unsupervised and supervised. Think of supervised word embeddings as "embeddings with a purpose." The example code we provide shows how to obtain supervised word embeddings.*

Here is a video reviewing the assignment and jump-start code:

**Rasa Resources.**There are extensive online materials for Rasa tools for building conversational agents (chatbots).

Start by reviewing tutorial materials under Working with Rasa by Mady Mantha at  
[$CANVAS\_COURSE\_REFERENCE$/modules/items/g5225c694d4a52c96f673c6016e84c828](https://canvas.northwestern.edu/courses/112790/modules/items/1448968)

Continue with an introduction to the Rasa Masterclass with Justina Petraityte at

[$CANVAS\_COURSE\_REFERENCE$/modules/items/g1f152d0cc0b4210c1ba9123a2ba47738](https://canvas.northwestern.edu/courses/112790/modules/items/1448969)

The Rasa Masterclass contains numerous videos and accompanying code examples.

Report Components

**Organize the paper as a research report** with each section answering basic questions as identified in bold type below:

* Abstract. An executive summary of the research.
* Introduction. **Why** are you conducting this research?
* Literature review. **Who** else has conducted research like this?
* Methods. **How** are you conducting the research?
* Results. **What** did you learn from the research?
* Conclusions. **So, what** does it all mean?

When drawing conclusions, make sure that you relate your study results back to the management problem you have identified. Which neural network methods would be most useful in addressing the management problem?

Grading Guidelines

* Data preparation, exploration, visualization (20 points)
* Review research design and modeling methods (20 points)
* Review results, evaluate models (20 points)
* Implementation and programming (20 points)
* Exposition, problem description and management recommendations (20 points)

See ***Grading Guidelines/Rubrics*** for additional grading information.

Deliverables and File Formats

Create a folder or directory with all supplementary files with your last name at the beginning of the folder name, compress that folder with zip compression, and post the zip-archived folder under the assignment link in Canvas. The following files should be included in an archive folder/directory that is uploaded as a single zip-compressed file. (Use zip, not StuffIt or any 7z or any other compression method.)

1. The text content of the paper should consist of two to five pages, double-spaced. Figures and tables may go in an appendix and do not count as part of the two to five pages. The paper should cover all five bullet points under Grading Guidelines. (The paper must be provided as an Adobe Acrobat pdf file. MS Word files are NOT acceptable.)
2. Files or links to files should be provided in the format as used by the Python program.
3. Complete program code in Python used to access and analyze the data. The code should be submitted as a plain text file, so it can be executed as a program (in a single Python program execution within an interactive Python console).
4. Output from the program, such as console listing/logs, text files, and graphics output for visualizations.
5. List file names and descriptions of files in the zip-compressed folder/directory. **Recognizing that long file names can pose problems on systems that have limited path lengths, ensure that file names are short but descriptive. For example, the file for the paper could have the form lastname-firstinitial-paper.pdf as in *miller-t-paper.pdf***

If you use a Jupyter notebook interface, include the notebook (ipynb) file itself, as well as a Python .py file. If you use the Anaconda Spyder interface, save a copy of your console listing as well as your Python .py file. These Python .py files should be included as plain text, not rich text, docx, or pdf files. Also, it is good to provide a full .html file or .pdf printed file, showing Python code and results.

**Covid-19 Option.** Part of your job in this assignment is to define a meaningful management problem or application. Covid-19 corpus research provides a possible management problem. Students are welcome to work on one or more tasks identified under the Kaggle-hosted competition. And, if students prefer, they can submit a solution to the Kaggle competition. The following links provide background information.

The competition link:  
[https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=568 (Links to an external site.)](https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=568)

Rules:  
https://www.kaggle.com/cord-19-rules-accept-form

Tasks:  
[https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks (Links to an external site.)](https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks)

Original White House Posting:  
[https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/ (Links to an external site.)](https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/)

Semantic-Scholar-Based Data Source:  
[https://pages.semanticscholar.org/coronavirus-research  (Links to an external site.)](https://pages.semanticscholar.org/coronavirus-research)

Video overview of what is known about Covid-19:  
[https://youtu.be/ddlRvqhGdPk (Links to an external site.)](https://youtu.be/ddlRvqhGdPk)[](https://youtu.be/ddlRvqhGdPk)

To learn more about the disease, there is a Coursera course at

[https://www.coursera.org/learn/covid-19?utm\_medium=email&utm\_source=marketing&utm\_campaign=tIo98GlqEeqd\_xFKDENJkw (Links to an external site.)](https://www.coursera.org/learn/covid-19?utm_medium=email&utm_source=marketing&utm_campaign=tIo98GlqEeqd_xFKDENJkw)