Project:

[due Dec 12, 11:59PM]

In the first assignment, we worked with text data and made a classifier to classify text. However, we did not focus on feature engineering. In the project, you need to focus more on feature selection/extraction step and do some more data preprocessing.
Follow the steps for the project:
1- Load your data using assignment one dataset and prepare your environment.
You have access to the following assets using your CSID that might be useful:
1- Bluenose: Bluenose.cs.dal.ca (undergrad and grads)
2-Hector: hector.cs.dal.ca (only grad students)
3-Gitlab: https://git.cs.dal.ca
2- write a function to cluster all the documents as a preprocessing step. This function assigns a cluster-id to each document. After clustering documents, we generate a different classifier for each identified cluster. (20 points)
3- Write a function to evaluate the quality of your clusters. You may apply any metric set that you like by providing a justification for that. (4 points)
4- Apply one of the following papers to enhance your feature selection method. Generate a set of features for each extracted cluster separately. (You may apply a feature extraction method from another paper by sending the paper to the TA and getting approval for that. Explain the method of your choice in a clear way.) (20 points)

- [Paper 1] Liang, Hong et al. "Text feature extraction based on deep learning: a review." *EURASIP journal on wireless communications and networking* vol. 2017,1 (2017): 211. doi:10.1186/s13638-017-0993-1
- [Paper 2] Meng, Q., Catchpoole, D., Skillicom, D., & Kennedy, P. J. (2017, May). Relational autoencoder for feature extraction. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 364-371). IEEE.
- [Paper 3] Maggipinto, M., Masiero, C., Beghi, A., & Susto, G. A. (2018). A Convolutional Autoencoder Approach for Feature Extraction in Virtual Metrology: Paper ID 259. *Procedia Manufacturing*, *17*, 126-133.
- [Paper 4] Katsuki, T., Ono, M., Koseki, A., Kudo, M., Haida, K., Kuroda, J., ... & Suzuki, A. (2018, June). Feature Extraction from Electronic Health Records of Diabetic Nephropathy Patients with Convolutioinal Autoencoder. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

5- Improve your classification approach by using new features and taking advantage of deep neural networks. Explain changes and the reason for changes in compare to your assignment one. (8 points)
6- Go to Brightspace and upload your notebook containing all of your work under the Project section.
7- Go to Brightspace and under the Quiz of Project section, upload your answer for each question separately.
Notes:

- Each question will be evaluated by
  - **Excellent**: Material is clearly grasped and convincingly presented. (4/4points)
  - **Good**: Most of the key ideas are present and clearly presented. (3/4points)
  - **Satisfactory**: Much of the assignment may be a summary or paraphrase of the material. There may be inaccuracies in the presented content. (2/4 points)

    - **Below Standard**: Has not grasped the material enough to answer the question effectively. (¼ points)
    - **Wrong:** No Answer, unrelated, or does not make sense. (0/4 points)
  - Delay to submit from 5 minutes to 24hours will deduct 25% of your mark. Between 24 hours and 48 hours deducts 50% of the mark and between 48 hours 52hours deducts 75%, and more than 52hours deduct 100% of the mark.
- Projects can be done in groups of a maximum of two members. If you submit as a group, you need to add a document to show the contribution of each member.
- If you have an issue executing your code because of memory issues, you can select a reasonable subset of data. You need to explain this as part of your submission.
- If you get the help from a TA in the learning center, explain in which part.
- If you have any questions about this assignment, post your question on the discussion form of the project on Brightspace.

Goodluck !