# Assignment 5 (Text Mining)

This assignment is **individual** effort.

# Problem Definition

This assignment has three objectives:

* Learn how Text Classification is used in an end-to-end scenario per the Research Paper provided to you for this assignment.
* Apply concepts you learned in Text Preprocessing, Text Vectorization, and Text Classification using MultinomialNB.
* Complete the analysis using SKLearn Pipeline and Grid Search concepts.

You have been provided with the following information:

* **ICIS2015MousaviRaghuFrey.pdf**: A Research Paper that applies Naïve Bayes to solve a Business Problem. *“Assessing Order Effects in Online Community-Based Health Forums”, R. Mousavi, T. S. Raghu, Keith Frey, ICIS 2015.*
* **HealthProNonPro.zip**: Health data used in the Research Paper (provided to you with permission from the authors of the paper).

# Requirement for this Assignment

Complete the following for this assignment:

1. Read in the data into a DataFrame.
2. Setup the data for Training/Testing. Use 20% for testing.
3. Use Spacy to preprocess the data. Explore and pick appropriate preprocessing steps.
4. Setup a Pipeline with TfidfVectorizer and Naïve Bayes.
5. Do Grid Search with 4-fold Cross Validation to search for the best values for the following two hyper-parameters (and any additional hyper parameters you may want to tune):
   * sublinear\_tf in TfidfVectorizer
   * alpha in Naïve Bayes
6. Use the Best Estimator resulting from the Grid Search for Prediction/Evaluation. Print the following evaluation metrics:
   * Accuracy score
   * Confusion matrix
   * Classification report
7. Extract the true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) using the following command.

TN, FP, FN, TP = metrics.confusion\_matrix(y\_true=ytest, y\_pred=ypred).ravel()

Then, using TN/FP/FN/TP write code to calculate the overall accuracy, the precision (for class 0 and class 1), the recall (for class 0 and class 1), and the f1-score (for class 0 and class 1). These should match what you are seeing in the accuracy\_score and classification\_report you printed above.

Hints:

* You should end up with an accuracy score of at least 92% for this assignment.
* Programming concepts you will need for this assignment were covered in ExerciseSet5a.ipynb, ExerciseSet5b.ipynb, ExerciseSet5c.ipynb, TextPreprocessing.ipynb, TextVectorization.ipynb, TextClassification.ipynb, TextTopicModeling.ipynb.

# Submission for this Assignment

Upload your final python script (either .py or .ipynb is fine). Use the following naming convention:

<CohortX><FirstName><LastName>Assignment5.py

*[Example: CohortAHinaAroraAssignment5.py]*

**NOTE: If you have used google colab for the assignment, open the file on spyder, save it as a .py file, and upload the .py file.**

# Appendix: Preamble from the Authors of the Paper

Many Question Answer platforms face the challenge of ensuring quality responses to ensure high user engagement. A low-tech approach to this, obviously, is to rely on the “wisdom of the crowd.” That is, either (a) let the person asking the question rate the best answer, or (b) let the crowd vote for the best answer. While this solution is good, it is not entirely optimal. For instance, if there was a way to automatically identify “high quality” answers as they come in, then we could present the predicted highest-quality answer as the first answer (instead of chronologically presenting all responses). This would improve the quality of the subsequent answers

In this paper, we set out to answer three questions:

* Does the textual quality of the answer chosen as the best answer by the asker differ from that of a regular answer?
* Does the textual quality of the first answer influence the textual quality of the subsequent answers?
* Does the textual quality of the first answer influence the number of the subsequent answers?

From an application perspective on text mining, we want you to note the following:

1. Text mining is not an end in itself. You often need to do additional steps to develop insights that can be applied to a business problem.
2. You can combine text mining with other empirical methods (e.g., discriminant analysis, regression analysis, etc.)
3. It helps to identify other similar datasets to help improve or even get started on your text mining application. For example, we augmented the Yahoo dataset with data from asktheDoctor.com. This was an essential step to answering the first basic question – “how do you know an answer is as good as that of a professional?”
4. But, we did not stop there, we then looked at other reliability indicators – such as readability index (a well known empirical approach to characterize “readability” of a block of text), and number of medical terms (that professionals are likely to use).

Read the paper to get an idea of how you progressively build insights by starting with text mining and then reverting to traditional regression-based models to evaluate your hypotheses. The implementation of an analytics backend to a QA forum would rely on the Naïve Bayesian classifier and the identification of other textual features as described in this paper.

CAUTION: This dataset is **UNFILTERED REAL DATA** from Yahoo! Answers and AskTheDoctor.com. Some of the content could make you uncomfortable. This assignment **DOES NOT REQUIRE** you to open the file and read the answers. If you are likely to be offended by the language in these files, just don’t read them.