Text Classification by using GraphLab

Due on Apr 19th

In this assignment, you will have chance to investigate and compare different machine learning algorithms (for text classification) by leveraging GraphLab API.

Dataset:
The dataset (reuters.csv) we use this time is from Reuters news, which contains 34 topics and 2,286 documents in total. Each topic has no more than 50 documents as well as no less than 20 documents.
We also attached a sample dataset (test.csv) here only for sample code usage. Sample code (assignment3_samplecode.ipynb) may be helpful to finish this assignment. You can borrow the idea in the sample code as well as derive your own idea.

Data set sample:
ipi, French industrial production fell a seasonally ….
reserves, French official reserves rose 258 mln francs….

In the sample above, The first column("*ipi","reserves*") is the class label and the second one("*French industrial production fell a seasonally…")* is the text content. By using machine learning algorithms, we will be able to predict the text label by using the news content.

Task 1:
Text cleaning. (1 point)
Currently, the text data can be noisy. In the first step, we need to clean the data by removing the stopword and all other vacabularies with frequencey less than 2 times from the text. More detailed information is available at https://dato.com/learn/userguide/text/analysis.html In roder to assist your work, please refer the example in the sample code "*Text Cleaning*" section.

Task 2:
Generate feature space for each document. (2 points)
By using Graph Lab text processing package (https://dato.com/learn/userguide/text/analysis.html) you should generate different kinds of features for each document in order for model calculation. Create at least 1, Word frequency; 2. TFIDF representation as new features.(You can add more features than that as long as you can find and calculate from the text content)
In roder to do that, please refer the example in the sample code "*Generating feature*" section.

Task 3:
Use classification model to predict topics based on text content and finish following 2 report questions. (7 points)

3.1 Build classification models. (2 points)

Please try different classification algorithms (https://dato.com/learn/userguide/supervised-learning/classifier.html) : 1. Nearest Neighbor Classifier; 2. Support Vector Machines, 3. Decision Tree Classifier and 4. Boosted Trees Classifier to build models to predict topics for each text.
In roder to do that, please refer the example in the sample code "*Use different classifiers*" section.

3.2 Report (5 points)
1. Which classifier(s) can be more useful for this task? (hint: some classifiers can't be used for this task)? And what kind of text representation (A.K.A features for the data) is more useful? Why? (3 point)
2. Model evaluation report. Compare different classifiers' result by comparing model accuracy, auc, precision and recall (which one is the best)? (2 points)
In roder to do that, please refer the example in the sample code "*Evaluation*" section.