MIE 1624 Introduction to Data Science and Analytics – Fall 2018

Assignment 1

Due Date: 11:59 pm, November 5, 2018

Submit via Quercus

Background:

Sentiment Analysis is a branch of Natural Language Processing (NLP) that allows us to determine algorithmically whether a statement or document is "positive" or "negative".

Sentiment analysis is a technology of increasing importance in the modern society as it allows individuals and organizations to detect trends in public opinion by analyzing social media content. As the airline industry become increasingly competitive, airline companies are always trying their best to satisfy customer experience. Through sentiment analysis, these companies can obtain information on customers' overall satisfaction and adopt business strategies to address aspects with low satisfaction.

The purpose of this assignment is to compute the sentiment of text information - in our case, tweets posted in 2015 regarding US airlines - and answer the research question: ***"What can public opinion on Twitter tell us about the US airlines in 2015?"*** The goal is to essentially use sentiment analysis on Twitter data to get insight into the people's opinions on US airlines.

Central to sentiment analysis are techniques first developed in text mining. Some of those techniques require a large collection of classified text data often divided into two types of data, a training data set and a testing data set. The training data set is further divided into data used solely for the purpose of building the model and data used for validating the model. The process of building a model is iterative, with the model being successively refined until an acceptable performance is achieved. The model is then used on the testing data in order to calculate its performance characteristics.

Produce a report in the form of an IPython notebook detailing the analysis you performed to answer the research question. Your analysis must include the following steps: data cleaning, exploratory analysis, model preparation, model implementation, and discussion. This is an open-ended problem: there are countless different ways to approach each part of the analysis and therefore the motivation for each step is just as important as its implementation. When writing the report, make sure to explain (for each step) what it is doing, why it is important, and the pros and cons of that approach.

Two sets of data are used for this assignment. The *generic_tweets.txt* file contains tweets that have had their sentiments already analyzed and recorded as binary values 0 (negative) and 4 (positive). Each line is a single tweet, which may contain multiple sentences despite their brevity. The comma-separated fields of each line are:

0	class	the polarity of each tweet (0 = negative emotion, 4 = positive emotion)
1	id	the id of the tweet (e.g. 2087)
2	date	the date of the tweet (e.g. Sat May 16 23:58:44 UTC 2009)
3	query	the query (e.g. lyx). If there is no query, then this value is NO_QUERY.
4	user	the user that tweeted (e.g. robotickilldozr)
5	text	the text of the tweet (e.g. Lyx is cool)

The second data set, *US_airline_tweets.csv*, contains a list of tweets regarding several US airlines. The comma-separated fields of each line are:

0	id	the id of the tweet
1	sentiment	can be “positive” or “negative”
2	negative_reason	reason for negative tweets. Left blank for positive tweets.
3	user	the user that tweeted
4	retweet_count	number of retweets
5	text	the text of the tweet

Both datasets have been collected directly from the web, so they may contain html tags, hashtags, and user tags.

Learning objectives:

1. Implement functionality to parse and clean data according to given requirements.
2. Understand how exploring the data by creating visualizations leads to a deeper understanding of the data.
3. Learn about training and testing and logistic regression.
4. Understand how to apply a machine learning algorithm (logistic regression) to the task of text classification.
5. Improve on skills and competencies required to collate and present domain specific, evidence-based insights.

To do:

1. **Data cleaning (30 marks):**
   The tweets, as given, are not in a form amenable to analysis -- there is too much 'noise'. Therefore, the first step is to "clean" the data. Design a procedure that prepares the Twitter data for analysis by satisfying the requirements below.
   - All html tags and attributes (i.e., /<[^>]+>/) are removed.
   - Html character codes (i.e., &...;) are replaced with an ASCII equivalent.
   - All URLs are removed.
   - All characters in the text are in lowercase.
   - All stopwords are removed. Be clear in what you consider as a stopword.
   - If a tweet is empty after pre-processing, it should be preserved as such.

2. **Exploratory analysis (10 marks):**
   - Design a simple procedure that determines the airline of a given tweet and apply this procedure to all the tweets in the US airline dataset. A suggestion would be to look at relevant words and hashtags in the tweets that identify to certain airlines. What can you say about the distribution of the US airlines of the tweets?
   - Present a graphical figure (e.g. chart, graph, histogram, boxplot, word cloud, etc) that visualizes some aspect of the generic tweets and another figure for the US airline tweets. All graphs and plots should be readable and have all axes that are appropriately labelled.

3. **Model preparation (15 marks):**
   Split the generic tweets randomly into training data (70%) and test data (30%). Prepare the data for logistic regression where each tweet is considered a single observation. In the logistic regression model, the outcome variable is the sentiment value, which is either positive or negative. The independent variables or features of the model can be whatever you want. As a suggestion, you can use the frequency of each word as the features of the model, or alternatively, you can first tag each n-gram by its part of speech and then use the frequency of each part of speech as the features of the model.

4. **Model implementation (30 marks):**
   Train a logistic regression model on the training data and apply the model to the test data to obtain an accuracy value. Evaluate the same model on the US airline data. How well do your predictions match the sentiment labelled in the US airline data?

   Split the **negative** US airline tweets into training data (70%) and test data (30%). **Use the sentiment labels in the US airline data instead of your predictions from the previous part.** Train a multi-class logistic regression model to predict the reason for the negative tweets. There are 10 different negative reasons labelled in the dataset. Feel free to combine similar reasons into fewer categories as long as you justify your reasoning. Again, you are free to define input features of your model using word frequency analysis or other techniques.

5. **Discussion (15 marks):**
   Answer the research question stated above based on the outputs of your first model. Describe the results of the analysis and discuss your interpretation of the results. Explain how each airline is viewed in the public eye based on the sentiment value. For your second model, if there are any tweets for which the model failed to predict the correct negative reason, explain why. Justify your explanation with a few examples from the test sets. For both models, suggest one way you can improve the accuracy of your models.

Bonus:

We will give up to 10 bonus marks for innovative work going substantially beyond the minimal requirements. These marks can make up for marks lost in other sections of the assignment, but your overall mark for this assignment cannot exceed 100%. The obtainable bonus marks will depend on the complexity of the undertaking, and are at the discretion of the marker. Importantly, your bonus work should not affect our ability to mark the main body of an assignment in any way. Any bonus work should be explicitly labelled as "Bonus" in its own section. You may decide to pursue any number of tasks of your own design related to this assignment, although you should consult with the TA before embarking on such exploration. Certainly, the rest of the assignment takes higher priority. Some ideas:

- Explore alternative classification methods and compare their performance to that of the logistic regression model.
- While the exploratory analysis section requires only two figures, you can explore the data further. You can also display the results of the logistic regression model visually.
- Improve the performance of your model by tuning its hyperparameters.

Tools:

- **Software**
  - **Python Version 3.X** is required for this assignment. Python Version 2.7 is not allowed.
  - Your code should run on the Data Scientist Workbench (Kernel 3).
  - All libraries and built-ins are allowed but here is a list of the major libraries you might consider: Numpy, Scipy, Scikit, Matplotlib, Pandas, NLTK.
  - No other tool or software besides Python **and its component libraries** can be used to touch the data files. For instance, using Microsoft Excel to clean the data is not allowed.
- **Required data files**
  - **generic_tweets.txt:** classified Twitter data containing a set of tweets which have been analyzed and scored for their sentiment
  - **US_airline_tweets.csv:** Twitter data containing a set of tweets from 2015 on the US airlines, which needs to be analyzed for this assignment
  - The data files cannot be altered by any means. The IPython Notebooks will be run using local versions of these data files.

- **Optional data files**
  - **corpus.txt:** corpus containing a set of words and their associated sentiment values
  - **stop_words.txt:** file containing an extensive list of stopwords
  - You may use these files if you wish but you are not required to.

What to submit:

Submit via Quercus an IPython notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention:

lastname_studentnumber_assignment1.ipynb

Make sure that you comment your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase and underscores are used as shown. **A program that cannot be evaluated because it varies from specifications will receive zero marks. Late submissions will not be accepted.**

Tips:

1. You have a lot of freedom with however you want to approach each step and with whatever library or function you want to use. As open-ended as the problem seems, the emphasis of the assignment is for you to be able to explain the reasoning behind every step.
2. While some suggestions have been made in certain steps to give you some direction, you are not required to follow them. Doing them, however, guarantees full marks if implemented and explained correctly.

TAs:

Shayne Lin (Shayne.lin@mail.utoronto.ca)
Sanjif Rajaratnam (sanjif.rajaratnam@mail.utoronto.ca)
Alina Sienkiewicz (alina.sienkiewicz@mail.utoronto.ca)