Homework: Text Classification

Background

We have learned text classification (supervised machine learning) as a computational technique for extracting insights from text data. In this HW, you will demonstrate your skills in the following areas:

1. Text data preparation (e.g., collecting)
2. Text data preprocessing (e.g., cleaning, transformation, TFIDF): Feature engineering
3. Training and testing supervised machine learning algorithms (e.g., Naïve Bayes) for text classification → Building predictive models for text mining
4. Evaluation of predictive model (e.g., accuracy, ROC curve, true positive rate)
5. Interpreting the results of your predictive models and discuss technical and managerial implications

Requirements

1. Create the Ipython notebook (named as "yourlastname_firstinitial_ HW_TextClassification.ipynb") and complete the **tasks** below. **Each cell must be properly numbered and formatted using Markdown**.
2. For any questions regarding Python coding, you should consult lecture notes (Jupyter Notebooks) and **https://stackoverflow.com/**
3. Your Jupyter Notebook needs to be properly formatted using Markdown and comment.

Required procedures & tasks

- You are required to build a predictive model to classify human resource (HR) news and non-HR news. A template for the text classification model you need to build can be found from **https://www.cbinsights.com/research/team-blog/human-resources-news-classification-machine-learning/** You're required to read this article and understand the steps/tasks to build such a HR text classification model. In simple, your model should be able to categorize news articles as either HR or non-HR (binary).
- First, you need to demonstrate your skills in data collection and preparation. To build a text classification (or predictive) model, you need a dataset (csv file) that contains texts (news articles) and labels (either HR[1] or non-HR[0]). In general, the larger the dataset, the better for the model. However, you need to be realistic that this is a HW and you have a limited time.
    - HR data source: There are many HR news portals (or websites) on the Internet. Choose a "**static**[1]" website and collect (perhaps, a couple hundreds) HR-related

[1] A **static** website is a HTML-based website. It is easy to crawl HTML-based websites (e.g., IMDB.com, RottenTomatoes.com) using regular Xpath and for looping. On the other hand, a growing number of websites use Java Script (and other advanced web technologies) and these sites are called "dynamic" sites (e.g., http://www.koovs.com). Crawling dynamic websites need Scrapy and Splash (see https://www.youtube.com/watch?v=VvFC93vAB7U).

news articles. It would be even better to collect HR articles focusing on a specific topic area (e.g., staffing, recruitments) This will help training machine learning algorithms.

- Non-HR data source: Perhaps, you already have some news articles (from BBC.com) about sports or politics. If you want to collect new data, that would be fine too!
- Dataset: The dataset size is up to you.

- Second, you need to demonstrate your skills in text preprocessing and feature engineering. That is, you need to clean and transform your dataset using vectorization, document-term matrix, and other feature engineering techniques (e.g., stopwords, TFIDF).
- Third, you need to demonstrate your skills in building text classification models using different machine learning algorithms (minimum three different algorithms required) and selecting the best model. For this, you should know how to use pipelines (and grid search, if applicable) and to evaluate different models using performance metrics. You must employ multiple performance metrics to evaluate the models. Using accuracy alone could misguide you and your client. Explain about your best model (e.g., algorithm, parameters).
- Fourth, you need to demonstrate your skills in applying your best model to a new (unlabeled) dataset. For this, you need to prepare a small dataset (10 to 20 news articles) containing texts only (without labels): some articles should be HR-related and others non-HR. Report how good your model is in predicting a new dataset as either HR or non-HR.
- Fifth, demonstrate your skills in "storytelling". To me (and your client), this is the most important part of this HW. Provide your interpretations of the results and discuss technical and managerial implications (e.g., most important features) for your client. For example, you can discuss the benefits of your model as well as its limitations.
- Finally, demonstrate your aptitude to learn from online resources. In class, we covered three supervised machine learning algorithms: Naïve Bayes, KNN, and SVM. Now, you need to apply an algorithm called Neural Network[2] (NN) and report the performance of your NN-based text classification model. Googling "neural network sklearn" will return a large list of online resources. Understanding how neural network works is beyond the scope of this course.

This is semi open-ended project so grading will be somewhat subjective. I expect you to go above and beyond the requirements/tasks listed above. This is for your benefit!

---

[2] NN is also known as deep learning, which is the backbone of a growing number of industry applications including AlphaGo https://www.youtube.com/watch?v=SUbqykXVx0A

How to Format Your Jupyter Notebook:

- Start with K-State Honor Code "**On my honor, as a student, I have neither given nor received unauthorized aid on this academic work**."
- Include the questions (tasks) and the question numbers, using Markdown, prior to each cell containing python codes.
- Must be professional and neat (You are submitting this report for consideration by your upper managers

Submission

- Complete Ipython notebook in **HTML** version (yourlastname_firstinitial_ HW_TextClassification.html).
- If you want to turn in more than one Jupyter Notebook, then make sure to name them properly (e.g., yourlastname_firstinitial_**1**_HW_TextClassification_DataCollection.html; yourlastname_firstinitial_**2**_HW_TextClassification_BuildingTextClassificationModel.html)