**Eth. AI, HW5, Privacy**

**Due Week 13, Thursday, 11:59 PM**

**Goal:** Investigate and correct privacy leakage in terms of text classification.

**Resources:** Tutorials on python and sklearn library can be found both in [W3Schools](https://www.w3schools.com/python/default.asp) as well as [here (Python course Slides](https://ind657-my.sharepoint.com/:f:/g/personal/jrusert_pfw_edu/Ei_EBaJU-IJBgpwM2A7mmvUB-GrO9jUSvSq0CgMgwZgwYw?e=DJcAgX)). Also, sklearn has demonstrations of models and functions on the specific model pages as well (see below links).

**Models:** You will be working with 1 type of model for this homework, 1 from the sklearn python library:

[Logistic Regression](https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html)

This homework consists of 2 parts, but all parts use the same dataset for training and testing provided on brightspace. Even though there are 2 parts, there still should only be 1 final report and 1 .py file handed in.

## Part 1: Privacy Leak in Classification Model

### Part a: Exposing the Leak

Your company has been tasked with creating a classifier which can classify a text as either Republican or Democrat based on social media posts. Your coworkers and you have gathered data and split it into training (train\_politics.csv) and testing (test\_politics.csv). UserIDs have been removed as a column to protect privacy. Additionally, “Republican” and “Democratic” labels have been converted to 0 and 1, respectively. Your coworkers claim that this is enough to protect privacy. Your goal is to demonstrate that the model can still leak information as is.

**Goal:** Demonstrate to your coworkers that information can be leaked from the trained classifier. You will do this using your coworkers' usernames (**@WilliamD, @TinaK, @DonaldW**). You know that William is more associated with the Democratic party, Tina with Republican, and Donald flips back and forth on issues. Furthermore, you also have the knowledge that Republicans more often tweet about “**america**” and Democrats more often tweet about “**change**”.

**Coding goals:**

1. Train a LR model and test the LR model on the provided training and testing datasets.
   * You should use the TF-IDF Vectorizer as used in previous homework.
   * **Do not** pre-preprocess the text at all.
   * Find the scores of the training and testing sets.
2. Using the .predict() and .predict\_proba() functions, you should demonstrate to your coworkers, that the model leaks their political party associations.
   * Your claims should be backed up by numerical evidence.
   * Hint 1: As Donald is party neutral, he may provide a starting baseline to view the model through.
   * Hint 2: Use the knowledge about what Republicans and Democrats tweet more about to help explore the model outputs. Do not rely on the usernames alone as Donald is neutral and it will still only assign one label. (Use only the provided information, **do not** scan the training data or testing data).

**To be added to the report:**

* The scores of the training and test data of the provided datasets (in a table).
* A description of the privacy leak you found and the steps for how you found it. As well as the numerical values supporting your claims.

### Part b: Fixing the Leak

Now that you have shown a privacy problem. Your goal is to better anonymize the training and testing data. You should do this by anonymizing all usernames in each text. To make it simple, you can assume a username is anything with an ‘@’ symbol inside. You should not simply remove these but anonymize them somehow (left up to you).

**Goal:** Anonymize the usernames so that the privacy leak shown is no longer existent, while also not reducing classification ability too much.

**Coding goals:**

* Write a function, **anonymize\_text(text),** which takes in a single text and anonymizes all usernames in that text, returning the anonymized text.
* Use this function to anonymize the training and testing files. Create 2 new files “test\_politics\_anonymized.csv” and “train\_politics\_anonymized.csv”.
* Using the 2 new files, train and test a new LR on the anonymized data.
* Following the same procedure in part a, demonstrate that the privacy leak is no longer present. (The steps should be similar to those used in part a, but now the values show no privacy leak.)

**To be added to the report:**

* A description of how you chose to anonymize the usernames and your reasoning behind this choice.
* The scores of the LR on the anonymized training and testing (in a table with the previous LR (part a) scores).
* A description of why the privacy leak is no longer present (with numerical values to back up your claim).

## Part 2: Better Data Anonymization

Since bringing the above to your co-workers attention, they want to greatly increase anonymization. They aim to do this by removing all names from each text. They define a name as any word that starts with a capital letter. You worry that this is not a good approach to anonymization as it will cause accuracy drops in the model.

**Goal:** There are two goals: First, you will verify this drop in accuracy and second you will propose and test a second method which still anonymizes names but maintains accuracy.

**Coding goals:**

1. Verify that removing all names will cause accuracy to drop.
   * Further anonymize your anonymized files from Part 1 (train/test\_politics\_anonymized.csv) by removing all names from each text. For simplicity, a name is defined as a word that starts with a capital letter (this is broad but okay for this exercise). Create 2 new files: **test\_politics\_removed.csv** and **train\_politics\_removed.csv.** Note that if a text becomes an empty string (no words remaining) then you can simply not include it in the new files.
   * Train and test a new LR on the created files, finding the scores for both.
2. Create a new method, **better\_anonymized\_text(text)**, which replaces names in a smarter manner, rather than simply removing them.
   * Hint: Masking them equally may prove no better than removing them. Think about ways to anonymize, while still keeping uniqueness of values.
   * Create 2 new files: **test\_politics\_better.csv** and **train\_politics\_better.csv**
   * **Do not** change the definition of a name. The anonymization should improve not the name selection.
3. Train and test a new LR on the “\*\_better.csv” files.

**To be added to the report:**

* The scores of the LR on the “removed” data in a table with the previous sets of results, along with discussion showing that the coworker proposed method is hurting accuracy.
* A description of your better method to anonymize the names and the intuition behind it.
* The scores of the LR on the “better” data in a table with the previous sets of results. Along with a discussion showing that your method both anonymizes and maintains accuracy better.

## To Turn In:

* Python file (**USERNAME\_HW5.py**.) which contains functions related to the above parts. You should include comments and useful function names to differentiate the testing portions.
* Report which contains:
  + The above requirements for each part. Make sure you make it clear (via titles or subheadings) which part you are describing.
  + Discussion of issues or difficulties encountered.
  + A note of any AI tools used in coding or writing.
* Your **six** produced anonymized files. (train/test\_anonymized, train/test\_removed, train/test\_better).

**Additional Rules (MUST BE FOLLOWED):**

1. All Homeworks should follow the overall [homework guidelines](https://ind657-my.sharepoint.com/:p:/g/personal/jrusert_pfw_edu/EZxf1ZsRXjBEkQLNhEhdTgUBt6U64KiT1DJ1YHtkARgKwA?e=bwHrbf)
2. The code should be written in python 3.
3. If noted, the functions must follow the naming and number of arguments as demonstrated.
4. You should make your code modular to the different steps. (You may have more functions to help your main functions)
5. You should be adding comments to document your code. **If I can’t understand why you perform an action, then I can’t credit you for performing that action.**
6. The report should be readable and reference your code, **without explicitly including code.**
7. You should include your name and homework number in the comments at the beginning of the python file.

**Report**

The reports for the homeworks are necessary to communicate your learning and thinking through of the material. Examples of good reports can be found on brightspace under Additional Resources/Guides. Note that your report style may differ, but it is a good reference to start with.

**Grading**

Assignment will be graded as follows:

|  |  |
| --- | --- |
| **Description** | **Points** |
| **(.py file)** Part 1 coding requirements | 10 |
| **(.py file)** Part 2 coding requirements | 10 |
| **(Report)** Part 1 report requirements | 10 |
| **(Report)** Part 2 report requirements | 10 |
| **(Report)** Other Report Requirements | 5 |
| **(.py file)** Documentation/Code (Comments, functions, etc) | 5 |
| **Total:** | **50** |

* **If the code does not run, it cannot be graded well.** (Many points can be lost if the code cannot be run, as I will not be able to fully test the implementation of the functions).
* **Breaking of the additional rules can result in applied penalties.** (Always make sure you are checking against the rules)

**Suggestions**

* **Documentation is key for showing your effort in this homework.** Make sure you are noting why you make certain decisions all throughout your code.
* The slides for previous classes are posted, so please refer to these and the book for ideas during implementation.
* Start simple, build up complexity. You should always make sure your new ideas being added do not cause your program to crash. So starting simple is the best way to a) maintain the ability to keep your code running, b) add in comments for documentation and thought process as you add more code.
* Work through the homework yourself, rather than sharing ideas (especially not code) with other students. **As a reminder, plagiarism (or sharing) of code is strictly prohibited.** This assignment is complex enough that significant overlap between students will be suspicious.
* If you have not worked with python before, w3schools can help you translate your previous coding experience to python (<https://www.w3schools.com/python/default.asp>)
* Stop by office hours to discuss ideas. I am always happy to help you think through your process!