# NLP Assignment 1

Name: Vamsi Tallam UIN: 432001932

**How to run the code:**

1. Extract the SpamLord.py file from SpamLord.zip
2. Place SpamLord.py in the Folder (C:\Users\vamsi\OneDrive\Desktop\NLP\PA1)
3. PA1 folder has data_dev folder which contains a subfolder dev and devGOLD file.
4. Arguments to run the code are path to dev folder and devGOLD file
5. Using Spyder, we can run the code using the following command:
   runfile('C:/Users/vamsi/OneDrive/Desktop/NLP/PA1/SpamLord.py', args = 'data_dev/dev data_dev/devGOLD ', wdir='C:/Users/vamsi/OneDrive/Desktop/NLP/PA1')
   or we can also use on console
   python3 SpamLord.py data_dev/dev data_dev/devGOLD

**Results and Analysis**:

```
In [8]: runfile('C:/Users/vamsi/OneDrive/Desktop/NLP/PA1/SpamLord.py',
args  = 'data_dev/dev data_dev/devGOLD ', wdir='C:/Users/vamsi/OneDrive/
Desktop/NLP/PA1')
True Positives (56):
{('Ahmed', 'e', 'tanzir@tamu.edu'),
 ('Ahmed', 'p', '979-845-4908'),
 ('Amato', 'e', 'amato@tamu.edu'),
 ('Amato', 'p', '979-458-0722'),
 ('Amato', 'p', '979-862-2275'),
 ('Andersen', 'e', 'flemminglandersen@tamu.edu'),
 ('Andersen', 'p', '979-845-3510'),
 ('Bettati', 'e', 'bettati@cs.tamu.edu'),
 ('Bettati', 'p', '979-845-5469'),
 ('Chai', 'e', 'jchai@cs.tamu.edu'),
 ('Chai', 'p', '979-845-3510'),
 ('Chaspari', 'e', 'chaspari@usc.edu'),
 ('Chaspari', 'p', '213-740-3477'),
 ('Choe', 'e', 'choe@tamu.edu'),
 ('Choe', 'p', '979-845-5466'),
 ('DaSilva', 'e', 'dilma@cse.tamu.edu'),
 ('Daugherity', 'e', 'daugher@neo.tamu.edu'),
 ('Daugherity', 'p', '979-845-1308'),
 ('Davis', 'e', 'davis@tamu.edu'),
 ('Davis', 'p', '979-845-4094'),
 ('Furuta', 'e', 'furuta@cs.tamu.edu'),
 ('Furuta', 'p', '979-845-3839'),
 ('Gooch', 'e', 'gooch@cse.tamu.edu'),
 ('Gooch', 'p', '979-845-5534'),
 ('Gu', 'e', 'ccs17tutorials@gmail.com'),
 ('Gu', 'e', 'guofei@cse.tamu.edu'),
 ('Gu', 'p', '979-845-2475'),
 ('Gutierrez-Osuna', 'e', 'rgutier@cse.tamu.edu'),
```

```
 ('Gutierrez-Osuna', 'e', 'rgutier@cse.tamu.edu'),
 ('Gutierrez-Osuna', 'p', '979-845-2942'),
 ('Hammond', 'e', 'hammond@tamu.edu'),
 ('Hammond', 'p', '979-353-0899'),
 ('Hu', 'e', 'hu@cse.tamu.edu'),
 ('Hu', 'e', 'xiahu@tamu.edu'),
 ('Hu', 'p', '979-845-8873'),
 ('Ioerger', 'e', 'ioerger@cs.tamu.edu'),
 ('Ioerger', 'p', '979-845-0161'),
 ('JHuang', 'e', 'jeff@cse.tamu.edu'),
 ('JHuang', 'e', 'jeffhuang@tamu.edu'),
 ('JHuang', 'p', '979-458-0722'),
 ('JHuang', 'p', '979-845-5485'),
 ('Jafari', 'e', 'jjackson@tamus.edu'),
 ('Jafari', 'e', 'lmcdow@tamu.edu'),
 ('Jafari', 'e', 'rjafari@tamu.edu'),
 ('Jafari', 'p', '979-458-9808'),
 ('Jafari', 'p', '979-862-4413'),
 ('Jafari', 'p', '979-862-8098'),
 ('Jimenez', 'e', 'djimenez@cs.tamu.edu'),
 ('Jimenez', 'p', '979-845-2434'),
 ('Juan', 'e', 'garay@cse.tamu.edu'),
 ('Juan', 'p', '979-845-4359'),
 ('Kim', 'p', '979-845-3660'),
 ('Klappenecker', 'e', 'klappi@cse.tamu.edu'),
 ('Klappenecker', 'p', '979-458-0608'),
 ('Lee', 'e', 'hlee@cse.tamu.edu'),
 ('Lee', 'p', '979-845-2490'),
 ('deWitte', 'e', 'paula.dewitte@tamu.edu')}
False Positives (2):
{('JHuang', 'p', '979-458-0718'), ('Amato', 'p', '979-458-0718')}
False Negatives (1):
{('Kim', 'e', 'ejkim@cs.tamu.edu')}
Summary: tp=56, fp=2, fn=1
```

Summary:

1. There are 56 instances of true positives
2. 2 cases of false positives – incorrectly captured 2 fax numbers as phone numbers
3. 1 case of false negative – could not capture kim's email id

**Known Issues and Limitations:**

- I started with a simple regex which could not capture all the cases, based on the cases that I missed I investigated the corresponding files, and I have added more expressions to capture.
- I tried my best to generalize them as much as possible. My code may fail to capture new instances which are not similar to the input files.

- Captured some miscellaneous cases as corner cases, to make the code more robust I converted text to lower case.
- Also tackled the case where the phone number is in multiple lines.
- I also noticed that there can be instances where the code can pick multiple phone numbers. This may not be an issue as we can pick the number correctly.

The above-mentioned cases are known issues:

1. Cases: Amato (fax) and Jhuang (fax)
   Reason: As discussed on the canvas these cases can be excluded.
2. Case: Kim (email)
   Reason: removing tags may result in unexpected issues, e.g., some email addresses are inside tags.