# Homework 1 – SpamLord

- Submitted by Pragya Shah (SC15B107)

**Problem Statement**: Build regexes (regular expressions) that extract phone numbers and email addresses from the text of given files and output it in the following format

e-mail: jurafsky@stanford.edu

Phone no.: 650-723-0293

**Solution:**

Using the given starter code, build up the regular expression for extracting emails and phone numbers, tackling one case at a time by analyzing the output of the program.

The function process_file(name, f) has been edited to incorporate code for the extraction of phone numbers (which was absent earlier). Some additional changes were made to process the obtained email from the text files into their desired output format.

NOTE: The directory "python" contains –

SpamLord.py – the original starter code without any changes

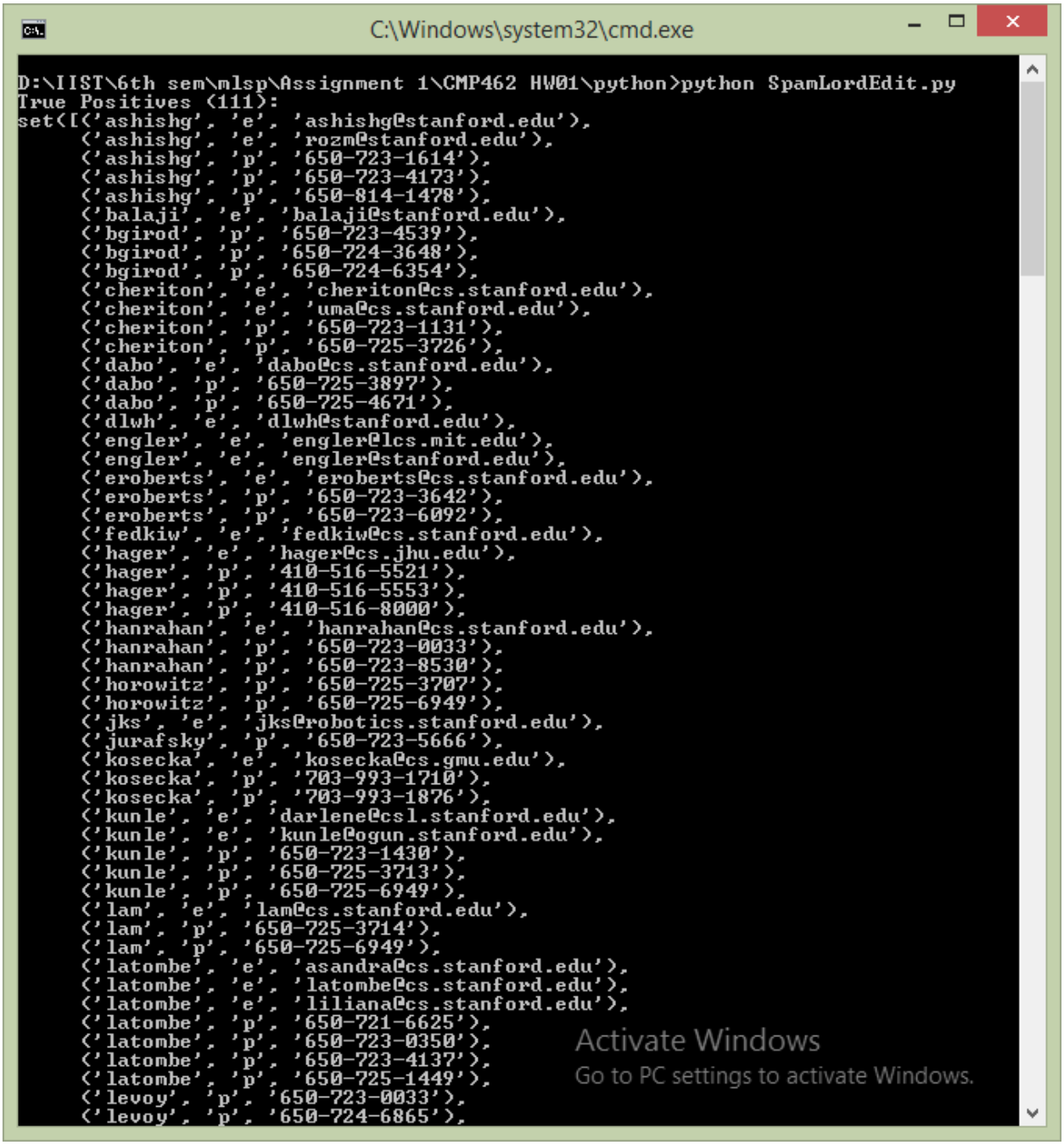
SpamLordEdit.py – the edited version of the above to give desired output. Run this to assess the homework.

The code uses Python 2.7.12 and might not be compatible with Python 3

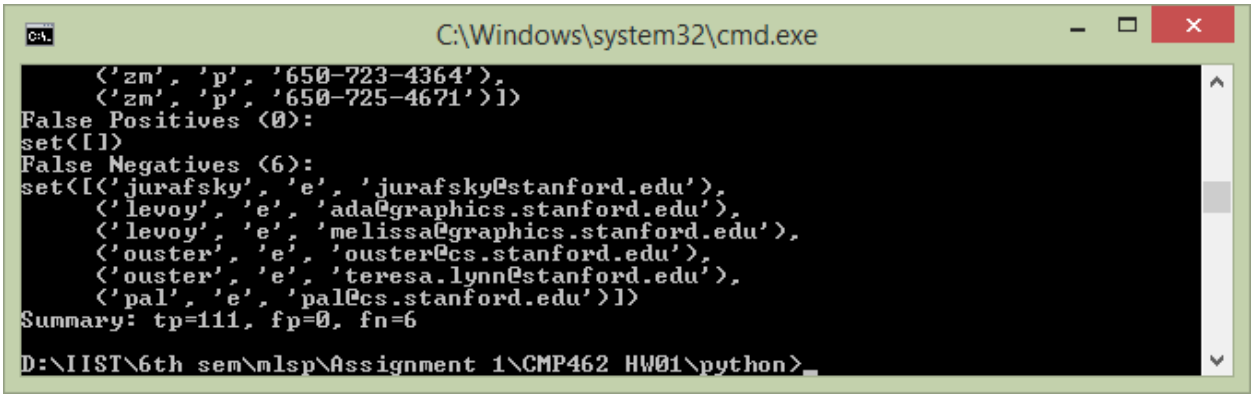
**Source code:**

Look at SpamLordEdit.py in the python directory.

**Output:**


C:\Windows\system32\cmd.exe
D:\IIST\6th sem\mlsp\Assignment 1\CMP462 HW01\python>python SpamLordEdit.py
True Positives (111):
set([('ashishg', 'e', 'ashishg@stanford.edu'),
('ashishg', 'e', 'rozm@stanford.edu'),
('ashishg', 'p', '650-723-1614'),
('ashishg', 'p', '650-723-4173'),
('ashishg', 'p', '650-814-1478'),
('balaji', 'e', 'balaji@stanford.edu'),
('bgirod', 'p', '650-723-4539'),
('bgirod', 'p', '650-724-3648'),
('bgirod', 'p', '650-724-6354'),
('cheriton', 'e', 'cheriton@cs.stanford.edu'),
('cheriton', 'e', 'uma@cs.stanford.edu'),
('cheriton', 'p', '650-723-1131'),
('cheriton', 'p', '650-725-3726'),
('dabo', 'e', 'dabo@cs.stanford.edu'),
('dabo', 'p', '650-725-3897'),
('dabo', 'p', '650-725-4671'),
('dlwh', 'e', 'dlwh@stanford.edu'),
('engler', 'e', 'engler@lcs.mit.edu'),
('engler', 'e', 'engler@stanford.edu'),
('eroberts', 'e', 'eroberts@cs.stanford.edu'),
('eroberts', 'p', '650-723-3642'),
('eroberts', 'p', '650-723-6092'),
('fedkiw', 'e', 'fedkiw@cs.stanford.edu'),
('hager', 'e', 'hager@cs.jhu.edu'),
('hager', 'p', '410-516-5521'),
('hager', 'p', '410-516-5553'),
('hager', 'p', '410-516-8000'),
('hanrahan', 'e', 'hanrahan@cs.stanford.edu'),
('hanrahan', 'p', '650-723-0033'),
('hanrahan', 'p', '650-723-8530'),
('horowitz', 'p', '650-725-3707'),
('horowitz', 'p', '650-725-6949'),
('jks', 'e', 'jks@robotics.stanford.edu'),
('jurafsky', 'p', '650-723-5666'),
('kosecka', 'e', 'kosecka@cs.gmu.edu'),
('kosecka', 'p', '703-993-1710'),
('kosecka', 'p', '703-993-1876'),
('kunle', 'e', 'darlene@csl.stanford.edu'),
('kunle', 'e', 'kunle@ogun.stanford.edu'),
('kunle', 'p', '650-723-1430'),
('kunle', 'p', '650-725-3713'),
('kunle', 'p', '650-725-6949'),
('lam', 'e', 'lam@cs.stanford.edu'),
('lam', 'p', '650-725-3714'),
('lam', 'p', '650-725-6949'),
('latombe', 'e', 'asandra@cs.stanford.edu'),
('latombe', 'e', 'latombe@cs.stanford.edu'),
('latombe', 'e', 'liliana@cs.stanford.edu'),
('latombe', 'p', '650-721-6625'),
('latombe', 'p', '650-723-0350'),
('latombe', 'p', '650-723-4137'),
('latombe', 'p', '650-725-1449'),
('levoy', 'p', '650-723-0033'),
('levoy', 'p', '650-724-6865'),
Activate Windows
Go to PC settings to activate Windows.


C:\Windows\system32\cmd.exe
('levoy', 'p', '650-724-6865'),
('levoy', 'p', '650-725-3724'),
('levoy', 'p', '650-725-4089'),
('manning', 'e', 'dbarros@cs.stanford.edu'),
('manning', 'e', 'manning@cs.stanford.edu'),
('manning', 'p', '650-723-7683'),
('manning', 'p', '650-725-1449'),
('manning', 'p', '650-725-3358'),
('nass', 'e', 'nass@stanford.edu'),
('nass', 'p', '650-723-5499'),
('nass', 'p', '650-725-2472'),
('nick', 'e', 'nick.parlante@cs.stanford.edu'),
('nick', 'p', '650-725-4727'),
('ok', 'p', '650-723-9753'),
('ok', 'p', '650-725-1449'),
('pal', 'p', '650-725-9046'),
('psyoung', 'e', 'patrick.young@stanford.edu'),
('rajeev', 'p', '650-723-4377'),
('rajeev', 'p', '650-723-6045'),
('rajeev', 'p', '650-725-4671'),
('rinard', 'e', 'rinard@lcs.mit.edu'),
('rinard', 'p', '617-253-1221'),
('rinard', 'p', '617-258-6922'),
('serafim', 'e', 'serafim@cs.stanford.edu'),
('serafim', 'p', '650-723-3334'),
('serafim', 'p', '650-725-1449'),
('shoham', 'e', 'shoham@stanford.edu'),
('shoham', 'p', '650-723-3432'),
('shoham', 'p', '650-725-1449'),
('subh', 'e', 'subh@stanford.edu'),
('subh', 'e', 'uma@cs.stanford.edu'),
('subh', 'p', '650-724-1915'),
('subh', 'p', '650-725-3726'),
('subh', 'p', '650-725-6949'),
('thm', 'e', 'pkrokel@stanford.edu'),
('thm', 'p', '650-725-3383'),
('thm', 'p', '650-725-3636'),
('thm', 'p', '650-725-3938'),
('tim', 'p', '650-724-9147'),
('tim', 'p', '650-725-2340'),
('tim', 'p', '650-725-4671'),
('ullman', 'e', 'support@gradiance.com'),
('ullman', 'e', 'ullman@cs.stanford.edu'),
('ullman', 'p', '650-494-8016'),
('ullman', 'p', '650-725-2588'),
('ullman', 'p', '650-725-4802'),
('vladlen', 'e', 'vladlen@stanford.edu'),
('widom', 'e', 'siroker@cs.stanford.edu'),
('widom', 'e', 'widom@cs.stanford.edu'),
('widom', 'p', '650-723-0872'),
('widom', 'p', '650-723-7690'),
('widom', 'p', '650-725-2588'),
('zelenski', 'e', 'zelenski@cs.stanford.edu'),
('zelenski', 'p', '650-723-6092'),
('zelenski', 'p', '650-725-8596'),
('zm', 'e', 'manna@cs.stanford.edu'),
('zm', 'p', '650-723-4364'),




**Summary:**

No of True Positive cases = 111

No of False Positive cases = 0

No of False Negative cases = 6

% Precision = 100%

% Recall = 94.87%

% Accuracy = 97.37%

**Inference:**

All phone numbers were successfully extracted from the text files and processed properly into the desired output format.

All but six emails were successfully extracted from the text files and processed properly into the desired output format.

The program was unable to detect 6 e-mails. This is because I was not able to find them in the original text files (probably they were written in some script format). Hence, I was unable to devise a regex for efficiently extracting them without being able to see their actual format in the file.

Through this assignment, I came to know how powerful regex are and the essential role they may play in file processing and information extraction. They are a very dense form of code and the programmer needs to give in a lot of thought in order to build a regex for performing even simple tasks.