Project Description:

This project tries to cover the knowledge of Python, spaCy library and text analysis. We want to protect sensitive information, so we need to redact text to hide the information we want.

The main steps are text analyzing and redacting data. The functions will be written in a python file named redactor.py. We will write test files about the functions from project1, and we will write a main.py to use all the functions.

Text analyzing. We will use spaCy library to do text analyzing. The main types we need to analyze are names, genders, dates, phones and address. For names, we directly use en\_core\_web\_sm model, which is pretrained for spaCy, and we use the label PERSON for it. For genders, we use a static gender pattern, that is use the words we frequently use for describing gender like him, we e.t. For dates, we use the DATE and CARDINAL labels build from the pretrained model. For phones, we use re library instead of spaCy. We use a phone pattern like (xxx)-xxx-xxxx to be the phone format. We use a phone pattern as r"[(]?\d{3}[)\s]?[-.\s]?\d{3}[-.\s]?\d{4}" to judge the phone numbwe. For address, we use the GPE and LOC labels build from the pretrained model.

Redacting data. We use replace method build-in Python str to replace the word we need to redact.

How to run

pipenv run python project1/main.py --input 'data/\*txt' --names --dates --phones --genders --address --output 'files/' --stats stderr

Data description

We extract three folders from the dataset, that is allen-p, arnold-j and arora-h. For simple test, we create three text files named 1.txt, 2.txt and 3.txt.

Functions

redactor.py \

replace\_text()- this function has a doc parameter to indicate the text we need to replace, its type is nlp. The second parameter has a matches parameter for the match result, we can get match result to get what content we need to replace. The third parameter is the character we need to do the replacement. It will return the replaced text.

phone\_encrypt ()- this function redact phone number in text to replacement character. The first parameter is text, its type is str. The second parameter is the character we need to do the replacement. The phone pattern is r"[(]?\d{3}[)\s]?[-.\s]?\d{3}[-.\s]?\d{4}". It will return the redacted text.

gender\_encrypt()-this function redact phone number in text to replacement character. The first parameter is text, its type is str. The second parameter is the character we need to do the replacement. The gender pattern is [{"LOWER": {"IN": ["aunt", "baby", "brother", "boyfriend", "bride", "cousin", "child", "dad", "daughter", "father", "father-in-law", "fiancé",

"fiancée", "friend", "girlfriend", "godchild", "godfather", "godmother", "grandchild", "grandchildren", "granddaughter", "grandfather", "granddad", "grandpa", "grandparent",

"grandparents", "grandmother", "grandma", "great-grandparents", "groom", "half-brother", "husband", "mother", "mother-in-law", "mum", "mummy", "mom", "nephew", "niece",

"parent", "parents", "sister", "son", "stepbrother", "twin", "twin-brother", "uncle", "wife",

"he", "his", "him", "himself", "she", "her", "hers", "herself", "I", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "they", "them", "their", "themselves"]}}]. It will return the redacted text.

address\_encrypt()- this function redact phone number in text to replacement character. The first parameter is text, its type is str. The second parameter is the character we need to do the replacement. We use the build-in label GPE and LOC as address in en\_core\_web\_sm model. It will return the redacted text.

date\_encrypt ()-this function redact phone number in text to replacement character. The first parameter is text, its type is str. The second parameter is the character we need to do the replacement. We use the build-in label DATE and CARDINAL as address in en\_core\_web\_sm model. It will return the redacted text.

name\_encrypt ()- this function redact phone number in text to replacement character. The first parameter is text, its type is str. The second parameter is the character we need to do the replacement. We use the build-in label PERSON as address in en\_core\_web\_sm model. It will return the redacted text.

encode()-this function is to redact file. The first parameter is filename, the second parameter is a config dict. We use the config dict to judge which field to redact and write the result to output file.

main.py \

parse\_config()- this function parse the command line to get config we need to deal with. It will return a config dict.

test\_download.py \

test\_download()- this function has a url parameter to indicate where to download data, this function will test fetchincidents function from project0.py.

test\_address.py \

this file is to test the address\_encrypt function from redactor.py.

test\_dates.py \

this file is to test the date\_encrypt function from redactor.py.

test\_genders.py \

this file is to test the gender \_encrypt function from redactor.py.

test\_names.py \

this file is to test the name\_encrypt function from redactor.py.

test\_phones.py \

this file is to test the phone\_encrypt function from redactor.py.

Bugs and Assumptions

Bugs:

The en\_core\_web\_sm model can’t do the text analysis perfectly. So the result of redacting is no so well as we want.

Assumptions:

As for text analysis, we directly use en\_core\_web\_sm model in spaCy as basic model. We use its build-in PERSON label to judge names. As for address, we use its build-in LOC label and GPE to judge it. For dates, we use its build-in DATE label and CARDINAL to judge it. For phone number, we use re library and create a re pattern to judge it, we think phone number format will be (xxx)-xxx-xxxx or xxx-xxx-xxxx. For genders, we use explicit words to judge it. That includes some common relative words. For redacting data, we directly use replace method in str type to replace the word we need to replace.

Output format

In config parameter, we use stats parameter to indicate where we output the information when we are running our program. We will output config parameters and the filename we will redact when we are running our program.