Assignment 1

March 20, 2021

Part A:Regex

1 Description:

This part you are required to write Regex patterns to match different email formats and phone numbers. The formats can be as the following:

1. `abc.yxz@stanford.edu`
2. `abc123@stanford.edu`
3. `abc @ stanford.cu.edu`
4. `abc at cs dot stanford dot edu`
5. `abc at stanford cu edu`
6. `abc at stanford.cu.edu`
7. `a-b-c-@-s-t-a-n-f-o-r-d.-e-d-u`
8. `(123) 444-8888`
9. `(123)444-8888`
10. `123-444-8888`
11. `123 444 8888`

- You are supposed to write two Regex patterns for the email part and two for the Phone Numbers. Please pay attention to the white spaces as its crucial as its calculated as a character
- You will find the assignment.zip file divided into two folders , data folder and code folder. You are supposed to modify the code folder and work at that directory.
- You are adviced to try https://regex101.com/ for trials.
- You are supposed to change `def process_file(name,f):` so that it returns list of emails and phone numbers. It is advised to use `match.extend(re.findall(additional_pattern,line)` if you want to use mutliple regex in the same time.
- Example: If we want to find any 2 letters followed by - and 3 letters followed by *:
```
match = re.findall(r'[a-z]{2}[\-]',line)
match.extend(re.findall(r'[a-z]{3}[\*]',line)
```
- This will be using the two regexes in the same time to return your data.

- The assessment is automatic using the score function you should achieve 0 False negative and 0 False positives.

Part B:Search Engine Query Parser

2 Description:Help the historian:

In this assignment you are required to help Mr.Ahmed the historian with a simple search engine to help him search in some documents about famous characters from the islamic golden age.The documents and the their index file are provided you are required to implement a query parser using the NLP techniques discussed in the tutorials and the lecture to parse the input query and retrieve the matched document.

3 Input Description

When you pull this assignment from the github course repo you will have the following:

- **10.txt files numbered from 0 to 9**
- **pickle file with name index.pkl,saving the index in dictionary format as follows: index[word]=(doc number,line number)**

4 Requirement

you are required to parse the following queries and return the matched documents.

queries

1. **" Who were famous women?"**
2. **"Poems and poets"**
3. **"What is the scientific work of islamic golden age? "**
4. **"Ancient Spain"**
5. **"A Famous Muslim Engineer"**
6. **"A woman with a kingdom"**
7. **"How were the prayer times determined?"**
8. **"A theologian"**

5 Assessment

After processing the following queries and retrieving the required documents answer the following questions:

1. **which query returned in exactly one document?**
2. **which query returned almost all the documents?**

3. **which queries returned around 80% relevant results(more than one)?**
4. **which queries returned somehow irrelevant results?**
5. **why do you think some queries were successful in returning relevant results and others are not?**
6. **if you read any of the documents which character was totally,new for you,what new information did you get?**