PYTHON ADVANCED PROGRAMMING

Chitra G M and P Rama Devi
Department of Computer Science Engineering

PYTHON ADVANCED PROGRAMMING

P Rama Devi

Department of Computer Science and Engineering

PYTHON ADVANCED PROGRAMMING

Groups

Searching for pattern matches is the basis of the powerful capabilities provided by regular expressions. Adding groups to a pattern isolates parts of the matching text, expanding those capabilities to create a parser. Groups are defined by enclosing patterns in parentheses ((and)).

For the names example we were seeing
If the pattern is written as re.compile(r'M(r|s|rs)\.?\s[A-Z]\w*')

The above will print all the names we were discussing

re.compile(r'(Mr|Ms|Mrs)\.?\s[A-Z]\w*')

PYTHON ADVANCED PROGRAMMING

Another Example:

```
import re
text='''
Mr. Sachin
Mr Virat
Mrs Karanam
Ms Sindhu
Mr. R

'''

pattern = re.compile(r'(Mr|Ms|Mrs)\.?\s[A-Z]\w*')
matches=pattern.finditer(text)
for match in matches:
    print(match)
```

```python
emails='''
ramadevi@gmail.com
rama.devi@pes.edu
p-rama-devi@my-work.net
'''

pattern = re.compile(r'[a-zA-Z.]+@[a-zA-Z]+\.(com|edu)')
matches=pattern.finditer(emails)
for match in matches:
    print(match)
```

Output:

ramadevi@gmail.com
rama.devi@pes.edu
p-rama-devi@my-work.net

urls=
'''

https://www.google.com
http://ramadevi.com
https://pep.com
https://www.isro.gov

Question: we want to capture domain name and top level domain name

Example: google.com

If we use re.compile(r'https?://(www\.)?\w+\.\w+')

Output all 4 will be matched. Lets see how to get the answer for the question

```
pattern=re.compile(r'https?://(www\.)?(\w+)(\.\w+)')
matches=pattern.finditer(urls)
for match in matches:
      print(match.group(2))
```

Output will be for domain

match.group(3) will give top level domain

match.group(0)

Will give all the matches

We can use back references for the previous example called
as substitution

```
pattern=re.compile(r'https?://(www\.)?(\w+)(\.\w+)')
subbed_urls=pattern.sub(r'\2\3',urls)
print(subbed_urls)
```

Here \2,\3 are back references
i.e domain and top level domain information is printed

THANK YOU

Chitra G M and P Rama Devi
Department of CSE

pramadevi@pes.edu
chitragm@pes.edu