「正規表現による文字列処理」演習および課題

名古屋大学 情報基盤センター
情報基盤ネットワーク研究部門
基盤ネットワーク研究グループ
嶋田 創

演習1

英字と数字からなるパスワード用文字列を入力させ、正規表現などで以下を判定してOK/NGを返すプログラム書け

- 「大文字」「小文字」「数字」の3種類が必ず含まれる
- 文字数が8文字以上
  - 正規表現ではなく、「len(文字列変数)」を利用した方が楽ではある

発展

- NG時に「何が足りないか」を明示する
- 記号もパスワードに入れさせるようにする
  - エスケープ処理しないと、エラーの原因となる記号がある点に注意
- Worst passwords top 100に含まれているパスワードを入力したら、さらに警告する(データは自分で探すこと)

演習2

正規表現を用いて、sample_auth_log2.txtから不正アクセスを試みたユーザ名を切り出して表示せよ

- ユーザ名は以下の書式のxxxの部分に記されるので、それを切り出すこと

  ...Invalid User xxx from ...

  - 正規表現でxxxの部分を切り出す表現を書けばOK

発展

- 不正アクセスを行ったユーザ名とその出現回数の表示

課題

sample_auth_log2.txtから、以下の情報をそれぞれ表示せよ

1. 不正アクセスを試みてきたIPアドレスの重複の無いリスト
   - Python上では「重複のないリスト = セット」なので、セットに投げ込んであげればOK
2. IPアドレスの第1オクテット、第2オクテット、第3オクテット、第4オクテットの重複のないリスト
   - 復習: IPアドレス(例: 133.6.90.249)のピリオドで区切った数字の左から右へそれぞれ第1～第4オクテット(例の第1オクテット: 133, 例の第2オクテット: 6, ...)

発展

- 1., 2.のそれぞれにおける、各要素の出現回数の表示

最終課題

不正アクセスの傾向の解析の基礎データを作るために、sample_auth_log2.txtから、以下の情報をまとめて表示せよ

1. アクセスを試みてきたIPアドレスとその出現数
2. アクセスを試みてきたユーザ名とその出現数
3. 各日(May 24, ..., Jun 4)におけるアクセスの数の合計
4. 毎時(0時台, ..., 23時台)におけるアクセスの数の合計
   - 各日のx時台のアクセス数を全て合計したもの

- どの数字が何を表しているか分かるように補足する文字も同時に出力すること(過去に数字だけ提示した人がいたので)
- 余裕があれば、出現率を(%とかで)併記するのもあり