C2184 Úvod do programování v Pythonu

Ukázka závěrečného testu

Úkolem tohoto testu je napsat skript pro zpracování souboru molekul a výstupů z Gaussianu.

1. Ze studijních materiálu si stáhněte `dataset_pka.zip`. Soubor můžete rozbalit pomocí libovolného nástroje, pokud použijete knihovnu v Pythonu, získate bonusové body.

2. Složka obsahuje dva typy souborů: `*.sdf` soubory obsahují informace o molekule (vaším úkolem je extrahovat hodnoty pKa) a soubory `*.log`, které obsahují informace o nábojích a vaším cílem je vyextrahovat maximální náboj na vodíku (H).
   pKa:

   ```
   > <E_NAME>
   NSC 3

   > <PKA>
   2.1

   > <E_STEREO_SPECIFIED>
   no_stereocenter
   ```

   q(max,H):

   ```
    Mulliken charges:
                   1
        1  F   -0.283823
        2  C   -0.167968
        3  C    0.297777
        4  C   -0.204856
        5  C    0.289886
        6  C    0.406838
        7  C   -0.246716
        8  N    0.358053
        9  O   -0.380521
       10  O   -0.397823
       11  O   -0.616303
       12  H    0.191590
       13  H    0.163136
       14  H    0.168775
       15  H    0.421955
    Sum of Mulliken charges =   0.00000
   ```

3. Spočítejte korelaci mezi **pKa** (y) a **q(max,H)** (x) podle [1]

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2)}} \tag{1}$$

[1] $\bar{x}$ ... průměr hodnot x, obdobně $\bar{y}$

4. Vypočíjte lineární regresi mezi **pKa** (y) a **q(max,H)** (x) podle[2]

$$y = ax + b \tag{2}$$

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \tag{3}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \tag{4}$$

5. Pomocí parametru přímky a, b predikujte novou hodnotu pKa (**predPKA**) a vypočítejte RMSE podle

$$e_i = \mathrm{pKa}_i - \mathrm{predPKA}_i \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum e_i^2} \tag{6}$$

6. Výsledky uložte do csv souboru s touto hlavičkou:

```
Molecule;q(max,H);pKa;predPKA;e
```

Hodnotící tabulka:

	max.	Hodnocení
Celková funkčnost		
Práce se soubory (čtení+zápis)	7+7	
Extrakce dat	10	
Analýza dat	10	
Obecné aspekty		
Vhodné použití komentářů	10	
Vhodné použití výjimek (hodnoceny max. 2 výskyty)	7+7	
Styl kódu a znovupoužitelnost[3]	12	
Bonusy		
Celkem	70	

[2] n ... počet prvků

[3] přehlednost kódu, srozumitelnost proměnných, nutnost upravovat kód pro jiný datový set, získavání informací od uživatele, ...