PROJEKT H 19. 12.2018

TERMIN ROZLICZENIA 31.12.2018

ALGORYTMIKA: PRACA NA DANYCH STABELARYZOWANYCH

PYTHON:

MODUŁ PANDA.PY

DANE STABELARYZOWANE DATA FRAME

https://pandas.pydata.org/pandas-docs/stable/10min.html

https://pandas.pydata.org/pandas-docs/stable/cookbook.html#cookbook

DataFrame : elementarna obsługa

Dane w postaci tabeli 2D o swobodnej zawartości

```
import pandas as pd

dane= [['ABBOT', 427, 448],
       ['ABER', 534, 600]]
```

Dodajemy nagłówki kolumn

```
dane_df= pd.DataFrame(dane, columns=['firma', 2015, 2016])
```

1_csv_0.py

```
print('rozmiar danych', dane_df.shape)
print('naglowki:', dane_df.columns)
```

Wyświetlanie zawartości

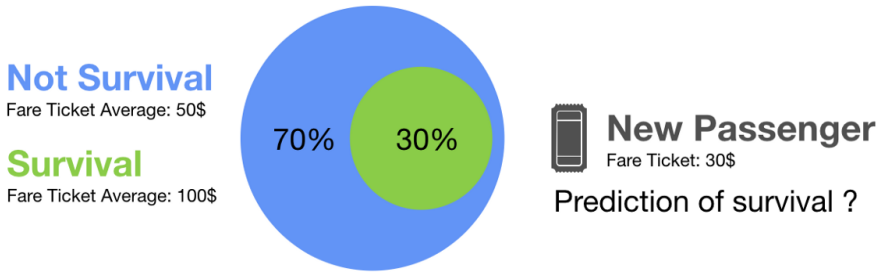
```
print(dane_df)

print(dane_df.ix[:,1])
print(dane_df.ix[1,:])
print(dane_df)
```

0_co_to_DataFrame.py

Titanic problem

Popracujemy na słynnych danych opisujących pasażerów statku Titanic w jego pierwszym i ostatnim rejsie. Dane te zawierają zarówno informacje osobiste, jaki czy dana osoba przeżyła.

Ciekawe jest, jak cena jaką pasażer zapłacił za bilet wpłynęła na przeżycia katastrofy.



https://blog.sicara.com/naive-bayes-classifier-sklearn-python-example-tips-42d100429e44

csv w Pythonie

nagłówki

rekordy pasażerów

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S

```
import pandas as pd
data = pd.read_csv('titanic.csv')
```

1_csv_2.py

Warto przećwiczyć

```
data.shape
data.columns
```

Oglądanie danych:

```
data.head()
data.tail()
```

Symbolizacja danych i czyszczenie danych

Zamieniamy opis na symbole:

```
data["Sex_cleaned"]=np.where(data["Sex"]=="male",0,1)
```

1_csv_3.py

Czyścimy dane:

Metoda *where* z numpy to wygodny rodzaj warunkowego przydziału wartości

```
data_poprawne=data[[
  "Survived",
  "Pclass",
  "Sex_cleaned",
  "Age",
  "SibSp",
  "Parch",
  "Fare",
  "Embarked_cleaned"
]].dropna(axis=0, how='any')
```

Pomijamy wiersze, w których choć raz pojawiło się *nan* w wybranych kolumnach

pandas.DataFrame.dropna

Metoda *dropna()* zmiennej typu DataFrame z bibliotek pandas

1_csv_4.py

Titanic problem

PROJEKT H

Na dziś:

- Wczytać tabelę z pliku *titanic.csv*.
- Przeprowadzić symbolizację wartości kategorycznych.
- Usunąć rekordy o niepełnej informacji w badanych kolumnach.

- Przygotować zestawienia (histogramy i tabele):
  - ❑ rozkład wieku wśród tych co przeżyli i tych co nie przeżyli
  - ❑ rozkład cen biletów wśród tych co przeżyli i tych co nie przeżyli

Na później:

- Wczytać tabele z plików *Ada_BP.txt* i *Ada_RR.txt*
- Wyciągnąć czas z pliku Ada_BP.txt .
- Zestawić zdarzenia w jedną tabele uporządkowaną czasem