

Som príliš často na sociálnej
sieti, čo hovoria dáta ?

Jozef Kiselák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiselák, ÚMAT, PF UPJŠ
DATADAY 4

17.2.2024

Najprv si stiahneme dáta z facebooku:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

Ak nemáte facebook:

súbor: your_posts_1.json

✓ <https://github.com/kis81/dataday3>

✓ adresár: c:\...\Desktop\dataday4

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnite si svoje dáta

Početnosť

Empirická pravdepodobnosť


Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

JavaScript Object Notation (JavaScriptový objektový zápis, **JSON**) je spôsob zápisu dát (dátový formát) nezávislý na počítačovej platforme, určený na prenos dát, ktoré môžu byť organizované v poliach alebo agregované v objektoch.

Vstupom

je ľubovoľná dátová štruktúra (číslo, reťazec, boolean, objekt alebo z nich zložené pole), výstupom je vždy reťazec. Zložitosť hierarchie vstupnej premennej nie je teoreticky nijako obmedzená.

JavaScript Object Notation	
	
Filename extension	.json
Internet media type	application/json
Type code	TEXT
Uniform Type Identifier (UTI)	public.json
Type of format	Data interchange
Extended from	JavaScript
Standard	STD 90 ↗ (RFC 8259 ↗), ECMA-404 ↗ , ISO/IEC 21778:2017 ↗
Open format?	Yes
Website	json.org ↗

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

The screenshot shows the Facebook 'Download Your Information' interface. The browser address bar displays the URL `facebook.com/dyi/?referrer=yfi_settings`. The page title is 'Download Your Information'. The main content area is divided into two columns. The left column contains the following text: 'You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it. Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days. If you'd like to view your information without downloading it, you can [Access Your Information](#) at any time.' The right column has a 'Downloads' section with a 'Request a download' link. Below this is the 'Select file options' section, which includes three dropdown menus: 'Format' (set to HTML), 'Media quality' (set to High), and 'Date range (required)'. At the bottom of the right column is the 'Select information to download' section, which states 'You can download everything or select the types of information you want to download.' and 'Your activity across Facebook' (with a 'Deselect' link). The page footer shows a progress bar at 4/24 and navigation buttons for 'Back' and 'Forward'.

Activities Google Chrome Mar 13 19:27 sk

dataday - Online LaTeX E Do You Post Too Much? (10) Facebook

facebook.com/dyi/?referrer=yfi_settings

Search Facebook

Download Your Information

You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.

Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.

If you'd like to view your information without downloading it, you can [Access Your Information](#) at any time.

When you download your information your download won't include information that someone else shared, like another person's photos that you're tagged in.

You can still [view this information anytime](#).

Downloads

[Request a download](#) Available to you

Select file options

You can choose the file format, media quality and date range for your download. HTML format is easy to view, while JSON format allows another service to more easily import the file. Media quality is the quality of your photos and videos but also affects file size.

Format: **HTML**

Media quality: **High**

Date range (required):

Select information to download

You can download everything or select the types of information you want to download.

Your activity across Facebook

Information and activity from different areas of Facebook, such as posts you've created, photos you're tagged in, groups you belong to and more.

4/24

[Back](#) [Forward](#)

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

The screenshot shows the Facebook 'Download Your Information' page. The browser window at the top shows the URL https://www.facebook.com/dyi/?referrer=yfi_settings. The page content includes:

- Download Your Information**: A section explaining that users can download a copy of their Facebook information at any time, in either HTML or JSON format. It notes that JSON is easier to use with other services but affects file size.
- Select file options**: A dropdown menu for 'Format' with 'JSON' selected. Other options visible are 'HTML' and 'JSON' (with a checkmark). There is also a 'Date range (required)' dropdown.
- Select information to download**: A section explaining that users can download everything or select specific types of information.
- Your activity across Facebook**: A section explaining that users can download information and activity from different areas of Facebook, such as posts, photos, and tagged items.

At the bottom of the page, there is a progress bar showing '5/24' and two buttons: 'Back' and 'Forward'.

Som príliš často na sociálnej sieti, čo hovoria dáta?

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnite si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

Stiahnite si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická pravdepodobnosť v strojovom učení?

Activities Google Chrome Mar 13 19:39 en

dataday - Online LaTeX E Do You Post Too Much? (12) Facebook

facebook.com/dyi/?referrer=yfi_settings

Search Facebook

Download Your Information

You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.

Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.

If you'd like to view your information without downloading it, you can [Access Your Information](#) at any time.

When you download your information your download won't include information that someone else shared, like another person's photos that you're tagged in.

You can still view this information anytime.

Downloads

[Request a download](#) Available to you

- Last week
- Last month
- Last 3 months
- Last 6 months
- Last year
- Last 3 years
- All time
- Custom
- Date range (required)

Select information to download

You can download everything or select the types of information you want to download.

Your activity across Facebook

Information and activity from different areas of Facebook, such as posts you've created, photos you're tagged in, groups you belong to and more

6/24

Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

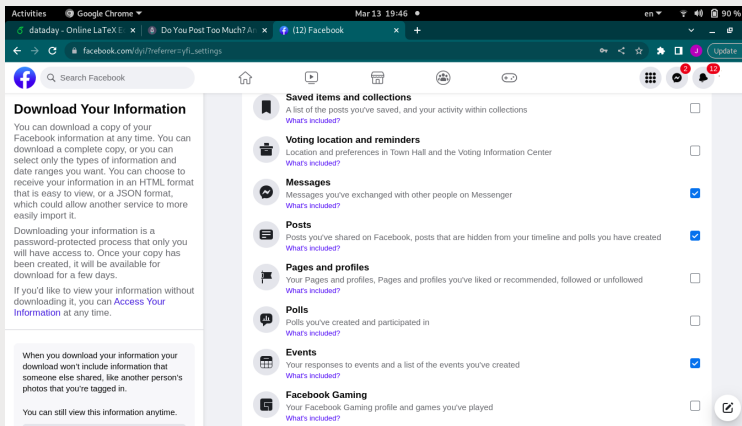
Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings



Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

The screenshot shows the Facebook 'Download Your Information' page. The browser address bar shows the URL https://www.facebook.com/dyi/?referrer=yfi_settings. The page content includes:

- Download Your Information**: A section explaining that users can download a copy of their Facebook information at any time, including a complete copy or specific types of information and date ranges. It also mentions that the download is a password-protected process and that users can view their information without downloading it by clicking 'Access Your Information'.
- Preferences**: A section titled 'Actions you've taken to customize your experience on Facebook' with checkboxes for 'Feed' and 'Preferences'.
- Ads information**: A section titled 'Your interactions with ads and advertisers on Facebook' with a checkbox for 'Ads information'.
- Start your download**: A section titled 'Your download may contain private information. You should keep it secure and take precautions when storing it, sending it or uploading it to another service.' with a 'Request a download' button.

8/24

Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnite si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická pravdepodobnosť v strojovom učení?

Stiahnite si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings

Activities Google Chrome Mar 13 19:59 en 80 %

dataday - Online LaTeX Do You Post Too Much? A (11) Facebook JSON - Wikipedia

facebook.com/dyi/?referrer=yfi_settings

Search Facebook

Download Your Information

You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.

Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.

If you'd like to view your information without downloading it, you can [Access Your Information](#) at any time.

Downloads

Request a download Available files

Sep 12, 2022 - Mar 13, 2023
Posts, Events and Messages (1.19 GB)
Requested on Mar 13 at 7:46 PM
JSON format
High-quality media
1 file
Expires Mar 17, 2023

Download ...

When you download your information your download won't include information that someone else shared, like another person's photos that you're tagged in.

You can still view this information anytime.

Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

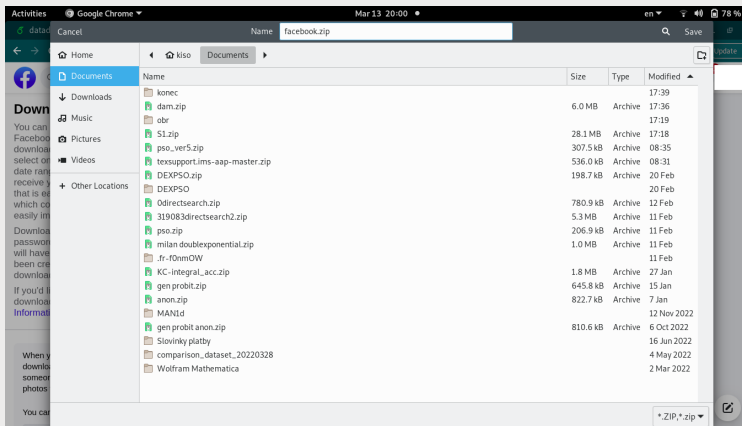
Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:

✓ https://www.facebook.com/dyi/?referrer=yfi_settings



Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

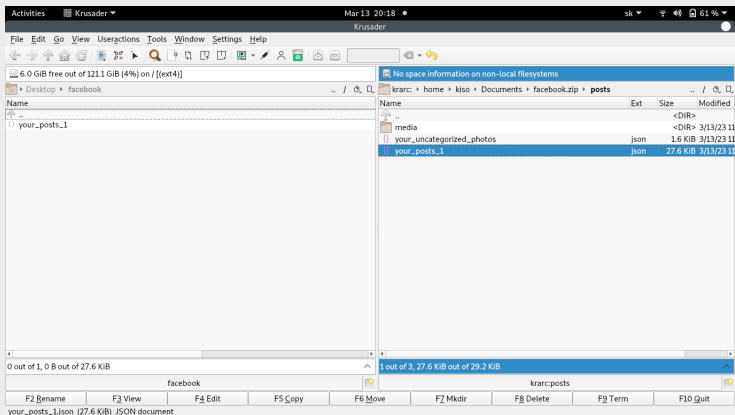
Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Stiahnime si svoje dáta:



11/24

Som príliš často na sociálnej sieti, čo hovoria dáta?

Back

Forward

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

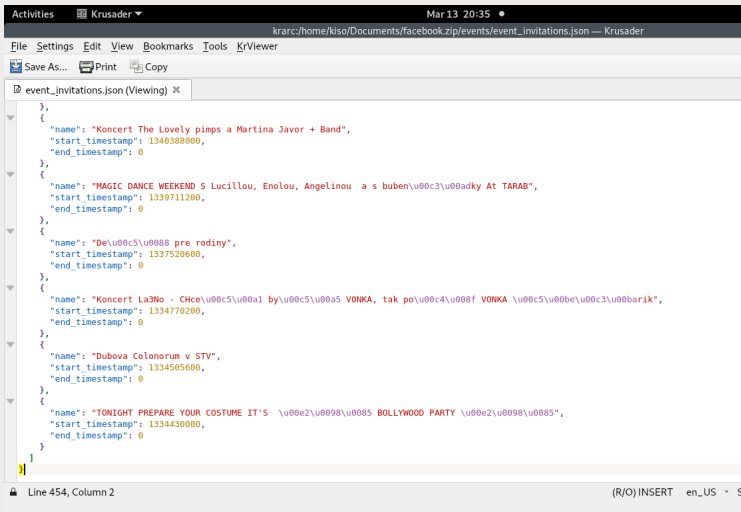
Stiahnite si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?



```
Activities  Krusader  Mar 13 20:35
krarc:/home/kiso/Documents/facebook.zip/events/event_invitations.json — Krusader
File  Settings  Edit  View  Bookmarks  Tools  KrViewer
Save As...  Print  Copy
event_invitations.json (Viewing) X
{
  {
    "name": "Koncert The Lovely pimps a Martina Javor + Band",
    "start_timestamp": 1340388000,
    "end_timestamp": 0
  },
  {
    "name": "MAGIC DANCE WEEKEND S Lucillou, Enolou, Angelinou  a s buben\u00c3\u00adky At TARAB",
    "start_timestamp": 1339711200,
    "end_timestamp": 0
  },
  {
    "name": "De\u00c5\u0088 pre rodiny",
    "start_timestamp": 1337520600,
    "end_timestamp": 0
  },
  {
    "name": "Koncert La3No - CHe\u00c5\u00a1 by\u00c5\u00a5 VONKA, tak po\u00c4\u008f VONKA \u00c5\u00be\u00c3\u00barik",
    "start_timestamp": 1334770200,
    "end_timestamp": 0
  },
  {
    "name": "Dubova Colonorum v STV",
    "start_timestamp": 1334505600,
    "end_timestamp": 0
  },
  {
    "name": "TONIGHT PREPARE YOUR COSTUME IT'S \u00e2\u0098\u0085 BOLLYWOOD PARTY \u00e2\u0098\u0085",
    "start_timestamp": 1334430000,
    "end_timestamp": 0
  }
}
Line 454, Column 2  (R/O) INSERT  en_US  S
```

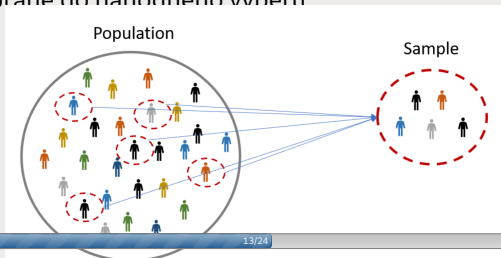
Som príliš často na sociálnej sieti, čo hovoria dáta ?

Back

Forward

Niekoľko základných pojmov.

- **základný súbor (populácia):** množina objektov (všetkých štatistických jednotiek), ktoré sú predmetom nášho záujmu a o ktorých sa majú robiť závery.
- **výberový súbor (náhodný výber/vzorka):** výber určitej veľkosti n zo základného súboru veľkosti N (t.j. podmnožina základného súboru).
- **reprezentatívnosť náhodného výberu :** všetky štatistické jednotky z celej populácie majú rovnakú šancu byť vybrané do náhodného výberu.



Základné typy (diskrétnych) premenných.

Ak NP X môže nadobúdať hodnotu $x_i \in \{kat_2, \dots, kat_p\}$:

- 1 **kvalitatívne (kategoriálne)** - popisujú vlastnosti štatistickej jednotky slovne a nemusia sa dať jednoznačne merať; v závislosti na tom, či tieto hodnoty vieme usporiadať ich ešte delíme na:
 - a) **nominálne** - jednotlivé kategórie síce vieme pomenovať, ale nie usporiadať; nedajú sa robiť žiadne matematické operácie okrem určovania početností; napr. farba očí, značka auta, ...
 - b) **ordinálne** - jednotlivé kategórie vieme aj pomenovať aj usporiadať; ani tu však nemožno robiť matematické operácie. Napr. počet valcov motora auta, kategórie odpovedí: určite áno - skôr áno - skôr nie - určite nie.

Početnosti (frekvencie)

Dáta získané napr. pri náhodnom výbere chceme redukovať, zhrnúť do niekoľkých číselných charakteristík, ktoré nazývame štatistiky.

Ak máme iba konečne (spočítateľne veľa) hodnôt).

- **Absolútna početnosť** n_i udáva, koľkokrát sa príslušná hodnota (znak) x_i v súbore vyskytla.
- **Relatívna početnosť** $f_i = \frac{n_i}{N} = \frac{n_i}{\sum_{i=1}^k n_i}$ je pomer absolútnej početnosti a celkového počtu pozorovaní vo výbere - početnosť relatívna vzhľadom k celkovému počtu prvkov výberu.
- POZOR percentuálne vyjadrenie nemusí priamo implikovať relatívnu početnosť.

Empirická pravdepodobnosť

- Uvažujme "náhodný pokus", ktorý má práve N možných výsledkov.
- Množina všetkých možných výsledkov je výberový priestor.
- Tento pokus môžeme opakovať a teda **pravdepodobnosť** intuitívne chápeme ako relatívnu početnosť v nekonečne veľkom počte pokusov.
- Každú podmnožinu výberového priestoru nazveme (náhodný) jav.
- Jednoprvkové podmnožiny nazývame tiež elementárne javy.
- Nech $x_1, \dots, x_n \in \mathbb{R}$. Hovoríme, že náhodná premenná X má empirické rozdelenie určené týmito dátami, ak je to diskrétna náhodná premenná a $P[X = x_i] = \frac{n_i}{N} = f_i$.

Histogram

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

- **Histogram** je grafické znázornenie rozdelenia dát pomocou stĺpcov (nie nutne) rovnakej šírky (vyjadruje šírku neprekrývajúcich sa po sebe nasledujúcich intervalov - tried), pričom výška stĺpcov vyjadruje početnosť sledovanej veličiny (v danom intervale).
- Všimnite si, že graf absolútnych i relatívnych početností musí vyzeráť rovnako, mení sa len mierka na vertikálnej osi.
- Pruhy v histograme však nie je možné preskupiť.
- Pre **ordinálne** dáta, nie je vhodné hovoriť o histograme.

Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť

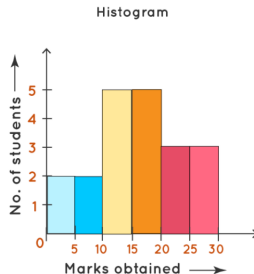
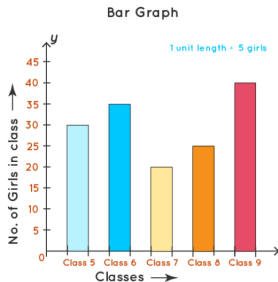
Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Histogram

Pozor na rozdiel medzi stĺpcovým grafom a histogramom!!!

Difference Between Bar Chart and Histogram



Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

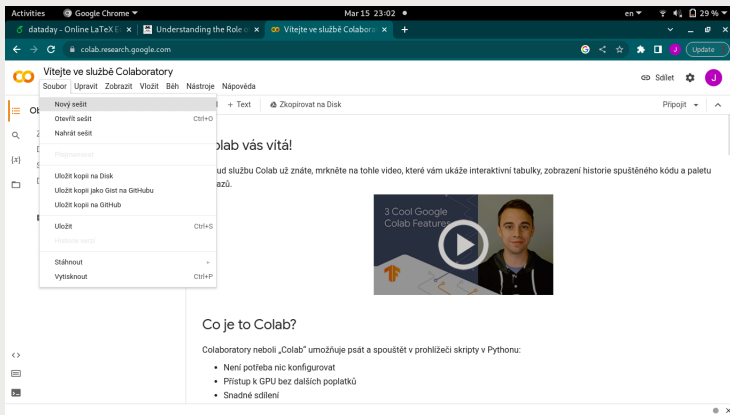
Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická pravdepodobnosť v strojvom učení?

Postujeme príliš často?

- ✓ <https://colab.research.google.com/>
- ✓ Súbor: *fcbook.py*



19/24

Som príliš často na sociálnej sieti, čo hovoria dáta ?

◀ Back

Forward ▶

Som príliš často na sociálnej sieti, čo hovoria dáta?

Jozef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

Empirická pravdepodobnosť


Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Postujeme príliš často?

✓ <https://colab.research.google.com/>

✓ Súbor: **fcbook.py**



```
Activities  Krusader  Mar 15 23:16  en  15%
file:///home/kisoi/Documents/fcbook.py — Krusader

File  Settings  Edit  View  Bookmarks  Tools  Kr/Viewer

Save As...  Print  Copy

fcbook.py (Viewing) X

import pandas as pd # for data manipulation and analysis, e.g. it offers data structures and operations for manipulating numerical tables and time series.
import matplotlib.pyplot as plt # for creating static, animated, and interactive visualizations
import seaborn as sns # data visualization library based on matplotlib, it provides a high-level interface for drawing attractive and informative statistical
graphics
import numpy as np # is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large
collection of high-level mathematical functions to operate on these arrays

# read the json file into a dataframe
df = pd.read_json('your_posts_1.json')
#event = pd.read_json('event_invitations.json')

df.head(3)
event.tail(3)

# rename the timestamp column
df.rename(columns={'timestamp': 'date'}, inplace=True)

#drop some unnecessary columns
df = df.drop(['attachments', 'title', 'tags'], axis=1)

#print(df)

# making sure it's datetime format
df['date'] = pd.to_datetime(df['date'])

df.head(3)
print(df.shape)
df.tail(3)

#Set the date column as the index of our DataFrame
#Select the column we want to resample by - in this case, is the date column.
#Use the .resample() function with the argument 'MS' (for "Month Start") to resample our data by month.
#Use .size() to specify what we want to measure each month - in this case, the number of rows. If a ...next date that fall within that month.

Line 1, Column 1  (R/O) INSERT  en_US  Soft Tabs: 4  UTF-8  Python
```


Som príliš často na sociálnej sieti, čo hovoria dáta ?

Jožef Kiseľák, ÚMAT, PF
UPJŠ
DATADAY 4

Stiahnime si svoje dáta

Početnosť

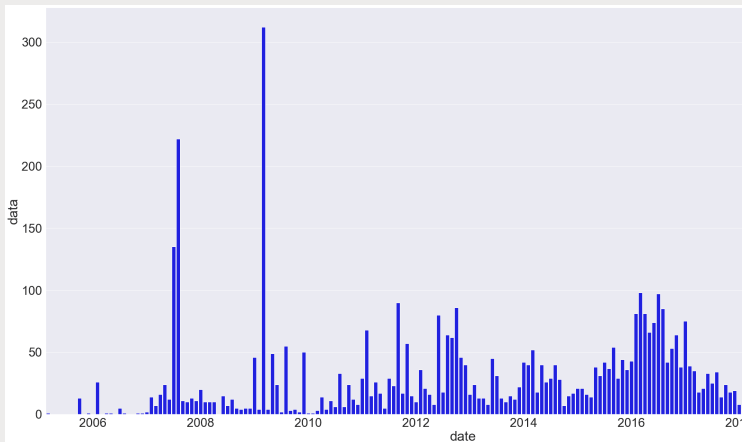
Empirická pravdepodobnosť

Postujeme príliš často?

Ako funguje empirická
pravdepodobnosť v
strojovom učení?

Postujeme príliš často?

- ✓ <https://www.dataquest.io/blog/analyze-facebook-data-python/>
- ✓ <https://www.dataquest.io/blog/how-much-spent-amazon-data-analysis/>



22/24

Som príliš často na sociálnej sieti, čo hovoria dáta ?

◀ Back

Forward ▶

Postujeme príliš často?

- ✓ Čo vlastne znamená "príliš často"?
- ✓ Potrebujeme nejakú referenčnú hodnotu.
- ✓ S ňou sa môžeme skúsiť porovnať.
- ✓ Ide o dáta meniace sa v čase: teória čaových radov (stochastických procesov).
- ✓ Proces vykazujúci stacionaritu: všetky alebo niektoré štatistické vlastnosti sú nezávislé na čase.
- ✓ Napr. strednú hodnotu (spriemernenú hodnotu).

Empirická pravdepodobnosť v strojovom učení?

Nachádza si cestu do strojového učenia rôznymi spôsobmi.
Napríklad:

- 1 v empirických Bayesových metódach sa zozbierané údaje aktualizujú na základe tzv. apriórnych pravdepodobností;
- 2 pred vytvorením tzv. posteriórneho rozdelenia - vytvorenie hyperparametrov podľa apriórneho predpokladu;
- 3 testovanie apriórnej pravdepodobnosti - "ladenie" rozdelenia pravdepodobnosti na zlepšenie približných hodnôt parametrov;