PROJEKTY PRE PRINCÍPY DÁTOVEJ VEDY 2023

Termín zostavenia skupín a výberu témy: 1.12.2023, 22:00
Termín odovzdania projektu: 8.1.2024, 22:00

Projekt je skupinový a riešite ho v skupinách po 3-5 študentov. **Skupiny si dohadujete medzi sebou,** každá skupina musí **do termínu zostavenia skupín** do Google Classroom postnuť správu, ktorá obsahuje: názov skupiny, zoznam členov a tému, ktorú ste si vybrali. Na jednej téme môže nezávisle pracovať viac skupín. **Projekt odovzdáva každý člen tímu tak, že do google classroom do príslušnej úlohy zapíše linku na githubový repozitár a odovzdá úlohu.**

Na projekte pracujete ako skupina samostatne, v prípade záujmu si môžete dohodnúť konzultáciu, nie je to však nevyhnutné a ani vám nevieme zaručiť, že naše nápady budú akokoľvek užitočné ;) Individuálna diskusia o vašom projekte bude tvoriť časť skúšky, každý člen skupiny by tak mal byť oboznámený s veľkou časťou aspektov projektu a mal by vedieť podrobne opísať aj svoj príspevok.

Projekty budete odovzdávať ako githubový repozitár - zvážte, či repozitár bude verejný (preferované) alebo do privátneho repozitára jednoducho prizvete vyučujúcich predmetu (napr. ak ste používali dáta, ktoré nepochádzajú z verejných zdrojov a nemožno ich zverejniť). Vo vašom githubovom repozitári by mali byť jednoznačne identifikovateľné nasledujúce časti:

- **Správa o projekte** je ucelený dokument, ktorý zhŕňa výsledky vášho projektu; môže to byť napríklad README.md alebo PDF dokument v hlavnom adresári. Musí obsahovať:
    - špecifikáciu otázok, ktorým ste sa venovali
    - prehľad dátových zdrojov, s ktorými ste pracovali
    - výsledky vašej analýzy (tabuľky, grafy a ich textovú diskusiu)
    - stručný popis použitých nástrojov, metód a technických výziev, s ktorými ste sa pri riešení stretli

  Tomuto dokumentu venujte patričnú pozornosť, keďže bude hlavným podkladom ku hodnoteniu projektu.

- **Dáta s ktorými ste pracovali** alebo popis, akým spôsobom tieto dáta získame z pôvodných zdrojov (nereplikujte pôvodný dátový zdroj, obzvlášť pokiaľ sa jedná o veľké súbory). Ak ste niektoré dátové súbory vyrábali manuálne z dokumentov a pod. tak aj tieto súbory

- **Váš zdrojový kód** so stručným popisom (môže byť aj vo forme notebookov)

Téma 1: Game of Thrones

Viete vypozorovať a/alebo predpovedať trendy vyhľadávania kľúčových postáv zo série Game of Thrones? Okrem základnej analýzy, viete vymyslieť aj nejaké zaujímavé typy analýz (napr. majú niektoré skupiny postáv výrazne odlišné trendy od iných skupín?) Nezabudnite zobrať do úvahy aj dátumy vydania jednotlivých kníh.

Možný zdroj dát:
Google Trends - https://trends.google.com/ (umožňuje aj bulk download dát)
IMDB obsahuje napr. dátumy prvého vysielania jednotlivých epizód
Fanúšikovské stránky a wikipédia obsahujú histórie jednotlivých postáv a v ktorých dieloch sa vyskytujú

Téma 2: Predikovanie počtu študentov na predmetoch

Ako presne by ste vedeli odhadnúť v júni resp. v novembri, koľko študentov si zapíše jednotlivé predmety v ďalšom semestri?

Možný zdroj dát:
Dáta, ktoré používajú rozvrhári pre tvorbu rozvrhov (poskytne vyučujúci na požiadanie)
Počty prihlášok na štúdium nájdete vo výročných správach FMFI UK
Ďalšie dáta, ktoré je možné vytiahnuť z AIS2, sprístupníme podľa dohody

Téma 3: Konštrukcia data setu áut

Vyrobte dataset, ktorý obsahuje fotky áut spolu s rokom výroby auta (môže byť aj rozsah, pokiaľ sa identický model vyrábal viac rokov). Dataset by mal byť čo najrozsiahlejší, t.j. mal by obsahovať autá rôznych značiek, ale dôležité je aby fotografie neboli len “výstavné” fotografie, ale aby boli v čo najrozmanitejších podmienkach (napr. za jazdy, za dažďa, v noci keď vidno len svetlá…, v rôznych prostrediach a pod.)

Téma 4: Čo robí filmy obľúbenými a neobľúbenými?

Pokúste sa identifikovať faktory, ktoré robia filmy obľúbenými alebo neobľúbenými. Faktory môžu zahŕňať napríklad žáner, štúdio, obdobie kedy bol film uvedený na trh, ale aj konkrétnych hercov či iných členov produkčného tímu (a čokoľvek iné, čo vám napadne ako vhodný faktor).

Možný zdroj dát:
Stránka IMDB obsahuje podrobné informácie o filmoch a ich popularite. Veľkú časť dát možno stiahnuť v ľahko spracovateľných formátoch, niektoré informácie sú však prístupné len cez stránku.

Téma 5: Ako sa šírila nákaza COVIDu medzi jednotlivými oblasťami Slovenska?

Na začiatku pandémie sa predpokladalo, že hlavným centrom šírenia choroby bude Bratislava a z nej sa nákaza bude šíriť do ostatných oblastí. Dáta však naznačujú, že táto predstava bola naivná a že šírenie COVIDu na Slovensku bolo podstatne komplikovanejšie a súviselo aj s väzbami regiónov na okolité štáty a na Veľkú Britániu.

Možný zdroj dát: Celosvetová databáza GISAID (kompletnú databázu poskytne vyučujúci) obsahuje informácie o tom, ako sa v jednotlivých regiónoch vyskytovali jednotlivé varianty COVIDu. Varianty COVIDu sa šírili vo vlnách a nastupovali v rôznych regiónoch v rôznych časoch, vďaka čomu možno sledovať časové následnosti.