PROJEKTY PRE PRINCÍPY DÁTOVEJ VEDY

Termín zostavenia skupín a výberu témy: 8.12.2024, 22:00
Termín odovzdania projektu: 6.1.2025, 22:00

Všeobecná organizácia:

- Projekt je skupinový a riešite ho v skupinách po 3-5 študentoch.
- 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 projekte pracujete ako skupina samostatne. V prípade záujmu si môžete dohodnúť konzultáciu, nie je to však nevyhnutné a ani vám neviem zaručiť, že moje nápady budú akokoľvek užitočné.
- Na jednej téme môže **nezávisle** pracovať viac skupín.

Odovzdávanie:

- Projekty budete odovzdávať ako githubový repozitár - zvážte, či repozitár bude verejný (preferované) alebo ma jednoducho prizvete do privátneho repozitára (napr. ak ste používali dáta, ktoré nepochádzajú z verejných zdrojov a nemožno ich zverejniť).
- Projekt odovzdáva každý člen tímu tak, že do Google classroom do príslušnej úlohy zapíše link na githubový repozitár a odovzdá úlohu.
- 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.

Obsah:

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 PDF dokument, prípadne README.md 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 postup, akým spôsobom tieto dáta získame z pôvodných zdrojov (nereplikujte pôvodný dátový zdroj, ak 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).

Poznámky:

- Ak niečo predikujete, mali by ste rozumne overiť, či vaše predikcie sú ozaj dobré. Podobne ak niečo modelujete (napr. kvôli zisťovaniu, ako iné premenné vplývajú na modelovanú premennú), tak by ste mali overiť, že váš model dobre fituje a predikuje dáta.
- Dôležité sú poučenia, vysvetlenia, interpretácie, zaujímavé vykreslenia, nie iba priamočiare „mám predikčný model a som spokojný".
- Okrem primárnej (typicky predikčnej) úlohy dáta zvyčajne poskytujú zaujímavé informácie aj o sekundárnych veciach. Vždy zahrňte do projektu aj sekundárne analýzy: čo zaujímavého sme sa navyše z dát dozvedeli. Nezabudnite, že okrem predikčných modelov sme preberali aj iné metódy (PCA, zhlukovanie, štatistické testy...).
- Uvedené zdroje dát sú iba tipy: smelo používajte aj iné a tie uvedené ani vôbec nemusíte použiť.

Témy:

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.

Možné zdroje 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: Predikcie volieb

Predikujte výsledky volieb na Slovensku na základe výsledkov prieskumov a prípadne ďalších charakteristík. Ako by na základe momentálnych prieskumov dopadli parlamentné voľby, keby sa konali o mesiac? A ako, keby sa konali neskôr: napr. podľa harmonogramu, t.j. v roku 2027? Nezabudnite zobrať do úvahy vzťah medzi prieskumami v minulosti a reálnymi výsledkami volieb: napr. niektoré strany (alebo „kategórie" strán) typciky dopadajú lepšie/horšie v realite než v prieskumoch.

Analýzu môžete prípadne spraviť pre prezidentské voľby v USA (a môžete sa tváriť, že ste v čase pred 5.11.2024), alebo pre inú krajinu, kde tušíte niečo o politike a viete získať dobré dáta. Pre prípad USA je taký predikčný model (ktorý je oveľa náročnejší, než vyžaduje tento projekt) načrtnutý napr. v https://abcnews.go.com/538/538s-2024-presidential-election-forecast-works/story?id=113068753

Možné zdroje dát:

Napríklad agentúra Focus (https://www.focus-research.sk/) archivuje svoje výsledky prieskumov preferencií od januára 2021, ale k starším prieskumom sa dá dopátrať buď cez „novinky“ na ich stránke, alebo googlením (napr. google: focus volebne preferencie politickych stran august 2016)
Agentúra AKO - https://ako.sk/
Štatistický úrad - https://volby.statistics.sk/

Téma 3: Analýza volieb po okresoch

Pokúste sa čo najlepšie modelovať výsledky parlamentných volieb v jednotlivých okresoch SR v závislosti od charakteristík týchto okresov (ako miera nezamestnanosti, národnostné zloženie...). Výsledky modelu (alebo sady modelov) by mali byť interpretovateľné: teda aby sa z modelu dal vyčítať pozorovaný vzťah medzi charakteristikami okresov a výsledkami volieb. Spravte aj ďalšie analýzy: napr. aké sú vzťahy medzi jednotlivými charakteristikami, vychádza model výrazne iný pre rôzne volebné roky (napr. 2023 vs 2020), ...?

Možné zdroje dát:
Štatistický úrad: voľby - https://volby.statistics.sk/
Štatistický úrad: Datacube - https://datacube.statistics.sk/
Sčítanie obyvateľov - https://www.scitanie.sk/

Téma 4: Popularita hudby

Analyzujte, čo robí populárnymi piesne na Spotify. Vedeli by ste predikovať úspešnosť piesne na základe jej iných charakteristík? Analyzujte aj ďalšie javy: napr. je závislosť popularity od charakteristík rovnaká pre rôzne žánre; ako sa menia charakteristiky piesní v čase; existujú hudobníci, ktorí v rámci žánru tvoria populárne, ale veľmi netradičné pesničky; …?

Možné zdroje dát:
Spotify API: na webe sú návody, ako tú API využiť na naťahanie údajov
Prípadne môžete použiť niektorý zo „Spotify“ datasetov na Kaggli. Môžete z nich vybrať vhodný alebo ich viacero skombinovať.

Téma 5: Žánre hudby

Celé zadanie je rovnaké ako téma 4, akurát namiesto analyzovania popularity analyzujte žánre hudby. Teda skúste skonštruovať model, ktorý čo najlepšie predikuje, do akého žánru daná pesnička patrí na základe jej charakteristík. Okrem toho, ako vždy, analyzujte ďalšie javy.

Téma 6: Dohodou

Ak vás zaujíma téma, ktorú nemáte na bakalárke alebo inom predmete, a je náročnosťou porovnateľná s predchádzajúcimi, tak sa mi ozvite a ja vám dám vedieť, či môže taká

téma byť. **Pozor**, téma nemusí byť schválená (môže byť napr. príliš ľahká, alebo príliš ťažká), takže jej návrh mi dajte vedieť v dostatočnom predstihu pred deadlinom na výber témy.