

PROJEKTY PRE PRINCÍPY DÁTOVEJ VEDY 2025

Termín zostavenia skupín a výberu témy: 9.12.2025, 22:00

Skupiny zapisujte sem:

<https://docs.google.com/spreadsheets/d/1uGmR74HVd-jETqEirMKwETYirB3uGN5f-fnKHUW8CpE/edit?usp=sharing>

Termín odovzdania projektu: 9.1.2026, 22:00

(odovzdanie do systému vektor)

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 vyššie uvedeného spreadsheetu zapísať názov skupiny, zoznam členov a tému, ktorú ste si vybrali. Na jednej téme môže nezávisle pracovať viac skupín.

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. **Projekt odovzdáva každý člen tímu tak, že do systému vektor do príslušnej úlohy zapíše linku na githubový repozitár a odovzdá úlohu.** 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, vizualizáciu a ich textovú interpretáciu / diskusiu)
 - stručný popis použitých nástrojov, metód a technických výziev, s ktorými ste sa pri riešení stretli
 - stručný popis štruktúry kódu

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 komentovaný zdrojový kód** so stručným popisom (môže byť aj vo forme notebookov). Kód by mal byť spustiteľný aj na našich počítačoch a pre demonštráciu jeho fungovania by ste mali poskytnúť aj nejaké malé demo dátové súbory.

Ďalšie poznámky:

- Ak niečo predikujete, vyhodnoťte aj aké dobré sú vaše predikcie. (Vyhodnotenie úspešnosti, porovnanie s nejakým jednoduchým baseline prístupom, ...)
- Diskusia a detailná analýza výsledkov sú rovnako dôležité, ak nie dôležitejšie, ako vybudovanie predikčného modelu.
- Zahrňte do diskusie aj sekundárne analýzy, ktoré ukazujú čo zaujímavé navyše oproti pôvodnej otázke ste sa z dát dozvedeli.
- Na riešenie úlohy môžete použiť aj iné zdroje dát, ako sú tie, ktoré sú uvedené v zadaní; naopak, tie zo zadania nemusíte použiť vôbec, ak nájdete lepšie.

Téma 1: 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 GISAIID (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.

Téma 2: StarTrek

Viete vypožorovať, ako konkrétne postavy v seriáloch StarTrek ovplyvňujú obľúbenosť alebo neobľúbenosť jednotlivých epizód?

Možný zdroj dát:

IMDB obsahuje zoznamy epizód, sledovanosť, hodnotenie, ako aj zoznamy postáv a hercov. Fanúšikovské stránky a wikipédia obsahujú histórie jednotlivých postáv a v ktorých dieloch sa vyskytujú.

Téma 3: Čo ovplyvňuje volebné výsledky v okresoch

Pokúste sa čo najlepšie modelovať výsledky parlamentných volieb v jednotlivých okresoch SR v závislosti od rôznych charakteristík týchto okresov (nezamestnanosť, priemerný plat, vekové zloženie...). Výsledné modely by mali byť interpretovateľné, čiže z modelu by sa mal dať vyčítať pozorovaný vzťah medzi charakteristikami okresov a výsledkami volieb. Môžete pracovať aj s prieskumami, ale myslite na to, že prieskumy a volebné výsledky sa môžu správať odlišne.

Možné zdroje dát:

Datacube <https://datacube.statistics.sk/>

Sčítanie obyvateľov <https://www.scitanie.sk/>

Štatistický úrad <https://volby.statistics.sk/>

Téma 4: Bratislavské cyklotrasy

V Bratislave sú umiestnené cyklosčítače na rôznych miestach v sieti cyklotrás. Využite tieto dáta na analýzu správania cyklistov v Bratislave. Sú všetky cyklocesty vyťažené rovnako v rámci dňa, týždňa, mesiaca, roku? Viete klasifikovať rôzne cyklotrasy na základe týchto údajov? Na vyťaženosť ktorých vplyva počasie najviac? Zmenilo pridanie cyklocesty na Vajanského nábreží trendy v správaní cyklistov?

Formulujte hypotézy o pozorovaných správaniach. Tieto otázky slúžia len na inšpiráciu, môžete analyzovať čokoľvek relevantné. Táto téma je výborná na vizualizáciu (pozrite si knižnicu osmnx).

Možné zdroje dát:

Cyklosčítače <https://data.bratislava.sk/pages/cykloscitace>

Meteostat <https://meteostat.net/en/place/sk/bratislava?t=2025-11-17/2025-11-24>

Téma 5: Ako veľmi používame ChatGPT pri písaní záverečných prác?

LLM začali byť populárne začiatkom roku 2023. Preto je teraz ideálna doba pozrieť sa na to, ako ho používajú študenti pri písaní záverečných prác: veľa študentov, ktorí odovzdávali svoju bakalársku prácu v roku 2022 (pred nástupom LLM) následne v roku 2024 (v čase už hojného používania LLM) odovzdávali diplomovku na podobnú tému. Pomocou analýzy štýlu písania (napríklad frekvencie k-tic písmen v nejakom úseku textu) skúste odhadnúť, za akú časť diplomových prác študentov sú zodpovedné LLM (ako referenciu štýlu použite bakalársku prácu toho istého študenta z obdobia pred nástupom LLM). [Začnite so zopár konkrétnymi študentami, následne si rozmyslite ako by ste urobili širšiu analýzu.]

Možné zdroje dát:

Centrálny register záverečných prác <https://opac.crzp.sk/>

obsahuje plné texty záverečných prác vo formáte PDF s extrahovateľným textom