PROJEKTY PRE PRINCÍPY DÁTOVEJ VEDY 2022

Projekt je skupinový a riešite ho v skupinách 3-5 študentov. **Skupiny si dohadujete medzi sebou,** každá skupina **najneskôr do 12.12.** musí 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. **Termín odovzdania projektu je najneskôr 2 dni pred vašou skúškou, 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 s ktorýmkoľvek vyučujúcim 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 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
- **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: Ako sa vysoké školy prispôsobujú zmenám

Ako sa vysoké školy prispôsobujú zmenám v počtoch študentov alebo zmenám v rozpočte? Konkrétne nás zaujíma, ako sa menia počty ich vedecko-pedagogických zamestnancov v rôznych stupňoch kariéry (lektor, odborný asistent, docent, profesor, prípadne výskumný pracovník) vzhľadom ku týmto externým ukazovateľom. Pracujú zamestnanci aj na viacerých pracoviskách a vidno v tom nejaké trendy alebo zhluky? Viete v dátach vypozorovať aj nejaké iné zaujímavé trendy?

Možné zdroje dát:

Portál vysokých škôl obsahuje register zamestnancov vrátane výšky úväzkov a histórie pracovných pomerov https://www.portalvs.sk/regzam/
Ministerstvo školstva má na stránke podrobné rozpočtové podklady https://www.minedu.sk/financovanie-vysokeho-skolstva/
Nezabudnite na infláciu ako faktor

Téma 2: Eigenfaces

Analýzu hlavných komponentov je možné aplikovať aj na dáta ako sú fotografie.
(Kľúčové slovo je eigenfaces, krátky popis je napr. tu: https://www.geeksforgeeks.org/ml-face-recognition-using-eigenfaces-pca-algorithm/)
Aplikujte túto metódu na fotografie spolužiakov alebo na fotografie učiteľov na matfyze. Pokúste sa pomocou nej vyhľadávať skupinky ľudí, ktorí sa navzájom podobajú alebo skúste vymyslieť iné použitia.

Budete potrebovať fotografie predspracovať tak, aby mali tváre normalizovanú pozíciu a veľkosť. V tomto vám môže pomôcť knižnica opencv (hľadajte kľúčové slová face alignment).

Možný zdroj dát: fotografie učiteľov viete nájsť na ich osobných stránkach, ktoré sú zalinkované zo stránok jednotlivých katedier

Téma 3: 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 4: Predikcia trendov vo vedeckých článkoch

Viete predpoveď, aké vedecké témy (kľúčové slová) budú cool na základe predchádzajúcich trendov? Môžete vychádzať napríklad z názvov vedeckých článkov a ich abstraktov. Môžete skúmať rôzne ochorenia, lieky, prípadne môžete skúmať aj kľúčové slová bez predchádzajúceho výberu.

Možné zdroje dát:
Databáza pubmed obsahuje informácie o biomedicínskych článkoch a umožňuje bulk downloads: https://www.nlm.nih.gov/databases/download/pubmed_medline.html