4IZ503
PROJEKTOVÝ SEMINÁŘ

Miloš Maryška, Petr Máša, Ota Novotný, Jan Rauch

Organizace

Požadavky na projekt

Termíny

CÍLE PROJEKTU

Cíle projektu

- Vyzkoušet si samostatně komplexní úlohu zpracování dat v kombinaci BI a data science
- Tato úloha bude i podkladem pro SZZ/zkoušku, na které proběhne rozprava a otázky budou vycházet z témat zpracovávaných v projektu (mohou být rozvinuté).
- Projekt má potvrdit, že student je schopen připravit malé BI řešení a využít data science. Měl by pokrývat všechny fáze od přípravy dat, přes zpracování, prezentaci až po interpretaci.

OBSAH A FORMA CVIČENÍ

Obsah a forma cvičení

- Cílem je připravit jeho projekt nebo podstatnou část (min 2/3), který bude sloužit jako podklad při SZZ
- Forma cvičení – nepravidelné
  - Úvodní cvičení
  - Konzultace 5. týden (17.10.)
  - Konzultace 9. týden (14.11.)
  - Případně konzultace dle potřeby

ODEVZDÁVANÁ PRÁCE

Základní principy 1/5

- Zvolte si libovolnou oblast a k ní obstarejte data (ideálně z oblasti práce nebo hobby, kde jste schopni business interpretace nebo máte možnost konzultovat s majitelem dat)
  - Možno volit datasety na UCI ML Library, Kaggle, KDNuggets či další online zdroje
  - Ideálně desítky tisíc řádků a nižší až střední desítky sloupců
- Úloha by měla obsahovat netriviální spojení více datových zdrojů, část BI, část data science a zdůvodnění jejich propojení, které bude srozumitelné pro majitele dat
- **Dosažení úlohy je shrnutí povinných předmětů (4IT403, 4iz450,4iz460)**

Základní principy 2/5

- Odevzdání úvodní zprávy, 2 konzultace a následné finální odevzdání (503)

1. Úvodní zpráva o úloze a o datech („znám oblast, vím jak data vypadají a vím co chci řešit za úlohy v oblasti data science a jak – LISp-Miner/Python“) kdy se odevzdává – pátý týden
2. Výsledný projekt („mám hotový projekt nebo jeho draft verzi ke státnici/ke zkoušce“)

- Důvodem kroků je zpětná vazba, případně nasměrování tak, aby výsledný projekt co nejvíce odpovídal požadavkům / splnění cíle

Základní principy 3/5
Požadavky za část BI

- Připravit návrh datového modelu navrženého, schéma architektury navrženého řešení včetně workflow zpracování dat
- Obrázek datového modelu implementovaného na základě provedeného návrhu a obrázek implementovaných datových pump
- Grafický návrh reportu v PowerBI či jiném reportingovém nástroji, který bude uživatelským rozhraní pro prezentaci výsledků zpracování dat a významných zjištění
- Identifikace klíčových problémů při návrhu procesu zpracování dat a zejména způsob jejich odstranění a problémy při návrhu reportingového řešení

Základní principy 4/5
Požadavky za části Data Science 1/2

- Cíl - ukázat možnosti doplnění výsledků BI pomocí prostředků data science, zejména interpretovatelných metod
  - Typicky pravidla, stromy s interpretací (ne strom s R=1.00 a proměnné A01-A29) a kontingenční tabulky
  - Možné též výjimky k základním vztahům nalezený pomocí BI
  - zajímavé vztahy platné v analyzovaných datech doplňující výsledky dosažené pomocí BI
  - Cílem je nalezení několika (5-10) zajímavých vztahů v datech (zejména kategoriálních) odpovídajícími metodami a jejich vizualizace
  - Součástí je i příprava dat pro data science, tedy příprava do vhodného formátu (matice), kategorizace, profilování, iniciální zhodnocení profilů/obsahu dat a dále následná interpretace nalezených vztahů
  - část Data Science je možno řešit pomocí systému LISp-Miner nebo v jazyku Python s využitím modulů CleverMiner a jiných interpretovatelných metod. Po dohodě lze použít i jiné nástroje, musí být však pokryty asociační pravidla, vybrané další interpretovatelné či interpretované metody a subgroup Discovery
  - Jiné než probírané metody jsou po předchozí konzultaci možné, pokud podkladová teorie nebyla zkoušena v rámci povinných předmětů, může být zkoušena při obhajobě projektu

Základní principy 5/5
Požadavky za části Data Science 2/2

- Lze vycházet i z práce z IZI460, nutno nakombinovat s minimálně 1 dalším zdrojem a nové analýzy
- Připravit vlastní (nové) úlohy, minimálně 1-2x z každé z procedur 4ft-Miner, CF-Miner, SD4ft-Miner, prediktivní model (s interpretací) NEBO analýza kontingenčních tabulek či jiná interpretovatelná metoda (ve formátu 1 slide slovní zadání + screenshot zadání, screenshot výsledků – souhrn + několik zajímavých pravidel, ústní interpretace
- Úlohy lze nahradit jinými interpretovatelnými metodami, u každé je však nutné docílit opravdové interpretace pravidel ve formě znalosti, která by šla aplikovat v rámci dané problematiky
- Souhrn nejzajímavějších nalezených znalostí (interpretovatelných, ne nutně všech – o několika slidů) spolu s prezentací interpretace výsledků a možností přímých aplikací (jak)je hlavní součástí hodnocení
  - Tyto slidy budou tvořit jádro prezentace u SZZ/zkoušky, proto jim věnujte náležitou pozornost
- Shrnutí výsledků prediktivního modelu jako podpůrný výsledek

Úvodní zpráva obsahuje

- Popis řešené oblasti
- Popis dat (jaká data se použila – zdroj + URL pokud lze, minimálně 2 netriviálně kombinované datasety z různých zdrojů, tj. vazba alespoň 1:n) – nemusí platit pro vlastní data (z firmy, ze senzorů, …)
- Cíle analýzy, doménové znalosti pokud jsou k dispozici
- Rámcové profily dat (typy proměnných – kategoriální/spojité, počet kategorií, počet proměnných pro jednotlivé typy)
- Návrhy oblastí k analýze (iniciální) – nemusí být následně dodrženo + ukázky výsledků BI
- Způsob řešení otázek pro data science – LISp-Miner / Python
- Doporučený Formát: Powerpoint odevzdat v pptx

Výsledný projekt obsahuje

- Popis zvoleného data setu a dat z úvodní zprávy, případně rozšířený o profily atributů (pokud nebylo v úvodní zprávě)
- Závěry BI analýzy
- Seznam analytických otázek navazujících na BI analýzu
- Dílčí řešení navazujících analytických otázek
  - Otázky vyřešeny a připraveny ve výsledném formátu – tzn. business popis, technické řešení, výsledky, business interpretace / doporučené akce
- **Souhrn nejzajímavějších nalezených znalostí (interpretovatelných, ne nutně všech – o několika slidů) spolu s prezentací interpretace výsledků a možností přímých aplikací (jak)je hlavní součástí hodnocení**
  - **Tyto slidy budou tvořit jádro hodnocení a prezentace u SZZ, proto jim věnujte náležitou pozornost**
- Popis kroků k dokončení projektu
- Požadované přílohy:
  - kód + data (PowerBI workbook + python kód/Jupyter notebook)
  - data + metabáze LISp-Mineru
- Doporučený Formát: Powerpoint – odevzdat v pptx + přílohy (zip)
- Jsou možné i odchylky od zadání, je nutné je včas konzultovat

Pilíře hodnocení

- Úloha, výsledky, jejich interpretace a (potenciální) přínos pro majitele
- Získání, příprava a kombinace dat, odvození netriviálních (vhodných) ukazatelů
- Popis datových pump, BI dashboardy
- Data mining úlohy – výsledky a interpretace
- Celkový storytelling

- Hodnotí se nejen technické splnění, ale vhodnost a zajímavé použití

Kam se jednotlivé práce odevzdávají

- Do odevzdávárny v INSISu

DŮLEŽITÉ – Forma zakončení

- Pro studenty, kteří mají VS zakončenou SZZ (přihlášení na VS před ZS 23/24)
  - na základě výsledné práce (alespoň 2/3 požadavků) udělí jeden z vyučujících známku do INSISu
  - Obhajoba práce bude v rámci SZZ
- Pro studenty, kteří NEMAJÍ VS zakončenou SZZ (přihlášení na VS v ZS23/24 a později)
  - Obhajoba práce bude v rámci ústní zkoušky
- Na ústní zkoušku z 4IZ503 se přihlašují PRÁVĚ A POUZE studenti nekonající SZZ. Pro ostatní je povinná SZZ. Forma a obsah bude obdobná. Jedná se o prezentaci výsledků a rozpravu.
- Znamená to tedy, že student bude obhajovat svou práci
  - Buď v rámci SZZ (studenti přihlášeni na VS před ZS23/24)
  - Nebo v rámci ústní zkoušky 4IZ503 (studenti přihlášeni na VS v ZS23/24 a později).
- Student je zodpovědný za (ne-)přihlášení se na ústní zkoušku podle termínu zahájení VS.

DŮLEŽITÉ - použití LLM

Použití LLM je povoleno v souladu s regulací na VŠE
Nutno popsat, na co bylo LLM použito a jak (vstupy, výstupy, jak byly zpracovány, jak byly ověřeny)

Na co si dát pozor:

- Vygenerování kódů pro úlohy je typicky chybné, generuje obvykle nesmysly
- Student odpovídá za správnost odevzdaných výstupů
- Použití jiných procedur než byly zkoušeny v rámci povinných předmětů znamená že je potřeba jim rozumět a budou obsahem zkoušení
- Proto doporučujeme použít LLM pro generování kódu omezeně (vibe programming vs. AI assisted programming)

K čemu může být LLM vhodné

- Interpretace pravidel
- Další náměty využití – volné pole působnosti (nutno popsat jak bylo použito a jaké jsou závěry)

TERMÍNY

Termíny

- Odevzdání
  - Sobota 5. týdne – 18.10. EOD – odevzdání úvodní zprávy o datech
  - Sobota 12. týdne – 6.12. EOD - odevzdání průběžného projektu (2/3 + informace co budete dodělávat) – POUZE PRO STUDENTY KONAJÍCÍ SZZ (pro účely klasifikace předmětu)
  - Minimálně týden před zkouškou / SZZ – odevzdání finální verze projektu
- Konzultace
  - Pátek 5. týdne – 17.10
  - Pátek 9. týdne – 14.11
- Možné změny harmonogramu budou oznámeny emailem

Pro SZZ/Zkoušku

- Práci odevzdávejte minimálně týden před konáním SZZ/zkoušky do odevzdávárny (nebo emailem zkoušejícím)
  - Formát odevzdání stejný jako finální zpráva
- Zároveň si připravte představení BI části a nejzajímavějších nálezů v data science (ne nutně všeho), budete mít 10 minut na prezentaci
  - Není nutné představit vše