



Posudek oponenta závěrečné práce

Oponent práce: Ing. Petr Kasalický
Student: Ostap Iudin
Název práce: Detekce poruch v provozních datech serverů
Obor / specializace: Znalostní inženýrství
Vytvořeno dne: 8. června 2022

Hodnotící kritéria

1. Splnění zadání

- [1] zadání splněno
- [2] zadání splněno s menšími výhradami
- ▶ [3] zadání splněno s většími výhradami
- [4] zadání nesplněno

Student měl za cíl zorientovat se v problematice monitorování infrastruktury, naimplementovat potřebné nástroje pro sběr dat, uložení a vizualizaci dat. To vše bylo provedeno a je to rozsáhle popsáno v práci. Dalšími body v zadání je analýza přístupů na detekci selhání (failures), jejich implementace (alespoň jednoho) a provedení experimentu na nasbíraných datech. Analýza popsaná v práci je však velmi povrchní, zvláště pro obor znalostního inženýrství. Vybrané detekční metody jsou představeny pár větami bez hlubší analýzy, jak skutečně fungují či proč byly vybrány zrovna ony. Samotný experiment na nasbíraných datech a jeho průběh je nepopsán. Chybí popis nasbíraných dat i způsob vyhodnocení kvality detekce (vyjma kódu 6.1). Navíc závěr práce uvádí "The anomaly detection system has yet to be finalized by conduction additional experiments", čili student sám přiznává, že detekce anomalit je nedokončená.

2. Písemná část práce

40/100 (F)

Rozsah práce odpovídá bakalářské práci. Text je psaný pěknou angličtinou bez výrazných chyb, vyjma občasných neformálních výrazů jako je "I'm". Avšak to je tak jediné pozitivní na písemné části práce.

Už před začtením do úvodu si lze všimnout sekce Declaration obsahující "FILL IN ACCORDING TO THE INSTRUCTIONS" a následované výplňkovým Lorem Ipsum textem.

Struktura i obsah práce pak spíše odpovídá manuálu k softwaru, než závěrečná práci z oboru znalostního inženýrství. Kapitoly jsou dělené podle probíraného tématu (Introduction, Collection of Data, Storage of Data, Alerting, Anomaly detection, Experiment) a tedy nedodržují standardní rozdělení vědeckých prací (Introduction,

Related Work, Our approach, Experiment, Results). V jednotlivých kapitolách student míchá motivaci k práci, popis použitých technologií, kusy kódu konfiguruující tyto technologie a snad i své řešení. Jako důsledek není jasné, kdy student představuje nějaké existující řešení, a kdy popisuje svůj přínos. Také logická stavba je porušena, když na vybraný algoritmus v rámci experimentů, který je více představen až v kapitole 6, je odkazováno už v sekci 5.5 s tím, že funguje dostatečně rychle.

Práce obsahuje výrazné formální nedostatky, konkrétně:

Listing 2.1, 2.2, 2.3, 2.4, 4.5 prokládají text, avšak do něj nejsou nijak zapracovány, text se na ně neodkazuje ani je nevysvětluje.

Figure 4.11, 4.12 nic nevysvětlují, pouze ukazují náhled Dashboardu v Grafaně

Code Listing 5.4 je také neodkazovaný a nepochopil jsem jeho účel.

V kapitole 5 věta "a script was written - Anomaly Detection 5.1" odkazuje na Figure se schématem.

Na stranách 11 a 33 je půlka stránky prázdná, jako důsledek velkého obrázku na následující straně, avšak určitě šlo text a obrázky lépe poskládat, aby text působil kompaktně.

Když už se student odkazuje na Figure/Listing/Section, tak v textu uvádí jen číslo (např. 5.1) a tedy nemusí být jasné, zda je odkaz na Figure, Listing nebo Sekci (zvláště když jsou blízko sebe).

V Sekci 6.1 začíná odstavec "The function accepts", ale předtím nebyla žádná funkce zmíněna a není tedy jasné, o jaké funkci je řeč.

Popisek "Figure 6.10: Prediction" je nic neříkající.

V Sekci 3.2 je zapomenutá poznámka, že chtěl student citovat InfluxDB, ale nenahradil ji odkazem na práci.

V kapitole Experiment chybí základní popis dat a metodologie experimentu. Místo toho jsou zde v podsekcích o délce jednoho odstavce představovány jednotlivé způsoby, jak je možné detekovat anomálie v časové řadě s krátkým slovním komentářem, jak metoda fungovala na nasbíraných datech. Samotné výsledky experimentu jsou pak uvedené v jednom odstavci v sekci 6.3: Result of experiment, která je doplněna o jednu tabulku a dva grafy.

Práce cituje 19 zdrojů, polovina z toho jsou manuály a dokumentace k použitým nástrojům (InfluxDB, Grafana, Prometheus, AlertManager) a pouze 9 zdrojů cituje odbornou literaturu týkající se problematiky detekce anomálií. Poslední větou celé práce je informace, že vzniklý projekt je open-source, avšak přiložený odkaz vede na GitHub repozitář společnosti Seznam a ne na repozitář s prací.

3. Nepísemná část, přílohy

80/100 (B)

Dodaný kód je netriviální a velmi rozsáhlý. Některé implementované funkce obsahují dokumentaci ve formě docstringů, avšak zdaleko ne všechny. Výpis přiloženého média uvádí, že jsou přiložena i data (pravděpodobně získaná během experimentu), avšak přiložené CD obsahuje jen 2 prázdné CSV soubory a prázdné složky.

4. Hodnocení výsledků, jejich využitelnost

70/100 (C)

Programy naimplementované během práce jistě najdou svou využitelnost v oblasti monitoringu. Zůstanou však jen jako softwarový nástroj pro sběr dat. Na samotnou

analýzu detekce poruch bude velmi náročné navázat, protože není v práci dostatečně popsána. Získaná data nejsou součástí přiloženého média.

Celkové hodnocení

50/100 (E)

Student prokázal, že nástroje používané k monitoringu nastudoval, rozumí jim a dokáže je použít v praxi. Také odvedl velké množství práce na jejich rozšíření za účelem zisku dat potřebných pro detekci anomálií. Avšak nasbíraná data vytěžuje je velmi povrchně, bez hlubší analýzy a srovnání stávajících řešení. Nejslabší částí práce je písemná část, která obsahuje spoustu nedodělků, dopouští se obrovského množství formálních a stylistických chyb a vypadá spíše jako manuál k softwaru než jako vědecká práce z oboru znalostního inženýrství.

S přihlédnutím k velkému úsilí v nepísemné části doporučuji bakalářskou práci k obhajobě a hodnotím ji 50 body (E).

Otázky k obhajobě

1. Úvod kapitoly 5 uvádí, že "Script ...detects anomalies using a scientific model". Co myslíte tím "scientific model"? Jedná se o model k detekování anomálií založený na datech, který jste měl v práci představit, nebo se jedná o blíže nepopsaný externí model, v jehož výsledcích hledáte anomálie?

2. Jako motivaci k použití detekce založené na datech uvádíte na straně 17, že měříte 150 různých metrik a manuálně vytvořit pravidla pro detekci anomálie ze všech metrik (a jejich kombinací) by bylo náročné. Avšak při experimentování s různými detekčními modely používáte jen jednu metriku a to "node_cpu_seconds_total". Jaká je tedy přidaná hodnota popsaných metod? Fungují vůbec lépe než manuálně nastavená detekce pro jednu metriku (tabulka 6.1 ji neuvádí)?

3. Proč nejsou součástí přiloženého CD i všechna data naměřené během experimentu (všech 150 metrik), aby na ně mohl někdo jiný lehce navázat?

Instrukce

Splnění zadání

Posudte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posudte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posudte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.