

# Posudek oponenta bakalářské práce

Autor/autorka práce: **Tomáš Pašek**

Název práce: **Automatické zpracování dat z meteostanic**

## Obsah práce

Aplikace má sloužit ke zpracování totožných meteorologických dat uložených v různých formátech – získaných z různých meteostanic. Práce se soustředí především na implementaci nástroje, který takové zpracování umožní. V textu práce mi chybí kontext – k jakému účelu mají regulární výrazy sloužit. U podobné práce bych čekal, že půjde o hromadné zpracování různých datových zdrojů, ale ukázková PHP aplikace tomu nenasvědčuje.

## Kvalita řešení a dosažených výsledků

Popis regulárních výrazů mi připadá matoucí, postavený na několika vybraných příkladech a bez zmínění základních vlastností / operací / operátorů (např. při popisu „skupinové konstrukce“ je zároveň demonstrován operátor „nebo“ i zápis libovolného malého písmene). Přijde mi také zvláštní mluvit u regulárního výrazu o tom, že něco vrátí, i když chápu, co je tím myšleno. Podobně je na tom i popis Xpath jazyka – pokud čtenář jazyk vůbec nezná, z popisu mu příliš neporozumí, pokud ho zná, tento text nepotřebuje. U kapitol 2 a 3 bych spíš očekával důraz na to jak tyto jazyky použít k získávání dat s konkrétními příklady, místo pokusu o jejich velmi stručný popis.

Kapitola 4 je pro mě nesrozumitelná – abych si udělal obrázek, potřeboval bych nejspíš vidět příklady vstupních souborů a dat jaká se v nich budou hledat. Není mi jasné, co je myšleno hned úvodní větou – „Aby se zabránilo nějakým nejednoznačnostem, bude nutné prakticky vždy popsat celý vstupní soubor“ – čekal bych spíš, že se budete v souboru snažit hledat nějaké konstrukce odpovídající strukturou požadovaným datům. Chybí mi zde rešerše existujících datových souborů, které máte zpracovávat (zejména když se pak v dalších kapitolách odkazujete na jejich typickou nebo očekávanou podobu nebo na problémy které v nich mohou nastat).

Popis aplikace se takřka okamžitě začne zabývat implementačními detaily, aniž by byla vidět nějaká analýza – co všechno má aplikace dělat, kdo a jak ji bude používat, jakou strukturu má aplikace mít?

Při testování je důraz kladen prakticky jen na rychlost zpracování, nenarazil jsem na žádné výsledky, které by se zabývaly přesností nebo úplností zpracování dat automaticky generovanými výrazy. Nikde jsem také neviděl, jestli byly testy provedeny na datech z několika různých meteostanic, není tedy poznat nakolik je popsán postup univerzální.

Aplikace je spustitelná, ale na CD nejsou přiloženy žádné ukázky dat, takže je její otestování poměrně obtížné. GUI je celkem přehledné a velmi jednoduché, políčko pro regulární výraz je ale zbytečně malé a soudě podle popisu v práci se do něj často nemusí vejít. Třída Window by si rozhodně zasloužila důkladnější dekompozici, dělá celou řadu věcí – od vykreslení okna až po načítání některých souborů. Na druhé straně zvýraznění vybraných dat v okně je uděláno pěkně a přehledně. Podobným problémem trpí i PHP aplikace, která nabízí tři malá formulářová okna a u všech je možné očekávat, že data v nich budou řádově větší. PHP aplikace ale demonstruje, že pro validní vstupy automaticky vytvořené výrazy dokážou data správně rozpoznat.

Při pokusech s regulárními výrazy mi přišlo, že navržené výrazy v podstatě odpočítají slova (mezery) k hledanému řetězci a není mi jasné, v čem jsou takto vytvořené výrazy lepší než pozice dat zapsaná počtem mezer. Regulární výraz je navíc délkou úměrný délce souboru a jakákoliv změna v počtu mezer nebo doplnění další mezery na konec souboru vede k tomu, že přestane fungovat. Na druhé straně je možné že takový druh problému se ve vstupních datech neočekává (a případná kapitola věnující se rozboru zpracovávaných souborů by to osvětlila). Podobně XPath adresy spočívají jen v odpočítání pořadí uzlu, ve kterém jsou data uložena (a data nesmí být v attributech uzlů, s těmi si program neporadí – opět je možné že něco takového se vzhledem ke vstupním datům neočekává).

### **Formální úroveň**

Text práce je dobře strukturovaný, ale chybí v něm řada důležitých bodů, zejména podrobnější analýza možných vstupních dat a rozbor možné variability vstupů.

### **Práce s literaturou**

Literatura se omezuje na internetové zdroje (s výjimkou jedné knihy o regulárních výrazech), zejména na návody k jednotlivým technologiím. Všechny použité zdroje jsou relevantní a jsou řádně citovány.

### **Splnění zadání**

Všechny body zadání jsou splněny, s tím že vstupní data jsou popsána jen velmi zběžně a testování se nezabývá tím, jaké variability ve vstupních souborech jsou pro vytvořené výrazy ještě zvládnutelné.

### **Dotazy k práci**

Jakým způsobem Vaše aplikace využívá Maven?

Proč jsou regulární výrazy vytvářeny v PHP syntaxi, ale aplikace, která je vytváří, je psaná v Javě?

Testoval jste možné výstupy z různých meteostanic, jak moc mohou být výstupní data variabilní? Pokud jsou ve výstupu stanice slovní popisy počasí, co se stane, když bude mít popis někdy dvě a někdy jedno slovo (např. „Falling Rainy“ vs. „Sunny“)

Vzhledem k uvedeným připomínkám navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni 22.5.2020

Ing. Richard Lipka, Ph.D.