Posudek oponenta bakalářské práce

Autor/autorka práce: **Gabriela Hessová**

Název práce: **Automatické získání historických údajů z webových zdrojů**

Obsah práce

Cílem práce bylo navrhnout a implementovat nástroj pro automatické získání vybraných historických údajů z webových zdrojů. Jako zdroj vyla vybrána otevřená encyklopedie Wikipedie, respektive dump její anglické verze, se kterým lze po stažení pracovat offline.

Autorka se v práci věnuje možnostem získávání údajů z textu a existujícím zdrojům historických údajů. Dále popisuje návrh a implementaci nástroje pro získávání historických dat a jeho testování.

Kvalita řešení (programová část bakalářské práce)

Vytvořený nástroj je funkční, ale není příliš uživatelsky příjemný, což není zásadní nevýhoda, vzhledem k tomu, že je určen k použití jednou za čas, ale přesto to není příjemné. Nelze například vybrat vstupní či výstupní soubor pomocí `JFileChooser`u, uživatel musí cestu a název specifikovat ručně. Při zadání vstupního souboru s příponou `.txt` (testovací soubor dostupný na DVD má tuto příponu) místo zřejmě očekávané přípony `.xml` se zneaktivní tlačítko *Pokračovat* a je nutné restartovat aplikaci. Informace o průběhu zpracování se navíc nemění příliš často, což může vést uživatele k dojmu, že aplikace zamrzla. Nástroj sestává z cca 26 zdrojových souborů (cca 61 kB). Zdrojový kód je celkem přehledný a je celkem komentován. Mnohé komentáře však chybí, především u proměnných třídy a instance. Mezi zdrojovými soubory se překvapivě nalézají dva textové (možná konfigurační?) soubory.

Kvalita řešení (text bakalářské práce a práce s literaturou)

Text sestává z 61 stran (řádkování cca 1.1) a má logickou strukturu. Poměr teoretické a praktické části je vyrovnaný. Text práce je vhodně členěn do kapitol a doplněn tabulkami, obrázky, úseky zdrojového kódu a poznámkami pod čarou. Autorka celkem pěkně popisuje vytvořený nástroj od návrhu, přes implementaci až po testování (včetně návrhů na možná vylepšení). V části návrh bych ale očekával přesnější specifikaci požadavků, diagram případů užití a lepší zdůvodnění některých rozhodnutí (např. nevyužití SAX na str. 20). V popisu implementace i v přílohách chybí UML diagramy tříd. ERA diagram vytvořené databáze rovněž chybí, je však zbytečný, protože databáze obsahuje pouze jednu tabulku a ta je v textu popsána dobře (str. 25). Testování je místy sice popsáno trochu vágně, ale provedené testy dobře ověřují různorodé vlastnosti vytvořeného nástroje. Přílohy tvoří uživatelská příručka a tabulky s nejčastějšími typy infoboxů.

K textu práce mám dále několik drobných výhrad. V práci jsou ve větách nejednotně používány pomlčky i spojovníky pro stejné účely (např. str. 17). Občas lze narazit na příliš krátký odstavec (např. str. 17) nebo nevhodnou formulaci/překlad (např. „defaultně“ na str. 25). V záhlaví Seznamu použitých zkratek (str. 38) je uvedeno „Závěr“.

Zdrojů je v práci lehce nadprůměrné množství (konkrétně 30) a jsou celkem důsledně odkazovány v textu. Mnohé odkazy jsou však umístěny až za poslední tečkou odstavce (např. str. 2, 4, 6 atd.), což se může zdát logické (pokud se odkaz vztahuje k celému odstavci), ale v technickém textu je to nezvyklé. Před odkazy uvnitř odstavců často chybí mezera (např. str. 2, 3 atd.).

Splnění zadání

Práce splňuje zadání.

Doplňující informace k bakalářské práci

Práce byla vytvářena v rámci projektu časové osy, který je vytvářen v rámci několika bakalářských a diplomových prací.

Dotazy k bakalářské práci

1. Co jsou to orientované označené grafy (viz str. 6 bakalářské práce)?
2. Jaké výhody a nevýhody by přineslo využití SAXu místo ručního čtení dumpu Wikipedie?
3. Lze očekávat, že nástroj bude správně fungovat i na novějších revizích dumpu Wikipedie?

Autorka práce vytvořila funkční nástroj pro získávání dat z dumpu Wikipedie. Vzhledem k problémům popsaným v posudku navrhuji známku **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 30.6.2015

Ing. Tomáš Potužák, Ph.D.