Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Tvorba rozsáhlých úložišť patentových dat

Plzeň 2022 Bc. Vojtěch Danišík

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V diplomové práci jsou použity názvy programových produktů, firem apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

V Plzni dne 13. dubna 2022

Bc. Vojtěch Danišík

Poděkování

Děkuji panu Doc. Ing. Daliboru Fialovi, Ph.D. za ochotu při vedení diplomové práce a rady s jejím vypracováním.

Abstract

Creation of large-scale patent data repositories. The aim of the diploma thesis is to get acquainted with the available national sources of patent data and to create extensive local repositories of patent data enabling their effective searching and mining. The first part of the thesis thoroughly describes the types of patents, existing national sources and file formats in which patents are stored. Subsequently, the applicable technologies for searching and mining are described. The second part of the thesis is devoted to the selection of usable data and the implementation of selected technologies. Several queries and scenarios have been created to test efficient mining. The results of the testing are part of this work.

Abstrakt

Cílem diplomové práce je seznámit se s dostupnými národními zdroji dat o patentech a vytvořit rozsáhlá lokální úložiště patentových dat umožňující jejich efektivní prohledávání a vytěžování. První část práce důkladně popisuje typy patentů, existující národní zdroje a formáty souborů, ve kterých se patenty ukládají. Následně jsou popsány použitelné technologie pro prohledávání a vytěžování. Druhá část práce se věnuje výběru použitelných dat a implementaci vybraných technologií. Pro otestování efektivního vytěžování bylo vytvořeno několik query a scénářů. Výsledky testování jsou součástí této práce.

Obsah

1 Úvod

zmínit deployment - docker atp, aby se to mohlo narvat do teorie

2 Patent

typy, uložiště dat + info o zdrojích (patentové úřady atp)
Formáty dat (JSON, CSV, XML, ...)

3 Databáze

Objektové, relační, síťová, hierarchická, ... -> popsat všechny
SQL, NoSQL, ...
Mongo, Mysql, Postgres, ...

4 Docker

5 Návrh struktury

6 Výběr dat

Zdroje (které ano, které ne + proč) + udělat stejnou tabulku jak v excelu + zmínit stránku ze který jsem čerpal informace (wipo.int)

Země	**Patentový úřad**	**Zkratka**
Anglie	Intellectual Property Office	IPO
Arménie	Intellectual Property Office	-
Austrálie	IP Australia	-
Bělorusko	National Center of Intellectual Property	NCIP
Bulharsko	Patent Office of Republic of Bulgaria	-
Česko	Industrial Property Office of the Czech Republic	-
Čína	China National Intellectual Property Administration	CNIPA
Dánsko	Danish Patent and Trademark Office	-
Egypt	Egyptian Patent Office	-
Estonsko	The Estonian Patent Office	-
Filipíny	Intellectual Property Office of the Philippines	IPOPHL
Finsko	Finnish Patent and Registration Office	PRH
Francie	National Institute of Industrial Property	INPI
Hong Kong	Intellectual Property Department	-
Chorvatsko	State Intellectual Property Office of the Republic of Croatia	SIPO
Indie	Office of the Controller General of Patents, Designs and Trade Marks	-
Indonésie	Directorate General of Intellectual Property	DGIP
Irsko	Intellectual Property Office of Ireland	IPOI
Island	Icelandic Intellectual Property Office	ISIPO
Israel	The Israel Patent Office	ILPO
Itálie	Directorate General for the Protection of Industrial Property	-
Japonsko	Japan Patent Office	JPO
Jižní Korea	Korean Intellectual Property Office	KIPO
Kanada	Canadian Intellectual Property Office	CIPO

Tabulka 6.1: Odřádkovací sekvence znaků

Země	Patentový úřad	Zkratka
Kuba	Cuban Industrial Property Office	OCPI
Litva	State Patent Bureau of the Republic of Lithuania	-
Lotyšsko	Patent Office of the Republic of Latvia	-
Maďarsko	Hungarian Intellectual Property Office	HIPO
Malajsie	Intellectual Property Corporation of Malaysia	MyIPO
Mexiko	Instituto Mexicano De La Propiedad Industrial	IMPI
Moldova	State Agency on Intellectual Property	AGEPI
Německo	German Patent and Trade Mark Office	DPMA
Nizozemsko	Netherlands Patent Office	-
Norsko	Norwegian Industrial Property Office	NIPO
Nový Zéland	Intellectual Property Office of New Zealand	IPONZ
Peru	National Institute for the Defense of Competition and Protection of Intellectual Property	INDECOPI
Polsko	Urząd Patentowy Rzeczypospolitej Polskiej	UPRP
Portugalsko	Portuguese Institute of Industrial Property	-
Rakousko	Austrian Patent Office	-
Rumunsko	State Office for Inventions and Trademarks	OSIM
Rusko	Federal Service for Intellectual Property	Rospatent
Řecko	Hellenic Industrial Property Organization	HIPO
Singapur	Intellectual Property Office of Singapore	IPOS
Slovensko	Industrial Property Office of the Slovak Republic	-
Slovinsko	Slovenian Intellectual Property Office	SIPO
Srbsko	Intellectual Property Office of the Republic of Serbia	-
Španělsko	Spanish Patent and Trademark Office	OEPM
Švédsko	Swedish Intellectual Property Office	PRV
Švýcarsko	Swiss Federal Institute of Intellectual Property	-
Turecko	Turkish Patent and Trademark Office	Turkpatent
Ukrajina	Ukrainian Intellectual Property Institute	Ukrpatent

Tabulka 6.2: Odřádkovací sekvence znaků

6.1 Atributy

6.1.1 Povinné atributy

Země	Název patentu	Rok přihlášky / patentu	Autor	ID patentu
Kanada	x	x	x	x
Česko	x	x	-	x
Litva	x	x	x	x
Portugalsko	x	x	x	x
Španělsko	x	x	x	x
Švédsko	-	x	-	x
Izrael	x	x	x	x
Itálie	x	x	x	x
Mexiko	x	x	x	x
Polsko	x	x	-	-
Anglie	x	x	x	x
Rusko	x	x	x	x
Peru	x	x	x	x
Francie	x	x	x	x

Tabulka 6.3: Odřádkovací sekvence znaků

6.1.2 Nepovinné atributy

Rovnou udělat kapitolu, ve který se aplikují všechny podmínky + se sepíše souhrn počtu patentů, z jakých zemí atp.

Země	Abstrakt	Slovník	Reference	Žadatel
Kanada	-	-	-	x
Česko	x	-	x	-
Litva	x	-	-	x
Portugalsko	x	-	-	x
Španělsko	x	-	-	x
Švédsko	x	-	-	-
Izrael	-	-	-	x
Itálie	-	-	-	x
Mexiko	x	-	-	x
Polsko	x	x	-	-
Anglie	-	-	-	-
Rusko	-	-	-	-
Peru	-	-	-	-
Francie	x	-	x	x

Tabulka 6.4: Odřádkovací sekvence znaků

Země	Adresa	Rodina patentů	Obor	Fulltext
Kanada	x	-	x	-
Česko	-	-	x	-
Litva	-	-	x	-
Portugalsko	-	-	x	-
Španělsko	x	-	x	x
Švédsko	-	-	x	x
Izrael	x	-	-	-
Itálie	-	-	-	-
Mexiko	-	-	x	-
Polsko	-	-	-	-
Anglie	-	-	x	-
Rusko	-	-	-	-
Peru	-	-	x	-
Francie	-	-	x	x

Tabulka 6.5: Odřádkovací sekvence znaků

7 Implementace

8 Rozšiřitelnost modulu

Zadání diplomové práce sice splněno bylo, ale v blízké budoucnosti mohou být požadavky na modul změněny. Jako příklad lze uvést podporu přidávání nových patentů do databází, zjištění autorů pro české patenty, automatické stahování dat z již ověřených patentových zdrojů. V této kapitole jsou popsány 3 možné návrhy na rozšíření modulu ohledně importu dat do již existujících databází.

8.1 Přidávání nových patentů

Cílem tohoto rožšíření by bylo automatické přidávání patentů z datových souborů jak do MySQL databáze, tak i do Mongo.

Rozšíření by se dalo realizovat jako aplikace ve vyšším programovacím jazyku (např. Java, C), kdy vstupem do aplikace by byl soubor v datovém formátu JSON/XML/CSV a jiné. Vstupní soubor by se následně:

- převedl na JSON řetězec (v případě že soubor není ve formátu JSON) a vložil do Mongo databáze
- rozparsoval a extrahovali by se všechny atributy, které se ukládají v MySQL databázi (viz mysql kapitola) TODO

Jelikož je dost časté, že každý národní zdroj dat používá odlišnou strukturu patentu, tak bude potřeba aplikaci neustále upravovat (ať už v rámci přidávání nových zdrojů, nebo v případě změny struktury patentu u již podporovaných zdrojů).

Jako další velký problém lze zmínit extrakci atributů patentu ze souborů. Tím, že různé patentové soubory mají odlišnou strukturu, to znamená hloubku zanoření specifických elementů, jiné názvy elementů, tak bude obtížné naimplementovat řešení extrakce pro všechny soubory. Tento problém by se dal řešit tak, že se vytvoří soubory se slovníkama, které by obsahovali názvy elementů pro daný atribut. Slovníky by se následně použily při extrakci.

8.2 Zjišťování autorů pro české patenty

Český národní patentový úřad poskytuje data o českých patentech, které ale neobsahují autora ani instituci. Pro zjištění autora nebo instituce, která

patent registrovala, je nutné použít oficiální vyhledávač. Cílem tohoto rozšíření by bylo vytvořit aplikaci ve vyšším programovacím jazyku, která se pro všechny české patenty bude snažit najít jejich autory za pomoci využití prohledávačů webů (web crawler). Postupů řešení může být mnoho:

- Zjišťování autorů by se provedlo pro všechny existující české patenty v databázi. Z MySQL databáze se zjistí všechny ID patentů pro české patenty, které se následně použijí jako vstup pro web crawler.
- Zjišťování autorů by se provedlo pro patent/y uložené v souboru, kdy aplikace by pro všechny patenty v souboru zjistila autory a následně je dopsala do příslušnýho elementu patentu v daném souboru.
- Stejný postup jako předchozí s tím rozdílem, že po zjištění autora se patent rovnou přidá do MySQL i Mongo databáze.

8.3 Automatické stahování dat z ověřených zdrojů

Aplikace, která si načte soubor s uloženýma ověřenýma zdrojema (např. XML), kdy by bylo definováno: název země, URL stránky kde se stahuje (např. https://isdv.upv.cz/webapp/webapp.pubsrv.seznam?purl=opendata/pt), html element obsahující soubor ke stažení (např. a href="/doc/opendata/pt/OpenData....") a poslední stažený soubor. Aplikace by projela všechny stránky, pokusila by se stáhnout nejnovější soubor (ten co se neshoduje s tím uloženým v xml) a uložit ho do specifický složky. Pokud je ke stažení novější, updatuje se XML.

Tohle všechny by bylo automatický pomocí např. Jenkinsu, který by jednou za čas spustil tuhle appku + by mohl informovat emailem o buildu + výsledku, že tolik se updatnulo atd.

9 Ověření efektivního vytěžování

K ověření efektivního vytěžování bylo připraveno několik scénářů jak pro MytextbfSQL, tak i pro Mongo + ElasticSearch.

9.1 Mongo + ElasticSearch

9.2 MySQL

Pro MySQL bylo připraveno 10 scénářů, které testují všechny vytvořené tabulky v databázi. Každý scénář obsahuje textový popis, SQL příkaz, rychlost vykonání příkazu, počet výsledků a OBRÁZEK VÝSLEDEK

9.2.1 Scénář č.1

Textový popis: Nejčastěji patentující instituce v Izraeli v roce 2015
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.2 Scénář č.2

Textový popis: Nejméně patentovaný obor v Kanadě od roku 2010
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.3 Scénář č.3

Textový popis: Obor s nejvíce patenty za rok 2008 ve Španělsku
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.4 Scénář č.4

Textový popis: Autor s největším počtem patentů ze všech zemí
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.5 Scénář č.5

Textový popis: Nejméně používaný jazyk pro patenty za rok 2003
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.6 Scénář č.6

Textový popis: Instituce / autor s patenty pokrývající největší množství oborů
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.7 Scénář č.7

Textový popis: Země s nejvíce patenty od roku 2000
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.8 Scénář č.8

Textový popis: Nejvíce používaný jazyk pro patenty ve Francii
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.9 Scénář č.9

Textový popis: Nejčastěji patentující instituce v Anglii v ekonomickém oboru za rok 2013
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

9.2.10 Scénář č.10

extbfTextový popis:
SQL:
Rychlost vykonání:
Počet výsledků:
OBRÁZEK VÝSLEDEK:

10 Závěr

A Uživatelská dokumentace

B Vzhled modulů