VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

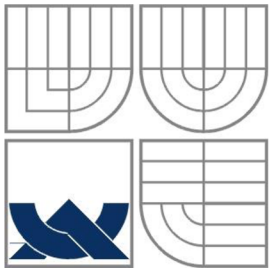
VYTVOŘENÍ ZNALOSTNÍ BÁZE ENTIT Z ČESKÉ WIKIPEDIE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

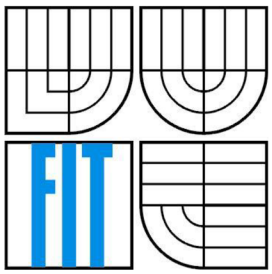
AUTOR PRÁCE
AUTHOR

MARTIN SYCHRA

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

VYTVOŘENÍ ZNALOSTNÍ BÁZE ENTIT Z ČESKÉ WIKIPEDIE

ENTITY KNOWLEDGE BASE CREATION FROM CZECH WIKIPEDIA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

MARTIN SYCHRA

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2014

Abstrakt

Cílem této práce je navrhnout a implementovat systém pro automatickou extrakci pojmenovaných entit z textů české Wikipedie, vytvořit znalostní báze těchto entit a vyhodnotit úspěšnost a výsledky vytvořeného systému. První část práce vysvětluje základní pojmy z této oblasti zpracování přirozeného jazyka a informuje o existujících systémech podobného charakteru. V ústřední části je popsán vlastní návrh několika metod extrakce a způsobu implementace těchto metod. K extrakci byly vybrány tyto entitní typy: osoby, místa, události a organizace. V závěru jsou popsány výsledky práce, tedy úspěšnost jednolitých metod u daného entitního typu a statistiky extrakce jednotlivých entit vztažené k celkovému složení české Wikipedie.

Abstract

The aim of this thesis is to propose and implement a system for an automatic extraction of named entities from Czech Wikipedia, to create a knowledge base consisting of these entities and to evaluate results of the created system. The first part explains basic notions of this field and discusses related work. The main part proposes several methods of extraction and details their implementation. The following types of entities are extracted: people, places, events and organizations. The final part of the thesis presents results, i.e., the success of the individual methods for each entity type and statistics on extraction of the individual entities in the whole Czech Wikipedia context.

Klíčová slova

Extrakce pojmenovaných entit, zpracování přirozeného jazyka, česká Wikipedie, znalostní báze, automatická extrakce

Keywords

Extraction of named entities, natural language processing, Czech Wikipedia, knowledge base, automatic extraction

Citace

Sychra Martin: Vytvoření znalostní báze entit z české Wikipedie, bakalářská práce, Brno, FIT VUT v Brně, 2014

Vytvoření znalostní báze entit z české Wikipedie

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D.
Další informace a pomoc mi poskytl Ing. Lubomír Otrusina.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Martin Sychra
21. května 2014

Poděkování

Děkuji vedoucímu, doc. Smržovi, za odborné vedení a pomoc při tvorbě bakalářské práce. Děkuji také Ing. Otrusinovi za významnou pomoc v počátečních fázích a celé Skupině znalostních technologií na Fakultě informačních technologií Vysokého učení technického v Brně, v rámci které bylo na tomto projektu pracováno, za umožnění vzniku projektu a za možnost využití výpočetních prostředků a zdrojů.

Obsah

1 Úvod

Zpracování přirozeného jazyka je velmi aktuální a rozvíjející se obor komputační lingvistiky, oblasti na pomezí informatiky a lingvistiky. Jednou z významných oblastí zpracování přirozeného jazyka je extrakce informací, tedy snaha o vyjmutí určitého druhu informace z textu v přirozeném jazyce do formy vhodné pro zpracování počítačem. Poznatky z oboru extrakce informací se využívají v mnoha aplikacích a nástrojích. Příkladem může být tvorba textových vyhledávačů či automatický strojový překlad.

Významnou podoblastí extrakce informací je extrakce pojmenovaných entit (a jejich rozpoznávání v textu), tedy názvů konkrétních lidí a věcí: jména osob či organizací, názvy geografických lokací či historických událostí, pojmenování konkrétních produktů a výrobků apod. Úkolem této disciplíny je na jedné straně získání strukturovaných dat o pojmenovaných entitách z textů v přirozeném jazyce, tedy např. z článku o určité entitě je možné zjistit její název, druh a další základní informace. Dále se tato disciplína zabývá rozpoznáváním těchto entit v textu, tj. identifikací názvů entit v rámci vět či celých textů, rozhodováním, o jakou konkrétní entitu se v daném kontextu jedná, či nalezením dalších odkazů na tutéž entitu v daném textu.

S těmito úkoly souvisí tvorba znalostních bází – v nich jsou uloženy informace o entitách, které byly dříve z textů získány, a z nich se také informace získávají při rozpoznávání již známých entit v nových textech. Znalostní báze mohou vznikat na základě zpracování určitých zdrojů informací. Příkladem zdroje může být např. nejrozšířenější svobodná webová encyklopedie – Wikipedie.

V předkládané bakalářské práci je popsán vývoj jedné znalostní báze, a to na základě extrakcí pojmenovaných entit z české verze Wikipedie, která dává pravidelně k dispozici ke stažení celý obsah ve zdrojové podobě. Tento obsah lze analyzovat a za pomoci určitých metod v něm vyhledat pojmenované entity, kategorizovat je a podle jejich druhu k nim extrahovat několik základních informací. Z těchto informací lze posléze vytvořit znalostní bázi, která je využitelná v dalších projektech.

V práci se nejprve pokusíme teoreticky vymezit oblast extrakce pojmenovaných entit z textu a představit významné projekty, které se touto oblastí zpracování přirozeného jazyka v současné době zabývají (srov. např. projekt DECIPHER[1]).

Stěžejní část práce představí návrh vlastního systému pro extrakci pojmenovaných entit ze zdrojových souborů české Wikipedie. Práce popisuje několik různých metod, kterými je možné pojmenované entity z textů extrahovat – tj. jak lze klasifikovat konkrétní článek české Wikipedie a zařadit ho k určitému entitnímu typu. Některé z těchto metod se také zabývají extrakcí základních informací o dané entitě.

Práce zároveň rozebírá koncepci výsledného programu, tj. jak budou jednotlivé metody fungovat v rámci systému, z jakých komponent se bude systém skládat a za co budou jednotlivé komponenty zodpovědné.

Práce následně představuje implementaci programové části, tj. přináší ucelený popis implementace celého systému a podrobněji se zaměřuje na několik vybraných částí. V této části je také prezentována implementace navržených metod v rámci extrakce jednotlivých entit.

V další části práce jsou představeny a vyhodnoceny výsledky činnosti sestrojeného systému. Je zde podrobně popsána extrakce čtyř vybraných entitních typů: osob, míst (tj. států a sídel, jako jsou např. města, vesnice apod.), událostí a organizací, dále jsou uvedeny počty a statistiky extrahovaných

[1] http://decipher-research.eu/, viz kapitola 3.3.

entit a jejich atributů, tj. základních informací o těchto entitách. Práce poté vyhodnocuje vhodnost všech použitých metod pro extrakci konkrétního typu pojmenované entity.

Statistiky jsou přitom dány do kontextu celé české Wikipedie – je tedy popsáno, jakou část tvoří jednotlivé entity z celkového množství článků na Wikipedii a jaký typ článků tvoří zbylé texty (tj. články, které nebyly extrahovány).

Závěrečná část práce celkově shrnuje dosažené výsledky automatické extrakce pojmenovaných entit z textů české Wikipedie a představuje možnosti jejich využití v dalších navazujících projektech, jako je rozpoznávání pojmenovaných entit v textu či doplnění znalostních bází z jiných zdrojů.

2 Základní pojmy a použité zdroje

Tato kapitola přináší popis základních pojmů týkajících se oblasti zpracování přirozeného jazyka, zejména extrakce pojmenovaných entit z textu. Dále popisuje zdroj dat pro vytvářenou znalostní bázi, tj. internetovou encyklopedii Wikipedii. Zaměříme se přitom na českou verzi této encyklopedie, která je v našem systému využita.

2.1 Pojmenované entity v kontextu NLP

Zpracování přirozeného jazyka (dále NLP[2]) řeší mnoho úloh souvisejících s komunikací mezi člověkem a strojem (počítačem). Porozumění přirozenému jazyku umělou inteligencí je velmi náročný a komplexní problém, který se dotýká mnoha oblastí a je v současné době jedním z nejaktuálnějších úloh umělé inteligence (srov. např. kapitola 1.3 v [1] či předmluva a první kapitola v [2]).

NLP zahrnuje různé oblasti, jako je např. automatické překládání textu, sumarizace textu, vyhledávání klíčových slov apod. V této kapitole se zaměříme na jednu z nich, a to extrakci informací z textu, konkrétněji pak extrakci pojmenovaných entit. Podrobně představíme její cíle a vysvětlíme základní pojmy spojené s touto oblastí, jako jsou např. pojmenované entity či tvorba znalostních bází v kontextu NLP a počítačové lingvistiky.

Extrakce informací z textu (dále IE[3]) je široká oblast zahrnující postupy automatického získávání strukturovaných dat (vhodných pro logické uchování a zpracování) z nestrukturovaného textu (tedy z textu v přirozeném jazyce). Jejím cílem je nalézt v textu hodnoty pro položky určité předdefinované šablony, resp. přiřadit sémantické kategorie vybraným prvkům v textu ([1] str. 376 v kapitole 10.6.2). Při procházení textem se snaží identifikovat činitele a další důležité informace (str. 122, kapitola 3.3 tamtéž) či se pokouší o identifikaci určitého druhu informace v různých podobách a formátech (str. 131 v kapitole 4.2.2 tamtéž).

Další pohled na IE najdeme v [3], kde je problematika vysvětlena na obecném systému pro automatickou extrakci. Vstupem takového systému je přirozený text a výstupem shrnutí textu na základě specifikovaného tématu či oblasti. Systém nalezne v textu významné informace a převede je do strukturované podoby vhodné pro uchování v databázi a další použití.

V rámci této práce se jedná o extrakci specifického druhu informace: pojmenovaných entit. Entita je termín pro označení určitého objektu reálného světa, tj. něčeho existujícího[4]. Pojmenovaná (též jmenná) entita je pak pojem reprezentující určitý konkrétní objekt, který má jméno (v češtině a dalších jazycích začínající velkým písmenem), tedy např. jména osob, měst či organizací [4]. Základními podproblémy NER (rozpoznávání pojmenovaných entit[5]) je jejich určení v textu (na kterém místě se nachází výraz odpovídající pojmenované entitě) a určení typu entity a konkrétní entity v rámci tohoto typu (určení typu entity a následně konkrétní entity v rámci typu lze souhrnně

[2] Zkratka anglického pojmu *natural language processing*.

[3] Vychází z anglického pojmu *information extraction*.

[4] Tato definice vychází z několika slovníků cizích slov dostupných on-line, např. http://www.slovnik-cizich-slov.net/, http://www.slovnik-cizich-slov.cz/ či http://slovnik-cizich-slov.abz.cz/.

[5] NER je zkratka anglického *named entity recognition*.

zahrnout pod termín disambiguace [4]). Jinými slovy jde o důležitou součást IE, která spočívá v identifikaci frází, které reprezentují název určité entity, a v přiřazení těchto frází do určitých tříd [5].

K řešení těchto problémů lze využít nástroje, které obsahují informace o pojmenovaných entitách, tedy různé slovníky či znalostní báze. Znalostní bázi chápeme v kontextu NLP jako strojově čitelný zdroj informací, tedy informací strukturovaně uložených. Často bývá součástí systémů umělé inteligence, např. tzv. expertních systémů [6].

2.2 Wikipedie a její česká verze

Extrakce informací je prováděna vždy z určitého zdroje dat. Výběr tohoto zdroje souvisí s druhem informací, které chceme získat, ale především se způsobem, jakým budeme data extrahovat. V předkládané bakalářské práci byla jako zdroj informací použita webová encyklopedie Wikipedie[6].

Wikipedie je otevřená internetová encyklopedie obsahující články v mnoha světových jazycích, jejíž hlavním cílem je volné šíření informací z nejrůznějších oblastí a oborů.

Jak již bylo zmíněno, hlavní charakteristickou vlastností Wikipedie je její otevřenost – její obsah vytváří široká veřejnost, články může psát či upravovat libovolný uživatel internetu. Tato skutečnost je důležitá i z hlediska automatického zpracování jejího textu. Výhodou je významný objem dat, který by nebyl schopen nashromáždit úzký okruh autorů (v současné době anglická verze Wikipedie obsahuje téměř 4 500 000 článků). Další předností otevřenosti Wikipedie je skutečnost, že jakýkoli její uživatel má možnost jednotlivé články upravovat či rozšiřovat, pokud původní obsah článku není dostačující.

Negativní stránkou otevřenosti Wikipedie je např. vyšší riziko chyb a nesprávných informací (články může do encyklopedie vkládat jakýkoli její uživatel bez ohledu na vzdělání či informovanost v daném oboru), a to i přes možnou kontrolu a opravu textů ostatními uživateli. Další nevýhodou (zejména pro automatické zpracování) je nejednotnost, a to především v oblasti stavby článků, zařazení do kategorií apod. I přesto, že existuje řada různých šablon, norem a doporučení, velké množství autorů se jimi neřídí. Pro některé prvky (např. řazení do kategorií či šablony článků) navíc normy či doporučení nemusí vůbec existovat.

V této práci byla zpracovávána česká verze Wikipedie[7], která sice neobsahuje takové množství dat jako verze anglická (přibližně 300 000 článků), nicméně lze z ní získat mnoho informací specifických pro český kontext, které v jiných jazykových verzích (alespoň v takovém rozsahu) nenajdeme. Díky tomu může vzniklá znalostní báze doplnit některou již existující bázi vytvořenou na základě Wikipedie (či jiného zdroje) pro jiný jazyk.

[6] http://www.wikipedia.org/

[7] http://cs.wikipedia.org/

3 Související projekty

Extrakcí pojmenovaných entit z nestrukturovaného textu se zabývá mnoho výzkumných pracovišť a existuje celá řada menších či větších projektů, z nichž některé pracují také s Wikipedií jako základním zdrojem dat. Příkladem může být projekt [7] prezentovaný v roce 2007 na konferenci EMNLP-CoNLL, kde J. Kazama a K. Torisawa představili, jak lze využívat úvodní věty článků na Wikipedii pro zlepšení činnosti systému pro rozpoznání pojmenovaných entit. Informace z této věty lze použít jako základní kategorizaci dané entity.

Dalším zajímavým projektem je měření a porovnávání několika systémů v této oblasti, na kterém pracují G.Rizzo, M. van Erp a R. Troncy [8]. Zde je prezentován především fakt, že každý ze systémů může být zaměřen (a tedy být úspěšnější) na jinou specifickou doménu (v jiném kontextu). Nejlepších výsledků je tedy dosaženo kombinací více takových extraktorů, které spolupracují v rámci jednoho systému – tento systém je autory prezentován a srovnáván se samostatně pracujícími systémy.

V následujících podkapitolách budou představeny a popsány tři významnější projekty, které se tématu práce dotýkají.

3.1 DBpedia

DBpedia[8] je rozsáhlý webový projekt, který je přímo součástí projektu Wikipedie. Zatímco úkolem Wikipedie je získávat a uchovávat nestrukturovaná data od uživatelů, činností DBpedie je získávání, uchovávání a zveřejňování strukturovaných data z Wikipedie. Data jsou extrahována na základě infoboxů[9] článků Wikipedie. DBpedia dává data volně k dispozici, a to jak ve formě celé báze ke stažení (resp. určité její verze), tak i on-line prostřednictvím SPARQL bodu.

Součástí tohoto projektu je i funkce DBpedia Spotlight, což je nástroj pro rozpoznávání a disambiguaci pojmenovaných entit v textu. Entity v textu identifikuje a přiřadí k nim konkrétní entitu z báze DBpedie (resp. Wikipedie). V případě více možných kandidátů vybírá nejvhodnější entitu na základě kontextu v dané větě.

3.2 Airpedia

Dalším souvisejícím projektem, který získává a dává volně k dispozici strukturovaná data získaná především z Wikipedie, ale i dalších zdrojů (podobné webové zdroje, které obsahují nestrukturované informace), je Airpedia[10]. Tento projekt navazuje na DBpedii – z ní přebírá základní data získaná z infoboxů Wikipedie. Poté je použito několik algoritmů k procházení a zkoumání článků, ze kterých byla data získána, a na základě získaných dat se systém učí automaticky získat data i jiným způsobem z textu článků, aby bylo následně možné získat data i ze článků, které infobox neobsahují (či dokonce z článků mimo Wikipedii).

[8] http://dbpedia.org/

[9] Více o této metodě v kapitole 4.2

[10] http://www.airpedia.org/

Prozatím je projekt ve fázi mírného rozšíření dat z DBpedie. Veškerá data jsou opět dostupná jak on-line – pomocí technologie SPARQL, tak také kompletně ke stažení.

3.3 DECIPHER

DECIPHER je zkratka z anglického Digital Environment for Cultural Interfaces; Promoting Heritage, Education and Research. Jedná se o mezinárodní projekt podpořený Evropskou komisí, na kterém spolupracuje výzkumná skupina z FIT VUT[11] spolu s několika výzkumnými pracovišti z Irska a Spojeného království (srov. např. Dublin Institute of Technology, National Gallery of Ireland aj.). Informace o tomto projektu lze nalézt v první řadě z webových stránek tohoto projektu[12] či např. ze zveřejněných a dostupných zpráv (např. [9]).

Projekt je zaměřen na získávání vědomostí a výzkum kulturního dědictví na základě příběhů a vyprávění. Ke zpracování nestrukturovaného textu ve vyprávěních je zapotřebí metod extrakce informací a právě na této části projektu pracuje výzkumná skupina z FIT VUT, především sestavením systému NER (Named Entity Recognizer) pro automatické rozpoznávání pojmenovaných entit v nestrukturovaném textu.

[11] Fakulta informačních technologií Vysokého učení technického v Brně.

[12] http://decipher-research.eu/

4 Návrh systému pro extrakci

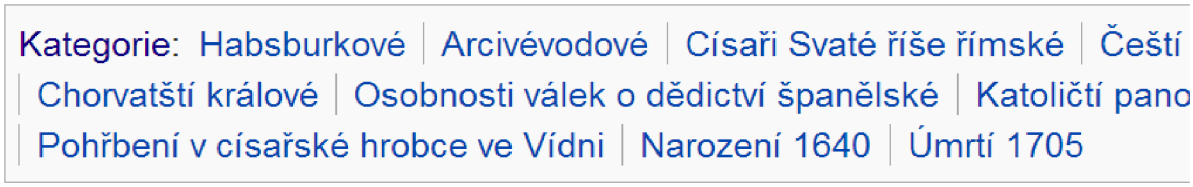
Tato kapitola popisuje návrh vytvářeného systému pro automatickou extrakci pojmenovaných entit ze zdrojových souborů Wikipedie. Bude navrženo a představeno několik různých způsobů, na základě kterých lze u článků na Wikipedii rozlišit entitní typ, tj. jak je možné jednotlivé typy entit z Wikipedie extrahovat. Tyto metody extrakce budou podrobně vysvětleny na konkrétních příkladech a zároveň budou zvažovány výhody a nevýhody jednotlivých metod a jejich vhodnost pro extrakci zvolených entitních typů (osoby, místa, události, organizace).

Posléze bude popsán návrh výsledného systému – ten bude tvořen především několika funkčními bloky, které budou odpovídat navrženým metodám extrakce.

Ve většině příkladů, na kterých budou vysvětleny jednotlivé metody extrakce, budou použity snímky a úryvky z článku o Leopoldovi I.[13] a dalších článků z české Wikipedie[14].

4.1 Extrakce na základě kategorií

Kategorie[15] tvoří základní klasifikaci článků na Wikipedii, a lze je tedy využít při rozlišování jmenných entit. Každý článek je přiřazen do jedné či více kategorií, do kterých podle názoru autora patří – kategorie je tedy v podstatě množina článků s podobným tématem. Seznam kategorií, do kterých je článek přiřazen, je uváděn v dolní části stránky článku jako skupina odkazů na jednotlivé kategorie. Příklad zobrazení kategorií u článku je na obrázku níže (Obrázek 4.1).



Obrázek 4.1: Zobrazení kategorií u článku Leopold I.

Vnitřní odkazy v rámci Wikipedie (tedy odkazy na nějaký jiný článek na Wikipedii) mají ve zdrojové podobě tento tvar:

```
[[název odkazované stránky na Wikipedii]]
```

Seznam odkazů na kategorie u určitého článku vypadá tedy ve zdrojovém textu článku např. tak, jak je ukázáno na následujících řádcích:

```
[[Kategorie:Habsburkové]]
[[Kategorie:Arcivévodové]]
[[Kategorie:Císaři Svaté říše římské]]
...
```

[13] http://cs.wikipedia.org/wiki/Leopold_I.

[14] http://cs.wikipedia.org/

[15] http://cs.wikipedia.org/wiki/N%C3%A1pov%C4%9Bda:Kategorie

```
[[Kategorie:Narození 1640]]
[[Kategorie:Úmrtí 1705]]
[[Kategorie:Muži]]
```

Prvním způsobem extrakce s využitím kategorií je tedy nejprve určit kategorii (kategorie), která obsahuje články hledaného entitního typu, a následně vyhledat všechny články, které v seznamu odkazů mají zvolenou kategorii uvedenou. Tento postup však není pro praxi příliš využitelný – každou kategorii je nutno určit zvlášť, nelze automaticky prohledávat strukturu kategorií do hloubky (tj. prohledávat i jejich podkategorie).

Kategorie tvoří velmi složitou stromovou strukturu – jak svou hloubkou, tak šířkou (rozvětveností). Každá kategorie může obsahovat velké množství článků, ale i dalších podkategorií. Například v kategorii Panovníci českého státu[16] žádného konkrétního panovníka nenajdeme, ti jsou zařazeni až v podkategoriích – např. kategorie Habsburkové[17] obsahuje jednotlivé osoby – panovníky, jako je Leopold I.

Doposud byl pro extrakci zvažován pouze soubor *cswiki-latest-pages-articles.xml*, který obsahuje základní podobu článků, tedy jednotlivé stránky Wikipedie ve zdrojové podobě. Tento soubor lze využít pro většinu operací v rámci automatické extrakce, u metody kategorií však pouze k prvnímu naznačenému způsobu, který není příliš vhodný (nelze se zanořovat do podkategorií). V úplném výpisu obsahu[18] (pro jednoduchost bude dále v textu používán anglický termín dump) lze však nalézt i další soubory, pro práci s kategoriemi je významný soubor *cswiki-latest-categorylinks.sql*. Tento SQL skript obsahuje data z tabulky veškerých odkazů na libovolnou kategorii[19]. Lze z něj tedy zjistit obsah jakékoliv kategorie, a to nejen informaci o všech článcích, které do této kategorie patří, ale i o všech podkategoriích, které daná kategorie zahrnuje. Lze tedy sestavit kompletní strom kategorií a pracovat s ním.

Složitost větvení a zanoření je velkou nevýhodou této metody. Před samotnou extrakcí je zapotřebí ručně listovat kategoriemi a hledat, ve kterých a na které úrovni zanoření se nacházejí články hledaného entitního typu. Při slepém zanoření do maximální hloubky budou často zahrnuty i články, které k danému typu nepatří. Příkladem tohoto jevu může být kategorie Města na Moravě[20], která obsahuje 128 článků o jednotlivých městech (tedy o entitách typu místo, konkrétněji město), zároveň také 123 podkategorií – téměř každé město má kromě článku také vlastní kategorii, ve které najdeme články týkající se daného města, např. o památkách atd.

4.2 Extrakce s využitím infoboxů

Infobox[21] je prvek, který umožňuje strukturovaně zapsat a zobrazit základní informace o článku na Wikipedii. Je to šablona, kterou autor článku zkopíruje, vloží do textu svého článku a

[16] http://cs.wikipedia.org/wiki/Kategorie:Panovn%C3%ADci_%C4%8Desk%C3%A9ho_st%C3%A1tu

[17] http://cs.wikipedia.org/wiki/Kategorie:Habsburkov%C3%A9

[18] Kopie veškerého obsahu, kterou dává Wikipedie pravidelně k dispozici ke stažení. Nejnovější výpis obsahu české Wikipedie je dostupný na adrese http://dumps.wikimedia.org/cswiki/latest/

[19] http://www.mediawiki.org/wiki/Manual:Categorylinks_table

[20] http://cs.wikipedia.org/wiki/Kategorie:M%C4%9Bsta_na_Morav%C4%9B

[21] http://cs.wikipedia.org/wiki/N%C3%A1pov%C4%9Bda:Infobox

následně vyplní její pole. Na stránce se zobrazuje vpravo od textu článku jako přehledná tabulka (ukázka zobrazení a porovnání se zdrojovou podobou infoboxu viz Obrázek 4.2). Vzhledem k tomu, že existuje mnoho různých šablon (podle typu článku), je úkolem autora vybrat a použít tu nejvhodnější podle typu vkládaného článku.

Leopold I.

císař Svaté říše římské, král český a uherský etc.

Leopold I. na dobovém portrétu.

Doba vlády	1657 – 1705
Korunovace	1656
Narození	9. června 1640 Vídeň
Úmrtí	5. května 1705 (64 let) Vídeň
Předchůdce	Ferdinand III.
Nástupce	Josef I.
Manželky	I. Markéta Habsburská II. Klaudie Felicitas

```
{{Infobox panovník
| jméno              =Leopold I.
| titul              =císař Svaté říše římské, k
| obrázek            =[[Soubor:Kaiser-Leopold1.j
| popisek            =Leopold I. na dobovém port
| vláda              =[[1657]] – [[1705]]
| korunovace         =[[1656]]
| tituly             =
| celé jméno         =
| předchůdce         =[[Ferdinand III. Habsburs
| následník          =[[Josef I. Habsburský|Jose
| partner1           =I. [[Markéta Habsburská (1
| partner2           =II.  [[Klaudie Felicitas
| partner3           =III. [[Eleonora Magdalena
| potomstvo          =[[Josef I. Habsburský|Jose
| rod                =[[Habsburkové]]
| hymna              =
| motto              =
| otec               =[[Ferdinand III. Habsburs
| matka              =[[Marie Anna Španělská (16
| narozen            =[[9. červen|9. června]] [[
| místo narození     =[[Vídeň]]
| úmrtí              =[[5. květen|5. května]] [[
| místo úmrtí        =[[Vídeň]]
```

Obrázek 4.2: Infobox – zobrazení na stránce a zdrojová podoba

Infoboxy jsou využity jako další ze způsobů rozpoznání článků o jmenných entitách a také k extrakci jejich základních atributů. Ke klasifikaci článku lze infoboxy využít tak, že zjistíme, jaký infobox článek obsahuje, a podle druhu tohoto infoboxu určíme typ entity, které se text článku týká. Pokud stránka infobox má, lze z něj také získat základní informace o dané entitě, tedy atributy této entity.

Metoda klasifikace článků pomocí infoboxů je velmi přesná. Pokud určíme vhodné infoboxy, články, které je obsahují, budou s velkou pravděpodobností náležet ke zvažovanému entitnímu typu (výjimkou může být skutečnost, kdy autor vložil do článku nevhodný infobox). Touto metodou lze také získat atributy dané entity: pokud určíme vhodná pole infoboxu, automatická extrakce těchto polí již není díky strukturovanému zápisu složitá (viz **Obrázek 4.2** – zdrojová podoba infoboxu).

Značnou potíží při extrakci entit touto metodou je velká rozmanitost a nejednotnost infoboxů – existuje mnoho druhů infoboxů a ty mají různé názvy atributů. Jedná se navíc pouze o šablonu, každý uživatel si tedy může infobox upravit podle svých představ (např. přidat či přejmenovat atributy) nebo dokonce vytvořit úplně nový infobox, pokud mu nevyhovuje žádný z již existujících. Další nevýhodou extrakce s využitím infoboxů je skutečnost, že infobox má na Wikipedii jen určité množství článků (zatímco např. kategorie by měly zahrnout všechny hledané články).

Metoda infoboxů je tedy spíše doplňující a slouží především k získání atributů entit.

4.3 Extrakce z textu článků

Třetí navrženou metodou extrakce je klasifikace článku podle vlastního textu. Využívá se syntax úvodní věty článku – ta je pro některé entity charakteristická, např. pro osoby často obsahuje jméno osoby napsané zvýrazněně („tučný text“), za ním jsou ve většině případů v závorkách základní údaje (datum, popř. místo narození a úmrtí) a text za závorkou obvykle pokračuje tvarem pomocného slovesa „být“ (např. „je“ nebo „byl/a”) následovaným základním, většinou stručným, až heslovitým popisem osoby (např. národnost a povolání). Příklad úvodní věty článku o osobě je na obrázku níže (viz **Obrázek 4.3**), ve zdrojové podobě vypadá úvodní věta článku takto:

```
'''Ladislav Michalík''' (* [[6. listopad|6. listopadu]] [[1941]]) je
bývalý [[Češi|český]] [[fotbalista]], [[Fotbalový útočník|útočník]].
```

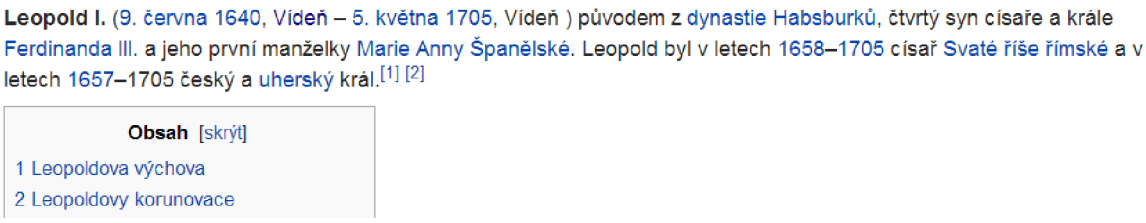
Díky charakteristické syntaxi lze pomocí první věty článku rozpoznat, o jakou entitu se jedná, v některých případech je možné navíc v další fázi extrakce získat základní údaje o entitě (tj. např. u zmíněných osob datum a místo narození, popř. úmrtí, stručný popis osoby aj.).

Obrázek 4.3: Úvodní věta článku se syntaxí charakteristickou pro osoby

Ne vždy však bude automatické zpracování úvodní věty úspěšné – hned v příkladu na obrázku (Obrázek 4.3) v závorce za jménem není uvedeno místo narození, pouze datum. U článku o Leopoldovi I. (viz Obrázek 4.4) jsou sice v závorce všechny informace obsaženy (včetně data a místa úmrtí), nicméně nenásledují stručné informace uvozené tvarem slovesa „být“. Syntax je zde tedy

odlišná. U jiných článků však nemusí být dodržena ani jedna z uvedených konvencí a tvar úvodní věty může být prakticky libovolný, obzvláště u jiných entitních typů. Tato metoda je tedy významná především pro typ osoba, pro další typy je použitelná s výrazně menším úspěchem.

Využití metody pro extrakci jmenných entit spočívá ve zvolení určitého regulárního výrazu, který by měl vyhovovat úvodním větám článků dané entity (současně ale žádným jiným), a následné aplikaci tohoto regulárního výrazu na úvodní věty všech článků na Wikipedii. Pokud regulární výraz vyhovuje, je daný článek zařazen do znalostní báze hledaného entitního typu. Dalším krokem je aplikace stejného (popř. upraveného) regulárního výrazu na úvodní věty již klasifikovaných článků (získaných touto metodou i metodami předchozími) za účelem získání atributů dané entity, tedy základních informací, které jsou v úvodní větě článku uvedeny.


Leopold I. (9. června 1640, Vídeň – 5. května 1705, Vídeň) původem z dynastie Habsburků, čtvrtý syn císaře a krále Ferdinanda III. a jeho první manželky Marie Anny Španělské. Leopold byl v letech 1658–1705 císař Svaté říše římské a v letech 1657–1705 český a uherský král.[1] [2]

Obsah [skrýt]

1 Leopoldova výchova

2 Leopoldovy korunovace

Obrázek 4.4: Úvodní věta článku o Leopoldovi I.

Prvním problémem tohoto postupu je co nejpřesněji identifikovat úvodní větu článku. Velmi často to totiž není první věta textu článku ve zdrojové podobě, může jí předcházet například mnohořádkový infobox (infoboxy se sice zobrazují vpravo od článku, podle konvence se ale ve zdrojovém textu umisťují na samý začátek článku). Je tedy zapotřebí sestrojit konečný automat, který prochází zdrojový text Wikipedie, zajišťuje identifikaci úvodní věty článku a zároveň znalost názvu (titulu) daného článku. Případný infobox (či jiné konstrukce na začátku článku) je přeskočen a teprve pak je určena první věta článku. Následně je na potenciální úvodní řádek aplikován regulární výraz a na základě něho vyhodnoceno, zda lze daný článek přiřadit ke hledané entitě.

Stěžejním problémem je však určení vhodného regulárního výrazu. Pokud zvolíme obecnější (volnější) výraz, nalezneme mnoho článků, které ke hledané entitě nepatří. Pokud však zvolíme naopak regulární výraz příliš konkrétní, nalezeme jen malý počet článků, protože (jak již bylo řečeno dříve) syntax úvodních vět není jednotná.

4.4 Překlad z anglické databáze

Čtvrtou a poslední metodou je doplnění báze o entity, které byly získány extrakcí z anglické verze Wikipedie, ale které mají svůj ekvivalent i v české verzi (a nebyly zahrnuty dosavadními metodami). Využití této metody je samozřejmě možné pouze tehdy, je-li k dispozici báze entit z jiného jazyka, což omezuje využití této metody. V této práci bylo využito báze, která je tvořena v rámci projektu DECIPHER (viz kapitola 3.2), konkrétně se jedná o soubor *KB.all*[22], který obsahuje aktuální verzi znalostní báze projektu DECIPHER NER.

Články (resp. entity) jsou samozřejmě v obou jazycích různě pojmenovány a často neexistují v obou jazykových verzích. Při řešení tohoto problému je využito odkazů mezi jazykovými verzemi,

[22] Tento soubor je v době tvorby této práce k dispozici ke stažení na http://athena3.fit.vutbr.cz/kb/KB.all

tzv. langlinks[23]. Seznam všech odkazů z určité jazykové verze Wikipedie lze nalézt v dalším souboru v dumpu, a to v souboru *cswiki-latest-langlinks.sql*. Opět se (stejně jako v případě categorylinks – viz kapitola 4.1) jedná o SQL skript, který obsahuje data z tabulky, ve které jsou uchovávány odkazy na články v jiných jazykových verzích. V této práci se zabýváme těmi, které odkazují na články z anglické verze Wikipedie.

4.5 Návrh výsledného systému

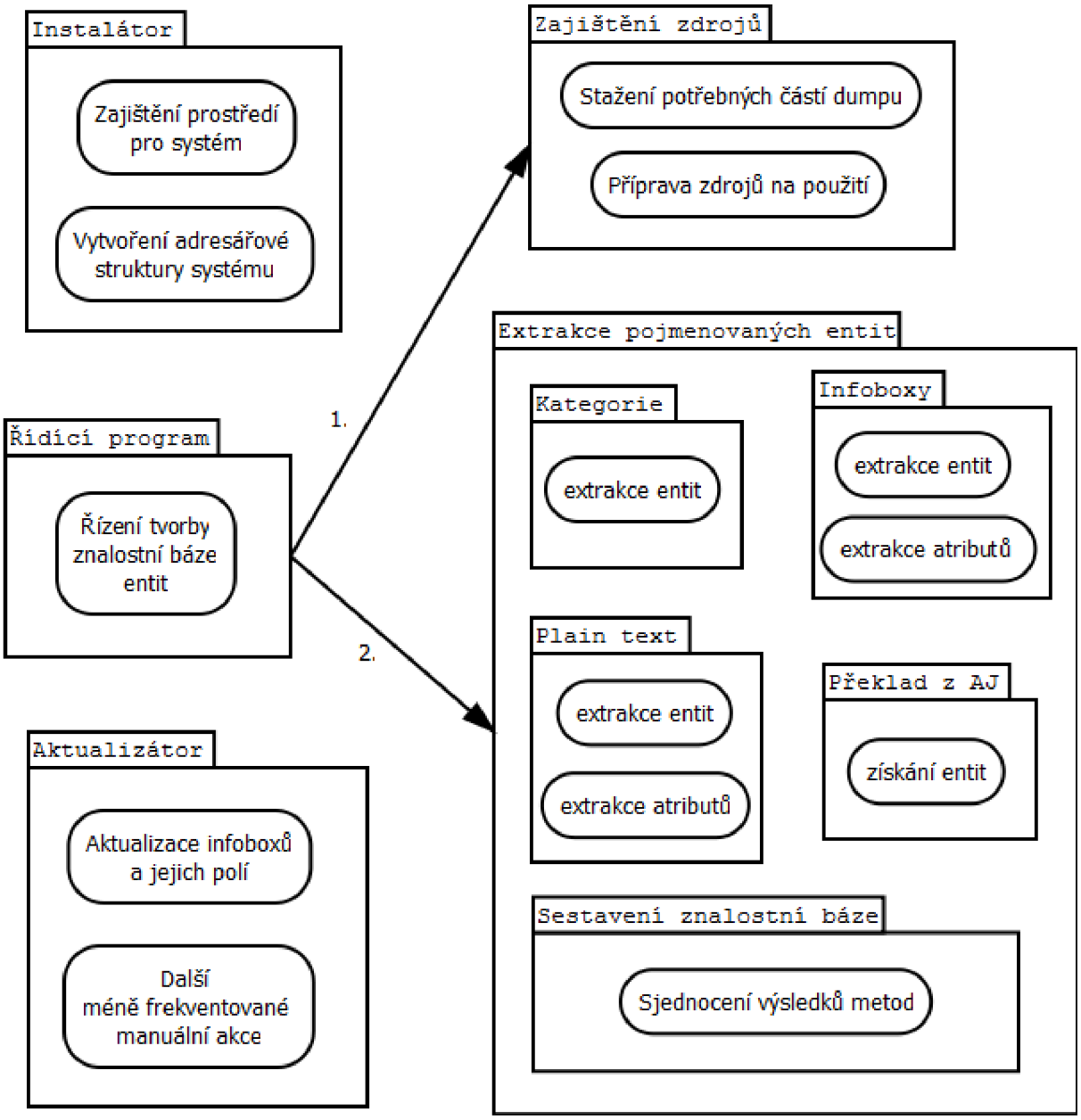
Výsledný systém je složen z několika komponent. Kompletní schéma systému je zobrazeno na obrázku pod touto podkapitolou (Obrázek 4.5) včetně popisu činnosti jednotlivých komponent.

Stěžejní částí je vlastní extrakce pojmenovaných entit – tato komponenta je složena z několika složek, z nichž každá je zodpovědná za jednu metodu extrakce. Před extrakcí pojmenovaných entit je však potřeba připravit nové zdrojové soubory – tedy stáhnout a připravit k použití potřebné části dumpu české Wikipedie či znalostní bázi z projektu DECIPHER pro získání entit z anglické Wikipedie. Tomuto úkolu se věnuje druhá komponenta. Obě tyto části systému jsou spouštěny řídícím programem, který tak zajišťuje jejich návaznost a korektní aktualizaci znalostní báze pojmenovaných entit.

V rámci hlavní komponenty systému, která bude entity extrahovat, se postupně na zdrojový soubor Wikipedie aplikují všechny navržené metody extrakce a poté jsou výsledky sjednoceny do jedné znalostní báze daného entitního typu.

Kromě výše zmíněných součástí jsou v systému další dva prvky: instalační a aktualizační. Instalátor je první část systému, kterou je zapotřebí spustit před všemi ostatními v prostředí, kde byl systém umístěn a kde ještě žádná extrakce neproběhla. Vytvoří adresářovou strukturu systému a zajistí tak vhodné prostředí pro běh systému. Aktualizační komponenta zahrnuje akce, které není potřeba provádět pravidelně, ale pouze jednou za čas – pro aktualizaci jednotlivých metod – přičemž je následně potřeba manuální zásah. Příkladem takové akce je vyhledání všech infoboxů na Wikipedii a jejich seřazení podle počtu. Poté je zapotřebí, aby byl výsledek ručně prozkoumán a aby byly vybrány významné infoboxy k jednotlivým entitním typům. Ty se totiž mohou měnit, např. se dva infoboxy sjednotí v jeden, který bude použit ve všech článcích, kde původně byl jeden ze dvou dosavadních. Podobně je zapotřebí zkontrolovat výskyt a vyplnění polí infoboxů.

[23] http://www.mediawiki.org/wiki/Langlinks_table



Obrázek 4.5: Schéma výsledného systému

5 Stručný popis implementace

V této kapitole bude popsána implementace navrženého nástroje. Nastíníme základní strukturu programové části práce a charakterizujeme její jednotlivé součásti. Představíme ucelený popis implementace a u vybraných součástí zmíníme významné implementační detaily. Zaměříme se také na to, jak byly jednotlivé metody aplikovány v rámci extrakce konkrétních entitních typů – tedy jaké byly zvoleny kategorie, infoboxy či regulární výrazy pro extrakci z textu článků.

Většina funkcí pro zpracování zdrojových souborů Wikipedie byla napsána v jazyce Python[24]. Tento interpretovaný skriptovací jazyk je v dnešní době jedním z nejvyužívanějších jazyků pro práci s textem, ale i pro mnoho dalších činností. V našem nástroji bude použita novější verze tohoto jazyka: Python 3. [10] [11] [12] [13] [14]

Implementace vychází z návrhu systému (viz kapitola 4.5) a dělí se na několik hlavních částí, které budou popsány v následujících podkapitolách. Nejprve bude popsána implementace nejdůležitější části aplikace – vlastního systému pro tvorbu znalostní báze pojmenovaných entit. V poslední podkapitole zmíníme další komponenty systému – instalační a aktualizační.

5.1 Systém pro tvorbu znalostní báze entit

Toto jádro programové části práce zajišťuje vlastní extrakci pojmenovaných entit z aktuálních zdrojových souborů Wikipedie a je tvořené třemi součástmi: řídícím skriptem *run.sh*, který řídí a spouští další dvě součásti, a to skript *zdroje.sh* (zajišťující stažení a připravení potřebných částí dumpu české Wikipedie) a *extrakce.sh* (vlastní extrakce entit jednotlivými metodami). Skript *extrakce.sh* spouští skripty v Pythonu 3, které ve většině případů reprezentují jednotlivé metody extrakce.

Implementace těchto metod bude popsána dále, popis je členěn do podkapitol právě podle metod. Jejich činnost je v některých případech implementována v jediném souboru (skriptu), v jiných případech rozdělena do více souborů podle logického členění metody (např. extrakce infoboxů a jejich atributů). V následujících podkapitolách bude také popsán postup, jak byly dané metody využity pro jednotlivé entitní typy.

5.2 Extrakce na základě kategorií

Tato metoda extrakce je implementována v souboru *kategorie.py*. Pro extrakci jednotlivých entit existuje vždy zvláštní funkce se jménem ve tvaru `kategorie_entita()`, tedy např. pro extrakci organizací `kategorie_organizace()`, protože postup je pro danou entitu vždy specifický, jednak výběrem kategorií, ale i způsobem jejich využití – do jaké hloubky se v kategoriích zanořovat či z jakých úrovní získávat články.

Před funkcemi zpracovávajícími jednotlivé entitní typy jsou volány tři důležité funkce, které vytvoří potřebné prostředky a ty jsou pak využívány dalšími funkcemi. První je funkce `odkazy_kategorie()`, která s využitím souboru *cswiki-latest-categorylinks.sql* zjistí a vrátí

[24] https://www.python.org/, české stránky http://python.cz/

veškeré odkazy na kategorie ve formátu vhodném pro další zpracování. Druhou funkcí je `vytvor_strom_kategorii()`, která na základě odkazů vrácených předchozí funkcí vybuduje strukturu reprezentující strom kategorií. Implementačně jde o slovník, kde je klíčem název kategorie a hodnotou dvojice množin: množina článků v této kategorii a množina podkategorií této kategorie. S touto strukturou lze velmi rychle a efektivně pracovat a procházet kategorie. Třetí funkcí je `cisla_na_clanky()`, ta pracuje se souborem *cswiki-latest-page.sql* a na základě informací z tohoto souboru vytvoří slovník „číslo:název" pro všechny články české Wikipedie, kde „číslo" je ID daného článku a „název" titul tohoto článku. Tento překladový slovník je zapotřebí, neboť v odkazech na kategorie se vyskytují pouze ID článků či kategorií, přičemž do znalostní báze chceme znát jejich názvy (tituly).

V rámci funkcí pro zpracování jednotlivých entit jsou ještě využívány především tři důležité funkce pro zjištění obsahu kategorií. Funkce `co_ma_kategorii(kategorie)` vrátí dva seznamy, které reprezentují obsah zadané kategorie, v prvním jsou ID článků v dané kategorii a ve druhém ID podkategorií zadané kategorie. Druhou funkcí je `vsechny_clanky()`, která prozkoumá kategorii do zadané maximální hloubky zanoření a vrátí všechny články, které jsou v ní a v jejích podkategoriích (do zadané maximální hloubky) obsaženy. Poslední funkce je `vse_z_max_hloubky()` a pracuje obdobně, ale vrací všechny články z neomezené hloubky zanoření.

Kromě výše uvedené funkčnosti obsahuje *kategorie.py* také prostředky pro zjištění celkového obsahu české Wikipedie, což zajišťuje funkce `obsah_kategorii()`. Ta zjišťuje obsah základních kategorií Wikipedie.

V následujících podkapitolách bude popsáno, jak byla metoda kategorií implementována pro jednotlivé entitní typy, tedy jaké kategorie byly pro extrakci využity a jakým způsobem (do jaké hloubky zanoření apod.).

5.2.1 Kategorie typu osoba

Na Wikipedii existují dvě početné kategorie, které by měly teoreticky obsahovat všechny entity typu osoba: jsou to kategorie Muži[25] a Ženy[26]. Tato úvaha však vychází z předpokladu, že by všichni autoři byli důslední a každý článek o nějaké osobě by byl do těchto kategorií zařazen. Kromě těchto kategorií byly tedy vybrány některé další pro kontrolu, případně pro doplnění extrahovaného seznamu osob. V rámci těchto kategorií se není třeba zanořovat, nejsou totiž členěny na podkategorie, ale obsahují přímo články o osobách.

Další použitou kategorií, ve které opět nejsou žádné podkategorie, ale pouze články, je kategorie Žijící lidé[27], která by podle názvu měla obsahovat všechny žijící osoby. Naopak zesnulé osoby by měly být zařazeny do kategorií podle data úmrtí, a to jak podle roku[28], tak podle dne v měsíci[29]. Např. osoba, která zemřela 1. ledna 1942, by měla být obsažena v kategoriích „Úmrtí 1. ledna" a „Úmrtí 1942". Každá osoba (ať už žijící, či zesnulá) by pak měla být zařazena v kategoriích podle data (opět rozdělené na kategorie podle roku a podle dne v měsíci) a místa narození – všechny

[25] http://cs.wikipedia.org/wiki/Kategorie:Mu%C5%BEi

[26] http://cs.wikipedia.org/wiki/Kategorie:%C5%BDeny

[27] http://cs.wikipedia.org/wiki/Kategorie:%C5%BDij%C3%ADc%C3%AD_lid%C3%A9

[28] http://cs.wikipedia.org/wiki/Kategorie:%C3%9Amrt%C3%AD_podle_let

[29] http://cs.wikipedia.org/wiki/Kategorie:%C3%9Amrt%C3%AD_podle_dne_a_m%C4%9Bs%C3%ADce

tyto podkategorie jsou obsaženy v kategorii Narození[30]. Oba tyto podstromy (Úmrtí a Narození) lze prohledávat do maximální možné hloubky, nevyskytují se v nich jiné články než ty, které hledáme (tedy příslušící k entitnímu typu osoba).

5.2.2 Kategorie typu místo

V rámci entitního typu místo byly vybrány dva jeho podtypy, a to „stát“ a „sídlo“ (tedy obce, města, vesnice apod.). Pro typ stát byla identifikována vhodná kategorie Státy podle kontinentů[31], resp. její jednotlivé podkategorie. V těchto kategoriích jsou již přímo články odpovídající státům, ale také podkategorie jednotlivých států – v nich články nesmíme vyhledávat, budeme tedy akceptovat pouze články z úrovně zanoření 2 (v kategorii Státy podle kontinentů).

Pro druhý podtyp, tedy sídla, bylo využito více stromů kategorií a v každém byla zvolena různá metodika zanoření, vždy podle toho, na které úrovni se nacházely články typu sídlo a na které už naopak články „nižší úrovně“, tedy např. články o důležitých/známých místech v určitém sídle. Pro města byl použit strom Města podle zemí, pro vesnice obdobně Vesnice podle zemí apod.

Zvláštní pozornost vždy vyžadují kategorie, které se týkají českých sídel (např. podkategorie Vesnice v Česku). Česká republika má na české Wikipedii pochopitelně významnější postavení než ostatní země, což například znamená, že v české Wikipedii nalezneme mnohem více sídel z České republiky než z jiných zemí. Z tohoto důvodu jsou příslušné kategorie členěny více než ostatní – obsahují např. podkategorie podle krajů (např. Vesnice v Olomouckém kraji) a v nich podle okresů (např. Vesnice okresu Prostějov). V kategorii Vesnice v Česku je tedy zapotřebí zanořit se do větší hloubky a obdobně je tomu tak v dalších kategoriích (Obce podle zemí a v rámci ní podkategorie Obce v Česku).

5.2.3 Kategorie typu událost

V základní kategorii Wikipedie Události[32] byly vybrány dvě významné kategorie (resp. stromy podkategorií), které by měly všechny články typu událost obsahovat, dokonce by se měly výrazně překrývat (většina článků bude obsažena v obou). Jedná se o kategorie Události podle roků a Události podle století.

Dále pak bylo ručně procházeno mnoho kategorií a článků typu událost. Členění je velmi složité, je zapotřebí uvažovat více stromů, nicméně vždy pouze do určité hloubky, aby nebyly zahrnuty jiné entity (např. lidé, kteří figurovali v rámci určité události). Po důkladné analýze bylo vybráno velké množství dalších stromů kategorií, které by měly dva základní doplnit. Některé tyto stromy (včetně dvojice základních) jsou uvedeny v tabulce (viz Tabulka 5.1), v pravém sloupci tabulky je pak číselně uvedená maximální povolená hloubka zanoření při hledání v daném stromu kategorií, nebo výraz „neomezeně“ v případě prohledávání do maximální možné hloubky.

Název kategorie (stromu)	Zvažovaná hloubka zanoření
Události podle století	3

[30] http://cs.wikipedia.org/wiki/Kategorie:Narozen%C3%AD

[31] http://cs.wikipedia.org/wiki/Kategorie:St%C3%A1ty_podle_kontinent%C5%AF

[32] http://cs.wikipedia.org/wiki/Kategorie:Ud%C3%A1losti

Události podle roků	3
Války podle století	3
Války podle typu	2
Války podle zemí	2
Bitvy	5
Volby podle let	neomezeně

Tabulka 5.1: Vybrané stromy kategorií typu událost

5.2.4 Kategorie typu organizace

Organizace jsou zařazeny ve stromu kategorií Organizace[33], a to především ve dvou rozsáhlých podstromech: Organizace podle zemí a Organizace podle typu. U těchto stromů byly zjištěny dvě základní vlastnosti, které ovlivní způsob implementace metody kategorií u tohoto entitního typu. První je to, že obsahují jen velmi málo nepatřičných článků, tedy článků, které nepojednávají o určité organizaci. Druhou důležitou vlastností je jejich hloubka – ačkoliv nejsou příliš rozsáhlé počtem článků, jsou poměrně rozsáhlé počtem (pod)kategorií a stupněm zanoření těchto (pod)kategorií.

První vlastnost by umožňovala použití obecné funkce, která vyhledává všechny články z dané kategorie (a všech jejích podkategorií), tedy bez specifického výběru hloubky zanoření. Kvůli druhé vlastnosti to však nebude možné – na určitých úrovních se již mohou nepatřičné články objevit. Z tohoto důvodu je extrakce kategoriemi u tohoto typu poměrně složitá a u mnoha různých kategorií je určena maximální hloubka či to, které její podkategorie budeme brát v potaz.

5.3 Extrakce s využitím infoboxů

Tato metoda je implementována dvěma soubory: *infoboxy.py* a *hodnoty_atributu.py*.

Skript *infoboxy.py* provádí dvě činnosti. První je vyhledání entit (tedy názvů článků) podle infoboxů a druhá je extrakce celých infoboxů ze zdrojových textů Wikipedie pro další zpracování – pro získání atributů z polí infoboxů. Extrakce celých infoboxů je prováděna funkcí `extrakce_celych_infoboxu()`, které se jako parametr zadá typ extrahované entity a název souboru, ve kterém je seznam infoboxů příslušících danému entitnímu typu. Tato funkce následně volá funkci `cele_infoboxy()`, která ve zdrojovém souboru Wikipedie vyhledá instance všech určených infoboxů a vrátí je v seznamu. Celé infoboxy jsou pak uloženy v souborech v adresáři *entita/infoboxy/*, tedy např. *organizace/infoboxy/*.

Druhou činností skriptu *infoboxy.py* je vyhledání názvů entit s využitím infoboxů. Funkce `nazvy_clanku_s_infoboxy()` opět přebírá jako parametry typ extrahované entity a soubor se seznamem infoboxů, následně volá funkci `najdi_infoboxy()`, která najde výskyty určených

[33] http://cs.wikipedia.org/wiki/Kategorie:Organizace

druhů infoboxů ve zdrojovém textu Wikipedie a vrátí seznam názvů článků, které instanci některého z infoboxů obsahují.

Soubor *hodnoty_atributu.py* má na starosti extrakci atributů z infoboxů. Využívá souborů s celými infoboxy extrahované skriptem *infoboxy.py* a extrahuje z těchto infoboxů různé atributy podle typu entity. Klíčová je funkce `hodnoty_atributu()`, která přebírá typ entity (odpovídající adresáři, ve kterém bude funkce pracovat), podtyp (např. u typu místo je to sídlo nebo stát), seznam infoboxů ke zpracování a seznam atributů (resp. názvů polí infoboxu), které byly pro extrakci vybrány. V posledním případě se jedná vždy o n-tici několika n-tic – atribut totiž může být v jednotlivých infoboxech pojmenován různě, např. „rozloha“ a „výměra“. Pro každý řádek souboru s infoboxy (tedy pro každý infobox) je volána funkce `zpracuj_infobox()`, která řádek analyzuje a vytvoří a vrátí slovník, kde je klíčem název pole (atributu) infoboxu a hodnotou hodnota tohoto pole. Funkce `hodnoty_atributu()` poté vybere ze slovníku atributy ze zadaného seznamu a ty uloží do souborů v adresáři *entita/*.

5.3.1 Infoboxy typu osoba

Konkrétní infoboxy pro extrakci entit typu osoba (obdobně i pro další typy) byly vybrány na základě výsledků skriptu *aktulalizace_infoboxu.py*, který je součástí aktualizační komponenty (viz kapitola 5.7). Tento skript vyhledá veškeré infoboxy na Wikipedii, ty jsou pak v souboru *infoboxy_sort.txt* seřazené podle výskytu (počtu jejich instancí) a jsou k dispozici pro ruční výběr a přiřazení k daným entitám.

Pro typ osoba bylo nalezeno a vybráno celkem 32 typů infoboxů. Nejvýznamnější z nich (s počtem instancí více než 1000) jsou uvedeny v tabulce níže (Tabulka 5.2). Na základě vybraných infoboxů bude provedena extrakce názvů článků (tedy entit typu osoba) i celých infoboxů pro další zpracování (pro extrakci atributů).

Název infoboxu	**Výskyt (počet instancí)**
Infobox Politik	8 945
Infobox - fotbalista	4 384
Infobox - osoba	3 947
Infobox - herec	2 607
Infobox - hokejista	1 196
Infobox panovník	1 001

Tabulka 5.2: Významné infoboxy typu osoba

Následně byla také provedena analýza vyplněnosti polí těchto infoboxů a na základě výsledků této analýzy byly vybrány atributy (resp. odpovídající pole infoboxů) pro extrakci. Je to především jméno dané osoby (ve většině infoboxů pole „jméno“, popř. „jméno hráče“), případně příjmení (pokud není zahrnuto v atributu „jméno“), a datum a místo narození a úmrtí (datum je někdy členěno do polí „datum“ a „rok“ a místo podobně do polí „město“ a „stát“). Je zapotřebí také počítat s různými názvy stejných polí – např. „datum narození“ a „datum_narození“. Krom těchto základních údajů bylo vybráno několik dalších, které doplní popis osoby, např. pole „znám jako“ či „povolání“,

která jsou v případě absence nahrazena názvem daného infoboxu, který ve většině případů specifikuje povolání dané osoby – tedy např. „fotbalista“ či „politik“.

5.3.2 Infoboxy typu místo

Opět bylo ze seznamu všech infoboxů na Wikipedii, který byl vytvořen aktualizační komponentou, vybráno několik infoboxů, které se (podle názvu) týkají entity typu místo a které mají významný počet výskytu (více než 100 instancí). V tabulce níže (viz Tabulka 5.3) jsou vypsány infoboxy vybrané podle těchto dvou kritérií a počet jejich instancí na Wikipedii.

Název infoboxu	Výskyt (počet instancí)
Infobox - sídlo světa	10 325
Infobox - sídlo	8 946
Infobox - česká obec	6 196
Infobox zaniklý stát	598
Infobox stát	305

Tabulka 5.3: Vybrané infoboxy typu místo

I u míst byla provedena analýza polí infoboxů a na základě jejich významnosti (vyplněnosti) byly vybrány ty, které budou extrahovány a které budou reprezentovat atributy tohoto entitního typu. Především bylo rozhodnuto, že bude typ místo (stejně jako u kategorií) rozdělen na dva podtypy, a to stát a sídlo, protože u těchto dvou typů entit jsou základní atributy rozdílné. U států budeme např. chtít znát hlavní město, u sídel naopak stát, do kterého sídlo patří. V případě českých obcí (především Infobox – česká obec) bude chybějící pole „stát“ nahrazeno polem „země“ (tedy Čechy, Morava, Slezsko). U obou podtypů byl mezi základní atributy zařazen počet obyvatel (pole „počet obyvatel“ nebo „obyvatel“), rozloha v km^2 (pole „rozloha“ či „výměra“) a základní atribut označený jako charakter nebo typ místa – pole „charakter“, „status“, případně nahrazeno názvem infoboxu (stát, sídlo světa apod.). U států pak budeme chtít znát státní zřízení, datum vzniku (a případně zániku), měnu a jazyk (popř. jazyky). U sídel zeměpisnou šířku a délku, nadmořskou výšku a představitele (starosta, starostka).

5.3.3 Infoboxy typu událost

I zde byl aplikován stejný postup jako u ostatních entit. V následující tabulce (Tabulka 5.4) jsou vypsány názvy vybraných infoboxů typu událost spolu s počtem jejich instancí na Wikipedii.

Název infoboxu	Výskyt (počet instancí)
Infobox - válka	1 009
Infobox Ročník fotbalového turnaje	449

Infobox Tenisový turnaj (konkrétní událost)	377
Infobox Ročník ligy ledního hokeje	286
Infobox - fotbalová sezóna	234
Infobox Tenisový turnaj ATP a WTA	160
Infobox Tenisový turnaj - ženy	72
Infobox Tenisový turnaj - muži	71

Tabulka 5.4: Infoboxy typu událost

Po provedené analýze infoboxů (resp. jejich polí) bylo stejně jako u míst rozhodnuto, že bude tento typ rozdělen na dva podtypy – na události typu válečný konflikt (jediný typ infoboxu s názvem Infobox – válka) a události typu sportovní turnaj (ostatní infoboxy). U válečných střetnutí budeme především chtít znát, do jakého většího celku patří, tedy např. u bitev (které tvoří většinu článků s tímto infoboxem) je v poli „konflikt“ uvedeno, v rámci kterého konfliktu (během které války) k bitvě došlo. V případě války, kdy tento atribut nebude uveden, bude nahrazen názvem infoboxu, tedy slovem „válka“. Dále mezi významné atributy u tohoto podtypu patří místo a čas střetnutí (pole „místo“ a „trvání“), které strany se boje zúčastnily (pole „strana1“ a „strana2“) a jak toto střetnutí dopadlo („výsledek“). U podtypu události – turnaje budeme chtít znát také místní (pole „místo“ či „země“) a časové určení (pole „čas“, „sezóna“, „ročník“ či „rok“) a především, o jaký turnaj/sport se jedná, což může být obsaženo v polích „soutěž“, „liga“ nebo „název“ či je případně nahrazeno názvem infoboxu (např. Tenisový turnaj ATP a WTA).

5.3.4 Infoboxy typu organizace

Opět byl zvolen stejný postup jako u předchozích entitních typů. Názvy vybraných infoboxů typu organizace (s počtem výskytů více než 100) a počty jejich instancí na Wikipedii jsou vypsány v níže uvedené tabulce (Tabulka 5.5).

Název infoboxu	**Výskyt (počet instancí)**
Infobox - farnost	1084
Infobox Fotbalový klub	960
Infobox - firma	923
Infobox politická strana	428
Infobox - střední škola	231
Infobox - českobratrská farnost	200
Infobox univerzita	175
Infobox Organizace	134

Tabulka 5.5: Vybrané infoboxy typu organizace

Následně byla provedena analýza těchto infoboxů (resp. vyplněnosti jejich polí) a stanoveny názvy polí, které budou reprezentovat atributy typu organizace a které budou extrahovány a uloženy do znalostní báze. V první řadě je to název organizace (velká rozmanitost v pojmenování odpovídajícího pole – např. „název“, „Název“, „jméno“, „název_strany“, „název fakulty“ apod.) a typ organizace (pole „druh organizace“, ve většině případů však chybí a je nahrazeno názvem infoboxu, tedy např. „farnost“, „Fotbalový klub“ či „politická strana“). Mezi další vybrané atributy patří např. sídlo organizace, a to město („město“, „sídlo město“, „sídlo“) a stát („stát“, „sídlo stát“), časové určení založení (velké množství názvů polí, např. „vznik“, „založení“, „Založení“ „založen“, „založena“, „rok založení“ či „zahájení“) a představitel organizace (např. „předseda“, „majitel“, „lídr_jméno“, „farář“, „rektor“, „děkan“, „ředitel“ apod.).

5.4 Extrakce z textu článků

Implementace této metody je opět rozdělena na dvě části: první je skript na extrakci entit (tedy pouze názvů článků) na základě textu článků (soubor *plain_text.py)* a druhou je skript pro extrakci informací (tedy atributů) z prostého textu článků (soubor *udaje_z_textu.py).*

Soubor *plain_text.py* obsahuje sestavené regulární výrazy pro úvodní věty jednotlivých entitních typů a dále funkci `plain_text()`, která provede extrakci na základě zadaného názvu entitního typu a daného regulárního výrazu. Funkce prochází zdrojový text Wikipedie a snaží se identifikovat libovolnou úvodní větu, na ni poté aplikovat daný regulární výraz a v případě úspěchu přidá titul článku do seznamu nalezených entit.

Skript *udaje_z_textu.py* také obsahuje množství regulárních výrazů, jsou však o něco složitější. Jejich funkcí je získat z úvodních vět článků několik základních údajů o entitě. Před extrakcí je tedy zapotřebí určit seznam entit (např. seznam osob získaných všemi metodami), posléze jsou tyto články ve zdrojovém souboru Wikipedie vyhledány a je opět co nejpřesněji určena úvodní věta článku. Na ni je pak postupně aplikováno několik regulárních výrazů, aby v případě neúspěchu nejpodrobnějšího výrazu mohlo být získáno alespoň menší množství údajů některým z méně podrobných.

5.4.1 Text článků typu osoba

Syntax úvodních vět osob bývá poměrně charakteristická a je popsána v kapitole 4.3. Při sestavování vhodného regulárního výrazu, který by této syntaxi odpovídal, bylo zapotřebí najít kompromis mezi výrazem příliš obecným (který by zahrnul i články jiných entitních typů, např. událostí, které v závorce mohou mít také letopočty jako osoby) a příliš podrobným (který by nalezl jen malé množství článků). Zvolen byl nakonec tento:

```
'''.+?'''[^.]{0,30}\(\*.+?\)
```

Skládá se ze tří hlavních částí: z názvu článku (resp. jména osoby), závorky a textem mezi nimi. Název článku je obklopený trojicemi jednoduchých apostrofů, které ve zdrojovém textu Wikipedie značí zvýrazněný (tučný) text. V závorce je povinný znak „*“ a za ním další znaky (např. data narození/úmrtí, místa apod.). Mezi názvem článku a závorkou mohou být libovolné znaky, ale

s výjimkou tečky (aby nekončila věta) a celkově jich může být maximálně 30 (aby závorka následovala v přijatelné vzdálenosti).

Entitní typ osoba je také jediný, u kterého se podařilo s úspěchem implementovat metodu pro získání atributů z textu. K tomu je použito několik složitějších regulárních výrazů, které vycházejí z výrazu výše zmíněného. První (tedy nejvíce obsáhlý) počítá se všemi informacemi, tedy že jsou v závorce údaje o narození i úmrtí, a to data i místa. V dalších výrazech hledané údaje postupně ubývají, v posledním je hledáno pouze datum narození. Pokud některý z výrazů uspěje, informace jsou extrahovány a další výrazy již aplikovány nejsou.

5.4.2 Text článků typu místo

Stěžejním problémem bylo v tomto případě stejně jako u osob zvolit takový regulární výraz, který by odpovídal co nejvíce úvodním větám článků o místech, ale který by současně nezahrnul žádné jiné články. Nejprve bylo ručně procházeno velké množství článků a byly hledány opakující se vzory. Výsledný výraz pro extrakci entit typu místo pak vypadá takto:

```
'''.+?'''[^.]*?\s(je|jsou)\s[^.]{0,30}(stát|republika|město|městem|městys|vesnice|ves|obec)(\s|,|\.)
```

Na začátku je očekáván (stejně jako u osob) název daného článku (v tomto případě místa) uzavřený mezi trojice apostrofů, což v syntaxi Wikipedie značí zvýrazněný text. Následuje libovolné množství znaků kromě tečky – chceme totiž výraz v rámci jediné věty. Dále výraz obsahuje mezerami oddělené sloveso „je“ nebo „jsou“ (kvůli pomnožným tvarům názvů některých míst, např. vesnic). Následuje sekvence znaků kromě tečky kvůli konci věty, navíc však omezujeme maximální počet znaků, aby klíčový řetězec označující typ místa (stát, město apod.) následoval v přijatelné vzdálenosti a nevztahoval se k jiné informaci (aby nevyhovoval např. řetězec „XY je známý hudebník, se kterým je úzce spojeno město YZ“). Poslední důležitou částí výrazu je určité specifické slovo, které označuje konkrétní typ místa, následované mezerou, čárkou nebo tečkou.

U tohoto entitního typu (stejně jako u všech ostatních s výjimkou osob, viz předchozí podkapitola) se nepodařilo použít této metody k získání atributů. Základní údaje o místu v úvodní větě nejsou ve většině případů obsaženy vůbec, a pokud některé ano, není to v rámci určité pravidelné konstrukce.

5.4.3 Text článků typu událost

I u entity typu událost byly nejprve ručně procházeny články o událostech a bylo zkoumáno, zda obsahují nějaký společný vzor či skupinu vzorů. Bohužel však bylo vyhodnoceno, že tomu tak není, články jsou uvedeny velkým množstvím různých způsobů, i když se jedná o stejný typ události. Pokud by byl přece jen nějaký vzor sestaven, byl by velmi složitý (a bylo by tedy výpočetně náročné zkoušet ho na úvodní části všech článků), a přesto by zahrnul jen malé procento článků, které se týkají událostí. Z těchto důvodů tato metoda nebyla v rámci entit typu událost aplikována.

5.4.4 Text článků typu organizace

Metoda extrakce na základě textu článku se ukázala jako obtížně použitelná i u entitního typu organizace. Nejprve byly ručně procházeny články o organizacích a byla zkoumána syntax jejich úvodních vět. Bylo zjištěno, že je velmi různorodá, samotných výrazů pro označení tohoto entitního typu je velmi mnoho, např. organizace, společnost, firma, sdružení, ale i konkrétnější formy – fotbalový klub, politická strana apod. Bylo tedy obtížné sestavit takový regulární výraz, který by nebyl příliš rozsáhlý (a tedy náročný na zpracování a vyhledávání), ale který by současně zahrnul větší množství článků. Jako konečná verze byl nakonec vybrán tento výraz:

```
'''.+?'''[^.]*?\s(je|jsou)\s[^.]{0,40}(organizace|politická strana|s
družení|společnost|firma|klub)(\s|,|\.|\])
```

Na začátku je očekáván (stejně jako u většiny článků na Wikipedii) název dané entity (v tomto případě název organizace) uzavřený mezi trojice apostrofů (zvýrazněný text). Následuje libovolné množství znaků kromě tečky – chceme výraz v rámci jedné věty. Dále výraz obsahuje mezerami oddělené sloveso „je“. Poté následuje opět sekvence znaků kromě tečky kvůli konci věty, i zde omezujeme maximální počet znaků, aby klíčový řetězec označující typ organizace (firma, společnost, klub apod.) následoval v přijatelné vzdálenosti a nevztahoval se k jiné informaci. Poslední důležitou částí výrazu je určité specifické slovo, které označuje organizaci, následované mezerou, čárkou, tečkou či případně uzavírací hranatou závorkou (pokud klíčové slovo nebo slovní spojení odkazuje na jinou stránku na Wikipedii, např. [[mezinárodní organizace]]).

Opět se nepodařilo použít metody extrakce z textu k získání atributů. Základní údaje o organizaci v úvodní větě ve většině případů vůbec nenajdeme, popř. je nelze plošně identifikovat a extrahovat.

5.5 Překlad z anglické databáze

Tuto doplňující metodu implementuje skript *preklad_z_aj.py*. Důležitou funkcí pro překlad je `zjisti_preklady()`, která ze souboru *cswiki-latest-langlinks.sql*[34] vyjme a připraví pro použití překladové informace, vytvořený slovník je pak používán dalšími funkcemi. Další potřebnou funkcí je `cisla_na_clanky()`, která vytvoří slovník pro překlad ID článků na jejich názvy (podobně jako u kategorií, viz kapitola 5.2).

Ještě před tímto zpracováním jednotlivých entit jsou v rámci tohoto skriptu volány funkce na aktualizaci anglické báze ve tvaru `angl_entita()`, tedy např. `angl_mista()`. V adresáři systému by měl v té době existovat soubor *KB.all* s kompletní bází z projektu DECIHER, který byl stažen v rámci skriptu *zdroje.sh*. Z něj jsou těmito funkcemi získány jednotlivé entity v anglickém jazyce a umístěny do adresáře *preklad*.

Stěžejní funkcí je `dopln_z_aj()`, která přijímá jako parametr typ entity (např. osoby, místa) a která načte výsledky anglické extrakce ze souboru *typ_entity.txt*, jehož umístění předpokládá v adresáři *preklad*. Pokud příslušný soubor v adresáři neexistuje, nebude u tohoto entitního typu metoda aplikována. Funkce následně získá články ze zadaného souboru (tedy anglické entity) a u

[34] Viz kapitola 4.4

všech rozhodne, zda existuje český ekvivalent. Pokud ano, články přeloží a přidá do stávající báze entit – do souboru *preklad_z_aj.txt* v adresáři právě zpracovávané entity.

5.6 Sestavení znalostní báze

Tento závěrečný krok je implementován v souboru *sestaveni_baze.py*. Jeho ústřední funkce `sestaveni_baze()` vyžaduje jako jediný parametr název pojmenované entity a je postupně volána pro vybrané entity (osoby, místa, události a organizace). Funkce najde a shromáždí výsledky získané jednotlivými metodami u dané entity, výsledky postupně sloučí, tiskne statistiky na standardní výstup a finální množiny entit umístí do adresáře *baze/*. Znalostní báze bude v tomto adresáři rozdělena do jednotlivých souborů podle entitního typu.

Dále skript obsahuje a spouští funkci `sestaveni_atributu()`, která shromáždí atributy získané ke všem entitním typům a také je umístí do adresáře *baze/* do souborů podle jednotlivých entitních typů (popř. podtypů, které mají rozdílné atributy – u míst a událostí, viz kapitoly 5.3.2 a 5.3.3). U osob budou navíc sloučeny atributy získané z obou metod (atributy z infoboxů a atributy získané z prostého textu článků).

5.7 Další součásti systému

Kromě jádra systému, které se stará o vlastní extrakci pojmenovaných entit a tvorbu znalostní báze, byly implementovány dvě další důležité komponenty vycházející z návrhu systému.

První z nich je tzv. instalátor, tedy instalační nástroj, který připraví prostředí pro běh systému. Jedná se o skript *instalator.sh*, který za pomoci standardních unixových příkazů ověří existenci potřebné adresářové struktury, a pokud neexistuje (či je neúplná), vytvoří ji. Rozmístí také základní podobu některých souborů, které jsou pro automatickou extrakci zapotřebí – např. základní seznamy infoboxů přiřazených k jednotlivým entitním typům.

Dalším prvkem systému je aktualizační komponenta, která bude spouštěna jednou za delší časový úsek a s pomocí ručního zásahu bude aktualizovat některé výchozí údaje k extrakci, jako je výběr infoboxů či jejich polí. Komponenta je implementována skriptem *aktualizator.sh*, který spouští především Python skript *aktulalizace_infoboxu.py*. V něm je stěžejní funkce `aktualizuj_infoboxy()`. Ta prochází zdrojový text všech článků Wikipedie a hledá a počítá v nich výskyt veškerých infoboxů libovolného typu. K tomu využívá funkci `zpracuj_radek_infobox()`, která je volána pro řádek obsahující název nějakého infoboxu, funkce tento název určí a započítá do počtu již nalezených, stejně pojmenovaných infoboxů. Výsledky jsou uloženy do souboru *infoboxy.txt*, řádky soboru reprezentují infoboxy s určitým názvem a jsou ve tvaru: `jmeno_infoboxu+'\t'+vyskyt_infoboxu (počet instancí)`. Následně je skriptem *aktualizator.sh* volán příkaz `sort` se specifickými parametry, který infoboxy seřadí podle výskytu od nejvýznamnějších k méně významným a takto seřazené výsledky této analýzy uloží do souboru *infoboxy_sort.txt*, kde jsou k dispozici pro ruční klasifikaci do jednotlivých entitních typů. V této fázi aktualizace je zapotřebí soubor ručně projít a vybrat infoboxy, které podle názvu přísluší k jednomu z určených entitních typů, a umístit seznam názvů těchto infoboxů do souboru *entita/infoboxy_entita.txt*, tedy např. v případě organizací do souboru *organizace/infoboxy_organizace.txt*. V těchto souborech tedy bude vždy seznam významných

(počtem instancí) infoboxů příslušících k danému entitnímu typu připravený pro použití při extrakci pojmenovaných entit metodou infoboxů.

6 Výsledky a vyhodnocení nástroje

V této kapitole budou demonstrovány výsledky, kterých bylo v rámci této práce dosaženo. V podkapitolách, které odpovídají zpracovávaným entitním typům, budou uvedeny počty extrahovaných entit jednotlivými metodami i celková velikost báze po sjednocení výsledků všech metod. V podkapitole 6.5 pak nastíníme celkový obsah Wikipedie a v podkapitole 6.6 shrneme výsledky extrakce a zhodnotíme rozsah záběru nástroje.

6.1 Extrakce entit typu osoba

Základní a nejúspěšnější použitou metodou je (obdobně jako u ostatních entitních typů) extrakce metodou kategorií. U entitního typu osoba bylo nejprve využito pouze kategorií „Muži“ a „Ženy“, které by měly teoreticky zahrnout všechny hledané osoby. Z kategorie „Muži“ bylo získáno **58 597** osob, z kategorie „Ženy“ **10 753** osob. Prostý součet těchto čísel by se rovnal číslu **69 350**, nicméně velikost báze po sjednocení těchto dvou kategorií je rovna pouze **69 349**, z čehož je zjevné, že článek o jedné osobě byl zařazen do obou těchto kategorií. Na první pohled jde o chybu, nicméně po identifikaci a ručním prozkoumání článku byl zjištěno, že zařazení do obou kategorií může být korektní – jedná se o Candy Darling[35], která je tzv. transwoman (operativně si přeměnila pohlaví z mužského na ženské).

Následně byly získány osoby z několika dalších stromů kategorií (viz kapitola 5.2.1). Z těchto kategorií bylo získáno **213** dalších článků, které z nějakého důvodu (většinou nedůsledností autora nebo se jedná např. o další transsexuální osoby) nebyly zařazeny do jedné ze základních kategorií osob rozlišujících pohlaví. Po sjednocení obsahuje báze osob získaných metodou kategorií **69 595** článků.

Extrakcí s využitím infoboxů (konkrétní využité infoboxy jsou zmíněny v kapitole 5.3.1) bylo nalezeno **29 771** článků, z nichž celkem **361** nebylo v bázi vytvořené na základě kategorií. V kategoriích nejsou tyto články zařazeny většinou kvůli nedůslednosti autora, popř. se autor rozhodl článek do dané kategorie nezařadit, např. pokud se jednalo o fiktivní osobu. Díky této metodě tedy vzrostla velikost báze osob na **69 956**.

Třetí aplikovanou metodou byla extrakce z prostého textu článků. Za použití regulárního výrazu navrženého v kapitole 5.4.1 bylo získáno **19 567** osob (resp. článků o osobách). Většina z nich již byla nalezena předchozími metodami, bylo však získáno celkem **72** nových osob, které předchozími způsoby nalezeny nebyly (autor článek nezařadil do vhodné kategorie a ani mu nepřiřadil odpovídající infobox) a které zvětšily obsah báze na **70 028** osob.

Poslední použitou metodou je překlad z báze osob, která byla vytvořena z anglické verze Wikipedie v rámci projektu DECIPHER[36]. V současné znalostní bázi tohoto projektu bylo nalezeno **1 093 021** osob, většina z nich pochopitelně v mnohem menší české verzi Wikipedie nemá ekvivalent, ten se podařilo nalézt u **41 899** osob, které tak byly přeloženy do češtiny a porovnány s naší bází. Bylo zjištěno, že z tohoto množství celkem **462** osob v naší bázi chybí, ve většině případů jde o různé starověké či mýtické postavy (např. biblické či historické). Tyto osoby se nepodařilo nalézt žádným z navržených metod extrakce – nejsou důsledně zařazeny do zvažovaných kategorií, nemají autorem

[35] http://cs.wikipedia.org/wiki/Candy_Darling

[36] Informace o této metodě v kapitole 4.4, informace o projektu DECIPHER v kapitole 3.2.

přiřazený infobox a syntax jejich úvodních vět neodpovídá navrženému regulárnímu výrazu. Přeložené entity byly tedy do báze přidány a velikost báze tak vzrostla na **70 490** osob.

Osoby jsou jediným entitním typem, na který se podařilo aplikovat obě metody získání atributů: za pomoci infoboxů i extrakcí z prostého textu článků. Metodou infoboxů byly extrahovány informace ke **29 771** článkům, což tvoří **42,23 %** z celkového množství entit typu osoba. Úspěšnější byla metoda extrakce z prostého textu článku, kterou byly získány základní informace k **54 070** osobám, touto metodou byly tedy získány atributy k **76,71 %** entit v bázi osob. Po sloučení výsledků obou metod jsou v bázi informace k **59 483** osobám, tedy k **84,39 %** osob.

6.2 Extrakce entit typu místo

První použitou metodou k extrakci míst byla stejně jako u osob metoda kategorií. Entitní typ místo uvažovaný v této práci je logicky členěn na dva podtypy – státy a sídla (města, vesnice apod.). V rámci podtypu stát bylo z vybraných kategorií popsaných v kapitole 5.2.2 získáno celkem **268** entit.

U podtypu sídlo byla extrakce prováděna postupně. Z jednotlivých stromů kategorií bylo získáno **6 691** měst, z toho je pouze **596** českých měst a **6 095** měst je z jiných zemí. U vesnic je situace opačná – nalezeno bylo **12 861** vesnic z České republiky a pouze **934** vesnic z ostatních zemí světa. U obcí byly extrahovány zvlášť i obce ležící na Slovensku, těch bylo nalezeno přesně **1 500**, zatímco českých obcí **6 276** a obcí z jiných zemí **5 559**. Průběžně byly odstraňovány duplicity, pokud bylo některé sídlo zahrnuto ve více kategoriích (např. jako vesnice i jako obec). Finální báze složená ze všech typů (města, obce, vesnice) obsahuje **24 975** sídel, která byla získána metodou na základě kategorií. Celkově spolu s podtypem stát bylo extrahováno **25 242** entit typu místo.

Následně byla aplikována metoda infoboxů, byly extrahovány články, které obsahují některý z infoboxů zmíněných v kapitole 5.3.2. Tímto způsobem bylo získáno **26 665** entit typu místo, z toho celých **4 352** článků nebylo získáno metodou kategorií. Velikost čísla ukazuje jednak na význam metody infoboxů u tohoto entitního typu (infoboxy tohoto typu mají na Wikipedii celkově nejvyšší výskyt), ale i na menší úspěšnost u extrakce metodou kategorií, způsobenou složitostí stromu kategorií.

Jako třetí byla stejně jako u osob využita metoda extrakce na základě úvodní věty textu článku, kterou bylo získáno **11 081** článků typu místo. Z tohoto množství bylo nalezeno **503** článků, které nebyly dvěma předchozími metodami nalezeny, tedy nebyly ve zvažovaných kategoriích a neobsahovaly některý z infoboxů. Tyto entity doplnily stávající bázi míst a zvětšily její obsah na **30 097** entit.

Poslední aplikovanou metodou byl překlad z anglické báze míst. Přeložit se podařilo **7 507** entit (tedy článků, které mají ekvivalent v české verzi Wikipedie), z toho však pouze **78** entit nebylo v dosavadní bázi a byly tedy do ní přidány. Celková velikost báze po sjednocení výsledků všech čtyř metod je **30 175** entit typu místo.

Atributy k entitnímu typu místo byly získávány pouze metodou infoboxů. Extrahovány byly informace ke **22 812** entitám typu místo, což tvoří **75,6 %** z celkového množství.

6.3 Extrakce entit typu událost

Extrakce entit typu událost začala opět získáním článků za pomoci kategorií, bylo použito několik postupů v prohledávání různých stromů kategorií zmíněných v kapitole 5.2.3. Ze dvou základních

stromů kategorií, tedy v kategoriích událostí podle datace, bylo získáno **1 547** událostí ze stromu Události podle roků a **1 918** událostí ze stromu Události podle století. Posléze byly získány články z velkého množství dalších stromů týkajících událostí, příkladem může být strom Války podle století, ze kterého bylo získáno **776** událostí typu válka, z kategorií Války podle zemí **235** válek a ze stromu Války podle typu **182**. Dále bylo získáno **1 061** událostí (bitvy, střetnutí apod.) ze stromu Bitvy a **324** volebních událostí ze stromu Volby. Mnoho událostí bylo zahrnuto vícekrát (např. podle typu i datace), po odstranění duplicit má báze událostí získaných z kategorií velikost **13 849**.

Druhou použitou metodou byla metoda s využitím infoboxů. Článků s některým z infoboxů zmíněných v kapitole 5.3.3 bylo na české Wikipedii nalezeno **2 630**, tyto články tedy podle návrhu metody patří pod entitní typ událost. Z tohoto množství nebylo metodou kategorií nalezeno **78**, což není (vzhledem k celkové mocnosti tohoto entitního typu) příliš velké množství a ukazuje na úspěšnost metody kategorií, a to i přes složitost kategorizačního členění u tohoto typu (viz kapitola 5.2.3). Sjednocením výsledků těchto dvou metod dosáhla velikost báze čísla **13 927**.

Metoda extrakce z prostého textu článků nebyla u tohoto entitního typu využita (zdůvodnění viz kapitola 5.4.3).

Jako třetí tak byla aplikována metoda překladu z anglické báze událostí. V ní bylo celkově nalezeno **80 089** entit, z toho se však podařilo pouze **6 212** přeložit do češtiny (byl tedy nalezen ekvivalentní článek na české Wikipedii). Poměrně velké množství (**553** entit) nebylo doposud v české bázi událostí, důvodem z části může být neúplný záběr nástroje, ale také nepřesné překlady, které mění entitní typ článku – např. článek Adnan Hadždž[37] na české Wikipedii pojednává o novináři a fotografovi tohoto jména (jde tedy o entitu typu osoba), ale anglická verze článku s názvem Adnan Hajj photographs controversy[38] se věnuje známé aféře tohoto fotografa a jedná se tedy skutečně o entitu typu událost.

Celkově bylo tedy zmíněnými metodami nalezeno **14 480** entit typu událost.

Následně byly extrahovány atributy událostí, a to metodou infoboxů. Atributy byly zpracovávány zvlášť ke dvěma zmíněným podtypům zmíněným v kapitole 5.3.3 (války a turnaje). U podtypu „války“ byly základní informace extrahovány k **998** entitám a u podtypu „sportovní turnaj“ k **1 619** entitám, celkově tedy ke **2 617** událostem, což tvoří pouze **18,07 %** z celkového množství událostí. Důvodem nízké úspěšnosti je nízký výskyt článků s infoboxy a fakt, že se nepodařilo aplikovat metodu extrakce z prostého textu článků, což se zdařilo pouze u osob.

6.4 Extrakce entit typu organizace

Jako první byla stejně jako u předchozích entitních typů aplikována metoda kategorií. Systémem popsaným v kapitole 5.2.4 byly články získávány ze dvou základních kategorií; ze stromu Organizace podle zemí bylo získáno **8 487** organizací a ze druhého stromu Organizace podle typu **5 463**. Je zřejmé, že se tyto dvě množiny částečně překrývají (mnoho organizací je zařazeno jak podle typu, tak podle země), ale některé organizace byly nalezeny pouze z jednoho stromu (např. mezinárodní organizace nejsou zařazeny podle zemí apod.). Po sjednocení bylo získáno **10 519** organizací extrakcí z kategorií.

Posléze byla i u entitního typu organizace využita metoda infoboxů. Na základě infoboxů týkajících se organizací (konkrétní jsou uvedeny v kapitole 5.3.4) bylo celkově nalezeno **4 418** článků. Z tohoto počtu nebyla téměř polovina (přesně **2 066**) nalezena metodou kategorií, z toho

[37] http://cs.wikipedia.org/wiki/Adnan_Had%C5%BEd%C5%BE

[38] http://en.wikipedia.org/wiki/Adnan_Hajj_photographs_controversy

většinu tvoří farnosti (církevní správní jednotky), které nebyly ve zvažovaných stromech kategorií zařazeny. Celkově tak bylo prvními dvěma metodami získáno **12 585** entit typu organizace.

Metodou extrakce z prostého textu článků bylo identifikováno **2 437** článků a **536** z nich nebylo obsaženo v dosavadní bázi dat. Tyto články byly tedy do báze přidány a zvýšily celkovou velikost báze na **13 121** organizací.

Toto číslo je současně i konečnou velikostí báze organizací, metoda překladu z anglické verze nebyla v době tvorby této práce aplikována, protože nebyla nalezena odpovídající anglická báze organizací.

I u entitního typu organizace byla pro získání atributů využita pouze metoda infoboxů. Základní informace byly extrahovány ze všech infoboxů typu organizace a byly tak získány k **4 079** entitám typu organizace, což tvoří **31,09 %** ze všech nalezených organizací. Větší úspěšnosti opět nebylo dosáhnuto kvůli nízkému výskytu infoboxů a nevyužití metody extrakce z prostého textu článků.

6.5 Obsah české Wikipedie

Abychom mohli zhodnotit úspěšnost nástroje, je třeba znát kromě počtů extrahovaných entit také vztah těchto počtů k celému obsahu Wikipedie. Bylo tedy zapotřebí prozkoumat obsah české Wikipedie. Díky funkcím pro práci se stromem kategorií implementovaným v souboru *kategorie.py* bylo možné prozkoumat obsah základních (tedy nejobsáhlejších) kategorií na Wikipedii. Kořenovou kategorií stromu, který obsahuje všechny obsahové články (tedy články se znalostmi, nikoliv články s metainformacemi, nápovědou apod.), je kategorie Základní kategorie[39]. Tato kategorie by tedy měla zahrnovat veškeré znalostní články české Wikipedie. Obsahuje dvacet dva podkategorií, které reprezentují dvacet dva základních oblastí a tvoří základní kategorizaci článků.

Vytvořeným nástrojem byly tyto kategorie prozkoumány do maximální možné hloubky a bylo zjištěno, kolik článků je v nich (a ve všech jejich podkategoriích) zahrnuto. Počty článků reprezentující rozsah kategorií jsou uvedeny v následující tabulce (Tabulka 6.1). Graf 6.1 pak znázorňuje výsledek graficky.

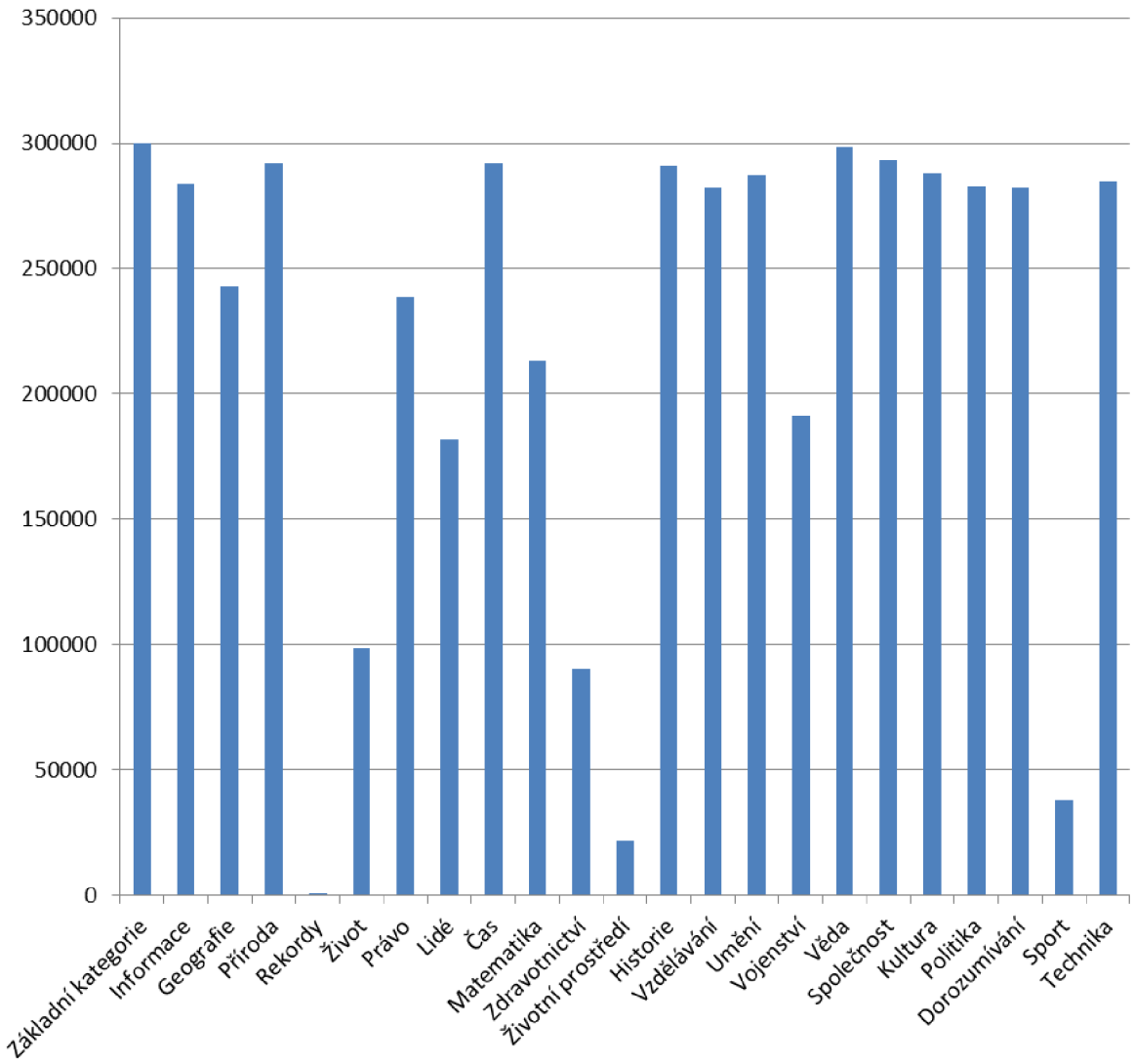
Název kategorie	**Počet článků**
Základní kategorie	300 033
Informace	283 744
Geografie	242 955
Příroda	291 779
Rekordy	176
Život	98 484
Právo	238 604
Lidé	181 973

[39] http://cs.wikipedia.org/wiki/Kategorie:Z%C3%A1kladn%C3%AD_kategorie

Čas	291 814
Matematika	213 404
Zdravotnictví	90 210
Životní prostředí	21 733
Historie	290 967
Vzdělávání	282 570
Umění	287 045
Vojenství	191 566
Věda	298 796
Společnost	293 183
Kultura	288 329
Politika	282 760
Dorozumívání	282 429
Sport	37 652
Technika	284 963

Tabulka 6.1: Základní kategorie Wikipedie a velikost jejich obsahu

Z výsledků je zjevné, že se většina kategorií výrazně překrývá – velmi často obsahují podobně vysoký počet článků blížící se hodnotě obsahu Základní kategorie, a tedy počtu všech článků na Wikipedii. Tento jev je způsoben provázaností jednotlivých oborů a složitostí sítě kategorií – některé články lze nalézt v určité hloubce u velkého množství základních kategorií. Kvůli tomu také nebylo možné použít metodu kategorií slepým prohledáváním do hloubky, ale bylo zapotřebí vybrat pečlivě konkrétní kategorie a hloubku, ve které budou entity hledány.



Graf 6.1: Znázornění velikosti obsahu základních kategorií Wikipedie

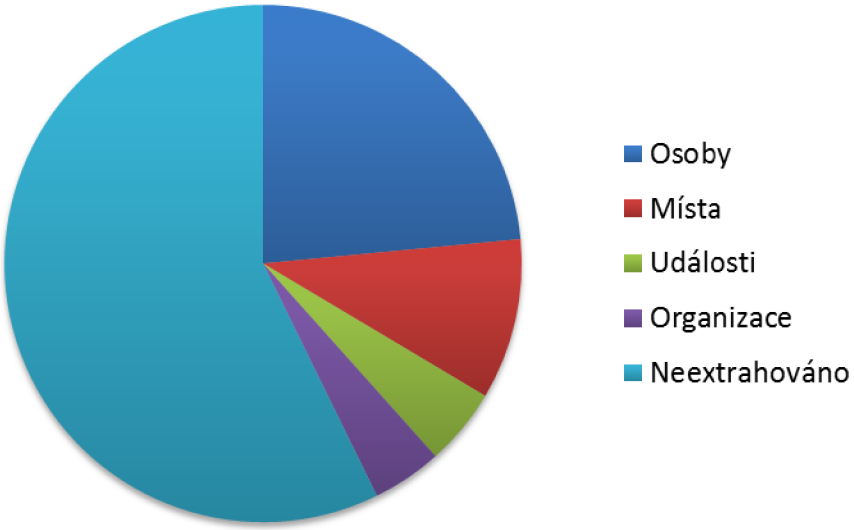
6.6 Shrnutí výsledků extrakce

V rámci práce byly extrahovány čtyři typy pojmenovaných entit, a to osoby, místa (státy a sídla), události a organizace. Využito bylo tří základních metod extrakce, prvním byla metoda kategorií, druhou využití infoboxů a třetí extrakce na základě textu článků. Počty extrahovaných entit jednotlivými metodami jsou podrobněji popsány v předchozích kapitolách, strukturovaně jsou pak uvedeny v tabulce níže (viz Tabulka 6.2). V době poslední extrakce bylo na Wikipedii celkově **300 033** článků, součet extrahovaných entit zvažovaných typů činí **128 266**, což činí cca **42,75 %** článků na české Wikipedii. Zbylých 57,25 % procent tvoří například obecné entity (vysvětlení pojmů), ale také další typy pojmenovaných entit nezvažované v této práci, jako např. produkty/výrobky, umělecká díla, množství geografických entit (řeky, moře, jezera, hory, pohoří, regiony apod.) a jiné. Graficky je podíl extrahovaných entit na celkovém množství článků v české Wikipedii zobrazen v níže uvedeném grafu (Graf 6.2). Články, které nebyly nástrojem zahrnuty a

identifikovány jako entita některého zvažovaného typu, jsou označeny jako „neextrahováno“ a tvoří zmíněných 57,25 %.

Entitní typ	**Zisk jednotlivými metodami extrakce**				**Celkem**
	Kategorie	***Infoboxy***	***Prostý text***	***Překlad***	
Osoby	69 595	29 711	19 567	41 899	**70 490**
Místa	25 242	26 665	11 081	7 507	**30 175**
Události	13 849	2 630	0	6 212	**14 480**
Organizace	10 519	4 418	2 437	0	**13 121**
Celkový počet získaných entit					**128 266**

Tabulka 6.2: Počty extrahovaných entit



Graf 6.2: Podíl extrahovaných entit na celkovém obsahu české Wikipedie

7 Závěr

V předkládané bakalářské práci byla řešena problematika extrakce pojmenovaných entit z české Wikipedie. Tato oblast byla nejprve představena (včetně několika současných projektů, které se jí zabývají) a posléze byly navrženy metody, kterými je možné entity z textů české Wikipedie extrahovat, a to kategoriemi (na základě zařazení článku do kategorií Wikipedie), s pomocí infoboxů (podle infoboxu, který je k článku přiřazen) a na základě prostého textu (využití syntaxe úvodní věty článku). K extrakci byly vybrány čtyři entitní typy: osoby, místa (státy a sídla, tj. města, obce, vesnice apod.), události a organizace. Navržené metody byly následně implementovány, práce tedy dále představila způsob jejich implementace a aplikace těchto metod na jednotlivé entitní typy.

V další části práce byly popsány dosažené výsledky extrakce – celkově bylo extrahováno 128 266 entit (z toho 70 490 osob, 30 175 míst, 14 480 událostí a 13 121 organizací), což z celkového počtu článků na české Wikipedii tvoří zhruba 42,75 %. Implementace všech tří metod lze teoreticky dále zlepšovat a získat tak více entit (např. podrobnější analýzou kategorií či vylepšením regulárních výrazů), ovšem zlepšení nebude pravděpodobně nijak výrazné. Procento extrahovaných pojmenovaných entit lze naopak výrazně zvýšit aplikací navržených metod na další entitní typy, např. produkty/výrobky, umělecká díla či jiné geografické typy (kromě zvažovaných míst, tedy států a sídel).

Znalostní bázi pojmenovaných entit, kterou lze díky vytvořenému nástroji sestavit, je možné využít v dalších projektech v oblasti rozpoznávání pojmenovaných entit. Konkrétním způsobem využití je doplnění jiné báze, např. sestavené na základě anglické verze Wikipedie, či sestrojení samostatného nástroje na rozpoznávání pojmenovaných entit v nestrukturovaném textu.

Literatura

[1] MANNING, Christopher D. a Hinrich SCHÜTZE. *Foundations of statistical natural language processing*. Cambridge: MIT Press, c1999, xxxvii, 680 s. ISBN 02-621-3360-1.

[2] PSUTKA, Josef. *Komunikace s počítačem mluvenou řečí.* Vyd. 1. Praha: Academia, 1996, 287 s. ISBN 80-200-0203-0.

[3] CARDIE, Claire. Empirical Methods in Information Extraction. *AI Magazine*. 1997, vol. 18, pp. 65-79. ISSN 0738-4602. Dostupné také z: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1322/1223

[4] LAŠEK, Ivo. Extrakce z nestrukturovaných dat. In: *Propojená data na webu* [online]. 2013-11-25 [cit. 2014-04-26]. Dostupné z: http://nb.vse.cz/~svatek/rzzw/extrakce_RDF_Lasek.pdf

[5] CHIEU, H.L., NG, H.T. Named entity recognition with a maximum entropy approach. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL* 2003 – Volume 4. pp. 160–163. CONLL '03, Association for Computational Linguistics (2003). Dostupné z: http://www.comp.nus.edu.sg/~nght/pubs/conll03.pdf

[6] ROUSE, Margaret. Knowledge base. In: *SearchCRM* [online]. 2007-03-?? [cit. 2014-04-26]. Dostupné z: http://searchcrm.techtarget.com/definition/knowledge-base

[7] KAZAMA, J., TORISAWA, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* Str. 698–707. Association for Computational Linguistics (June 2007). Dostupné také z: http://acl.ldc.upenn.edu/D/D07/D07-1073.pdf

[8] RIZZO, Giuseppe, ERP, Marieke a TRONCY, Raphael. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. *In: 9th edition of the Language Resources and Evaluation Conference (LREC'14),* Reykjavik, Iceland, 2014. Dostupné také z: http://www.di.unito.it/~rizzo/publications/Rizzo_Erp-LREC2014.pdf

[9] SMRŽ, Pavel, OTRUSINA, Lubomír, KOUŘIL, Jan a DYTRYCH, Jaroslav. DECIPHER Semantic Annotator. In: *Europe´s Information Society Thematic* Portal [online]. 2013-08-31 [cit. 2014-04-26]. Dostupné z: http://ec.europa.eu/information_society/apps/projects/logos/1/270001/080/deliverables/001_DecipherD431SemanticAnnotatorv01.pdf

[10] *Python* [online]. 2001 – 2014 [cit. 2014-05-15]. Python Software Foundation, oficiální stránky komunity. Dostupné z: https://www.python.org/

[11] JAVOREK, Jan. *Python CZ* [online]. [cit. 2014-05-15]. Stránky české komunity. Dostupné z: http://python.cz/

[12] *Python 3.0 Release* [online]. 2008-12-03 [cit. 2014-05-15]. Informace o Python verzi 3. Dostupné z: https://www.python.org/download/releases/3.0

[13] PILGRIM, Mark. *Dive into Python 3* [online]. 2011 [cit. 2014-05-15]. Dostupné z: http://www.diveintopython3.net/, česká verze: http://diveintopython3.py.cz/

[14] *Python 3.4.0 documentation* [online]. 2014-05-11 [cit. 2014-05-15]. Python Software Foundation, 1990 – 2014. Dostupné z: https://docs.python.org/3/

Seznam příloh

Příloha 1. CD s výsledným programem