VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

URČOVÁNÍ TYPŮ A ATRIBUTŮ ENTIT NAPŘÍČ JAZYKY

DETERMINING ENTITY TYPES AND ATTRIBUTES ACROSS LANGUAGES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **DANIEL ŠVUB**
AUTHOR

VEDOUCÍ PRÁCE **doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání bakalářské práce

Student: **Švub Daniel**
Program: Informační technologie
Název: **Určování typů a atributů entit napříč jazyky**
Identifying Entity Types and Attributes Across Languages
Kategorie: Umělá inteligence

Zadání:

1. Seznamte se s metodami a dostupnými nástroji pro určování typů a extrakci dalších atributů entit zmíněných v textu.
2. Zpracujte přehled jazykově závislých technik využívaných při extrakci a nástrojů pro různé úrovně analýzy jazyka v češtině a slovenštině.
3. Na základě získaných poznatků navrhněte a implementujte systém pro automatické určování typů entit zmíněných v článcích české a slovenské Wikipedie.
4. Vyhodnoťte výsledky systému na reprezentativním vzorku dat.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Pro udělení zápočtu za první semestr je požadováno:

- funkční prototyp řešení

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 15. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Cílem této práce je analýza článků na internetové encyklopedii Wikipedii a převod jejich textu psaného v přirozeném jazyce na strukturovanou databázi osob, míst a dalších entit. Podstatou implementovaného programu je určení typu entity na základě typických znaků, které ji charakterizují, a extrakce nejdůležitějších atributů této entity v českém a slovenském jazyce. Výsledkem práce je báze znalostí umožňující snadné vyhledávání a třídění informací. Díky snadné rozšiřitelnosti je možné do programu přidat identifikaci dalších typů entit a dalších vlastností, případně i podporu jiných jazyků.

Abstract

The target of this thesis is to analyze articles on the Wikipedia internet encyclopedia and to convert their text written in natural language into a structured database of persons, places and other entities. The essence of the implemented program is the determination of the type of entity based on its typical characteristics, and the extraction of the most important attributes of this entity in the Czech and Slovak languages. The result of this task is a knowledge base allowing simple searching and sorting of information. Thanks to its easy extensibility, it is possible to add identification of other types of entities and other features to the program, as well as a support of other languages.

Klíčová slova

Wikipedie, extrakce informací, analýza textu, atributy entit

Keywords

Wikipedia, information extraction, text mining, entity atributes

Citace

ŠVUB, Daniel. *Určování typů a atributů entit napříč jazyky.* Brno, 2019. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Určování typů a atributů entit napříč jazyky

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. Pavla Smrže. Další informace mi poskytli jeho kolegové Ing. Otrusina a Ing. Volf. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Daniel Švub
15. května 2019

Poděkování

Děkuji doc. RNDr. Pavlu Smržovi, Ph.D. za vedení a odbornou pomoc při tvorbě této práce. Rovněž děkuji Ing. Lubomíru Otrusinovi a Ing. Tomáši Volfovi za poskytnuté informace a zdroje.

Obsah

Kapitola 1

Úvod

Wikipedie je jedním z nejrozsáhlejších zdrojů informací na internetu. Obsahuje statisíce článků a některé z nich mohou mít desítky kapitol. Jak ale v tomto množství dat vyhledat konkrétní údaj? Pro člověka to obvykle bývá snadné, ovšem jak umožnit počítači, aby informaci přečetl za nás a dále s ní pracoval? Chceme-li využít Wikipedie například pro statistiku nebo automatizované vyhledávání, neexistuje snadný způsob, potřebujeme automaticky rozeznat články o osobách, městech, nebo hudebních dílech.

Tato práce se zabývá problematikou získání základních informací z článku a jejich uložení v organizované podobě umožňující rychlé vyhledání a třídění dat. Jednou z hlavních předností je rozšiřitelnost. Celá práce je funkční napříč jazyky a momentálně aktivně funguje pro český a slovenský jazyk (pro který se navíc jedná o vůbec první počin podobného typu).

Jak zjistit, zda článek pojednává o osobě, aniž bychom jej četli? Jak určit pohlaví nebo národnost této osoby? Právě tyto údaje mohou být velice užitečné pro účely statistiky. Je možné se například dozvědět, kolik procent žen na české Wikipedii pochází z Velké Británie. Můžeme evidovat, kolik obcí v České republice má status města a jaký je jejich poměr oproti obcím, které jej nemají.

Rovněž je možné mít přehled o tom, které články patří do jakých kategorií a případně obsah encyklopedie podle tohoto kritéria třídit. Existuje článek náležící dvěma konkrétním kategoriím zároveň? A pokud ano, o čem tento článek pojednává? Všechny tyto otázky jsme schopni zodpovědět pomocí výsledků této práce a poznatků získaných při její tvorbě.

Následující stránky popisují strukturu dat na Wikipedii, způsob jejich formátování, princip interních odkazů a šablon, zvyklosti wikipedistů při jejich psaní. Také pojednávají o způsobech zefektivnění práce s tímto webovým portálem a využití souborů, které její provozovatel exportuje. Možnostech téměř okamžitého získání znalostí, kvůli kterým by bylo obětovat desítky tisíc hodin studiu, a možnosti rozšíření této práce pro získání dalších a dalších znalostí.

Kapitola 2

Rozbor řešené problematiky

Účelem této práce je transformace článků na české a slovenské verzi internetové encyklopedie Wikipedie na strukturovanou bázi znalostí. Aby to bylo možné, je třeba pochopit problematiku získávání informací z textů psaných běžnou řečí, seznámit se s Wikipedií a jejím značkovacím jazykem, a především vytvořit algoritmus schopný extrakce informací, o které máme zájem, z jejích článků.

2.1 Extrakce informací z textu

Extrakce informací z textu je jedním z nejčastějších cílů tzv. textové analýzy („dolování v textech“, text miningu), což je z obecného hlediska zpracování dokumentů psaných v přirozeném jazyce bez zásahu člověka, jedno z odvětví oboru zvaného data mining. K tomuto zpracování využívá kombinaci strojového učení, statistické analýzy a předem daných algoritmů. Přirozený jazyk ovšem ze své podstaty není navržen pro čtení strojem. Hlavními problémy jsou absence pevně dané struktury, mnohoznačnost slov, obrazná vyjádření, ustálené slovní obraty a další okolnosti typické pro lidskou komunikaci. Právě překonáním těchto překážek se text mining zabývá.

Výstupem textové analýzy jsou smysluplné informace, které zná autor dokumentu, ale pro vnějšího pozorovatele nejsou patrné, dokud text nepřečte. Účelem je člověka čtení textu ušetřit. Oproti vyhledávání informací v textu, při kterém známe hledaný objekt a dotazujeme se na jeho přítomnost, případně sháníme další podrobnosti o něm, je tedy extrakce přesně opačný postup. Pomocí text miningu dostáváme informace doposud neznámé. Během analýzy je proto nejprve třeba určit, které informace jsou pro danou aplikaci podstatné.

Proces textové analýzy dokumentu lze rozdělit do dvou etap: předzpracování textu a samotného získávání informací. Během předzpracování (preprocessingu) je dokument pomocí různých metod převeden do standardizovaného formátu vhodného pro další zpracování. Těmito metodami můžou být převod do základního tvaru (tzv. lemmatizace), kdy je pro jednotlivá významová slova v textu za použití morfologické analýzy nalezen jejich slovníkový tvar (první pád jednotného čísla v případě podstatných jmen a zájmen, první pád jednotného čísla mužského rodu prvního stupně u jmen přídavných, v případě sloves potom infinitiv), či vytvoření frekvenčního slovníku, který obsahuje četnosti výskytu jednotlivých

slov, slovníku synonym nebo negativního slovníku, obsahujícího slova, která nenesou význam (předložky, spojky, některá slovesa). [2]

Textová analýza může mít mnoho různých aplikací, například [14]:

- **Kategorizace textu** – Rozdělování dokumentů do předem definovaných skupin na základě jejich vlastností (téma, autor, klíčová slova). Každý dokument může patřit do více kategorií.
- **Shlukování (clustering) textů** – Rozdělování dokumentů do skupin podle jejich vzájemné podobnosti. Každý dokument náleží do jednoho shluku. Dalším krokem obvykle bývá identifikace společných znaků jednotlivých skupin.
- **Shrnutí textu (summary extraction)** – Automatické vytvoření výtahu z dokumentu. Z textu jsou vybrány nejdůležitější části (přičemž důležitost je obvykle definována uživatelem).
- **Analýza citového zabarvení** – Rozdělování dokumentů do skupin dle množství emocionálního obsahu, které se v nich vyskytuje. Může být použito k vyhledávání nenávistných textů.
- **Identifikace jazyka** – Identifikování jazyka, ve kterém je dokument napsán.
- **Rozdělení dokumentů** – Rozčlenění vstupního dokumentu na menší tematické celky.
- **Extrakce informací** – Převod nestrukturovaného (či částečně strukturovaného) dokumentu do strukturované podoby vhodné k dalšímu zpracování. Právě tím se dále zabývá tato práce.

Množství nestrukturovaných informací na internetu (například internetové články nebo elektronické dokumenty) výrazně převyšuje množství informací dostupné ve strukturované formě (například databáze). Abychom tyto informace mohli použít pro účely vyhledávání, klasifikace nebo statistiky, musíme je do strukturované podoby převést. K tomu je nutné rozpoznat v textu pojmenovanou entitu a následně vyhledat a extrahovat její vlastnosti. Pojmenovanou entitou (named entity) rozumíme slova nebo slovní spojení vystupující v textu jako názvy konkrétních osob, míst, věcí, časových období a tak dále. Jedná se o jazykový prostředek jednoznačně odkazující ke konkrétnímu objektu.

Pojmenované entity jsou běžnou součástí textů v přirozeném jazyce a člověk je v dokumentu obvykle snadno a přirozeně rozezná, v případě čtení strojem je ovšem rozpoznání jazykového kontextu podstatně složitější problém. Kromě samotné existence entity v textu je také nutné rozpoznat další slova (obecná podstatná jména, zájmena), která na tuto entitu odkazují. Je rovněž užitečné identifikovat, že se celý dokument zabývá jednou konkrétní entitou a ostatní entity v článku s ní nějak souvisí. Nalezneme-li například v dokumentu pojednávajícím o osobě další vlastní jméno (v oboru textové analýzy lze chápat jako podmnožinu pojmenovaných entit), pak toto jméno může být jménem jiné osoby (evidentně v nějakém vztahu k původní osobě), názvem státu, z něhož osoba pochází, jménem města, ve kterém se osoba narodila nebo ve kterém působila, nebo například název historického období, se kterým je osoba spojována. [22]

Extrakce informací je obvykle tvořena následujícími činnostmi [15]:

- **Lexikální analýza** – Text dokumentu je rozdělen do vět a následně do tokenů (slov nebo interpunkčních znamének). Každý token je analyzován za účelem určení jeho slovního druhu a dalších parametrů.
- **Identifikace pojmenovaných entit** – Mezi tokeny jsou (za pomocí různých kritérií, mimo jiné kontextu, v jakém se slovo nachází, tedy částečné syntaktické analýzy) identifikovány pojmenované entity.
- **Porovnávání vzorů pro oblast zájmu** – Vlastní vyhledávání stěžejních informací. Za použití předem daných vzorů jsou extrahována data relevantní pro určitou oblast zájmu.
- **Identifikace referencí** – Identifikace slov odkazujících na jiná slova (například zájmeno zastupující podstatné jméno nebo synonyma popisující jeden objekt).
- **Spojování informací** – V některých případech může být informace rozdělena do více vět nebo i odstavců a věta samostatně nemusí dávat smysl. Pro extrakci celé informace je tak nutné zahrnout celý kontext.

Jak už bylo zmíněno, čtení přirozeného jazyka pro stroj představuje poměrně složitý problém. Přestože se většina jazyků řídí na první pohled striktními gramatickými pravidly, reálně může s využitím prvků, jako jsou ironie, obrazná pojmenování nebo nadsázka, jazyk popisovat naprosto jinou skutečnost, než která je popisována. Ačkoliv tedy pro některé účely může stačit kategorický pohled na strukturu jazyka, má své limity. V některých případech může být i naprosto spisovný jazyk nejednoznačný (stejné tvary slov, slova popisující více skutečností nebo taková, která mohou být více slovními druhy). [6]

Je také nutné zvážit lidský faktor. Autor dokumentu se mohl dopustit pravopisných, gramatických či typografických chyb, čímž je extrakce informací ještě více ztížena a často prakticky nerealizovatelná. Není-li jisté slovo ve slovníku, je možné jako řešení použít tzv. „guesser“, tedy nástroj poskytující návrhy slov ze slovníku, které se původnímu slovu podobají. Může se však stát, že slovo i s chybou dává smysl (přestože může být diametrálně odlišný). Tehdy je možné použít automatickou kontrolu gramatiky (například ošetření chyb ve shodě přísudku a podmětu v češtině pomocí určení rodu podstatného jména), ne vždy je ale možné uspět.

2.2 Wikipedie

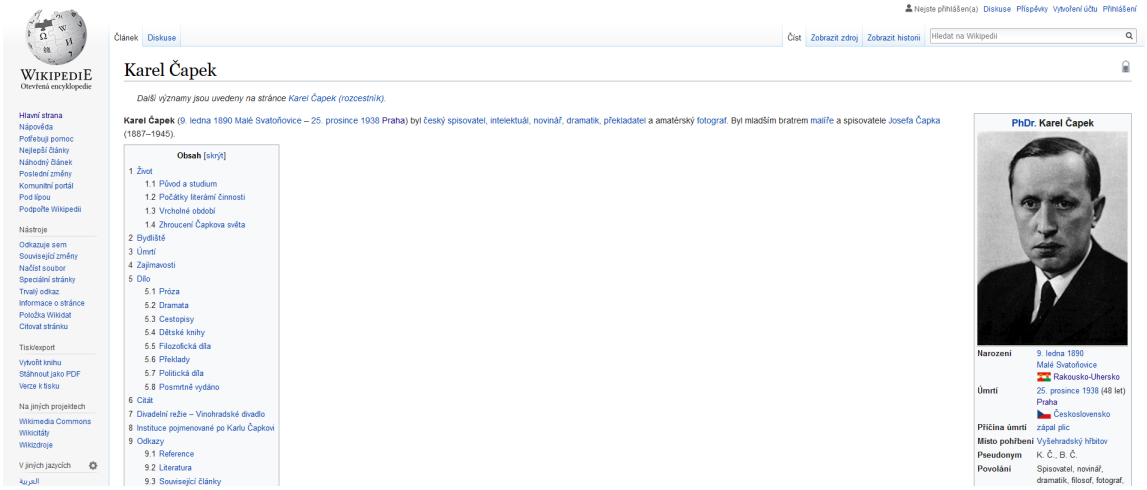
Wikipedie (v originále Wikipedia) je mnohojazyčná webová encyklopedie s otevřeným obsahem, na jejíž tvorbě spolupracují dobrovolní přispěvatelé z celého světa [12]. Jedná se o jeden z projektů nadace Wikimedia (Wikimedia Foundation) založené Jimmy Walesem běžících na systému MediaWiki, který nadace sama vyvíjí. Byla spuštěna 15. ledna 2001 a v současnosti jde o jeden z nejnavštěvovanějších portálů na celém internetu. Dalšími projekty nadace jsou například Wikislovník, Wikicitáty, Wikidata nebo Wikimedia Commons, přičemž mnohé z nich jsou s Wikipedií úzce propojeny. Celkem existuje 289 jazykových verzí Wikipedie.

Obrázek 2.1: Výběr jazykové verze na hlavní stránce Wikipedie.

K formátování stránek celá platforma využívá vlastní značkovací jazyk, wiki markup language [18]. Dokument psaný v tomto jazyce se nazývá wikitext. Existuje dvojí základní dělení používaných značek: dle lokace použití (některé značky lze použít kdekoli, jiné jen na začátku řádku) a dle párovosti. Nepárové značky se užívají například pro tvorbu seznamů (podle typu seznamu `*` nebo `#` na začátku řádku seznamu v počtu podle úrovně seznamu), přesměrování (`#REDIRECT` nebo lokalizovaný ekvivalent na začátku řádku), formátování tabulek (řádky začínající znakem `|` mezi dvojicemi znaků `{|` a `|}`) nebo vkládání interních proměnných (různý počet znaků ~ kdekoli v textu). Párové značky identifikují nadpisy (text ohraničený z obou stran znaky `=` v počtu podle úrovně nadpisu na začátku řádku), odkazy (text v hranatých závorkách, u vnitřních odkazů dvojitých), formátovaný text (text ohraničený z obou stran znaky `''` je psán kurzívou, text ohraničený znaky `'''` je tučný) nebo vložení šablony (název šablony ve dvojici složených závorek, s případnými dopňujícími parametry uvozenými znakem `|`). Je též možné použít vybrané značky jazyka HTML (komentáře, zalomení řádku, základní formátování a podobně), případně jiné značky ve formátu XML (`<nowiki>` pro vypnutí interpretace wiki značek, `<ref>` pro označení bibliografické citace a další).

Mimo běžných článků se na Wikipedii nachází také řada administrativních a informačních stránek o Wikipedii samotné. Jde například o seznamy nebo popisy šablon a kategorií, návody pro psaní článků, stránky o uživatelích. Každý článek má navíc svou diskusní stránku. K základnímu určení zaměření stránky slouží tzv. jmenné prostory (namespaces, zkráceně ns) [20], prefixy před názvem stránky. Každý z nich má vlastní číslo, například prefix „Wikipedie:“ je pod číslem 4, „Šablona:“ pod číslem 10 a tak dále. Obecně platí, že sudá čísla

patří typům stránek, lichá potom diskusím k nim. Pro účely této práce je validní pouze jmenný prostor 0, tedy stránky bez prefixu, články. Tento jmenný prostor ovšem vyžaduje další filtraci, a to z důvodu článků, které ze své podstaty nemohou pojednávat o žádné konkrétní entitě. Jde především o přesměrování na jiné stránky, seznamy nebo tzv. rozcestníky, což jsou rozlišovací stránky odkazující na články o věcech, které jsou obvykle popisovány stejným slovem nebo slovním spojením.



Obrázek 2.2: Stránka na české Wikipedii.

Články na Wikipedii jsou bohaté na vnitřní odkazy, které umožňují čtenáři zobrazit vysvětlení použitého pojmu nebo názvu jediným kliknutím. Tyto odkazy nejsou nijak vázány na konkrétní stránky, je tedy možné se okazovat na stránku neexistující (v takovém případě může odkaz nabýt validitu v budoucnu, pokud je stránka vytvořena). Speciální typ odkazů, dostupný ze spodní části zobrazovaných stránek, je odkaz na kategorii (stránku ve jmenném prostoru 14). Každý článek může spadat do libovolného množství kategorií, což umožňuje přibližnou identifikaci jeho obsahu.

Jak již bylo naznačeno výše, lze na Wikipedii využít šablon, tedy stránek, jejichž obsah lze vložit do jiné stránky. Pravděpodobně nejvýznamnějším typem šablony je tzv. infobox [11], tabulka obsahující základní informace o entitě, o níž článek pojednává. Těchto tabulek Wikipedie rozeznává větší počet a liší se nejen názvy buněk, ale i vizuálním stylem. Právě odsud lze získat nejvíce obecných informací o dané entitě. Ne vždy je to ovšem možné, neboť některé infoboxy získávají data ze spřízněného portálu Wikidata (což umožňuje konzistenci dynamických údajů napříč jazyky) a údaje se v takovém případě ve zdrojovém textu stránky fyzicky nenacházejí.

Přibližně dvakrát měsíčně nadace Wikimedia na svých stránkách dává k dispozici aktuální obsah celé Wikipedie ke stažení ve formě tzv. dumpů [19]. Jedná se o sadu XML souborů obsahujících určité druhy dat z určité jazykové verze k určitému datu. Existují dump soubory obsahující názvy všech stránek (soubor „all-titles“), jejich obsah včetně historie úprav (soubor „pages-meta-history“), informace o kategoriích (soubor „category“) či metadata médií na stránkách použitých (soubor „image“). Pro účely této práce je používán soubor s názvem „pages-articles“, v němž se nachází kompletní text aktuální verze všech stránek dané jazykové verze Wikipedie.

Dump soubory jsou ke stažení k dispozici v komprimované podobě, nejčastěji ve formátu `.gz` nebo `.bz2`. Rozbalený soubor (tedy vstup programu) pro českou verzi, která obsahuje celkem 400 tisíc článků, má velikost cca 3 GB, pro verzi slovenskou, obsahující článků téměř 225 tisíc, cca 1.3 GB. Nejrozsáhlejší, anglická verze pro srovnání obsahuje více než 5.5 milionu článků. Její dump soubor dosahuje velikosti kolem 60 GB.

2.3 Báze znalostí

Báze znalostí (knowledge base, KB) je centralizované úložiště komplexních strukturovaných i nestrukturovaných informací vedoucí k získání určitého stupně vědění. Jejím účelem je shromažďování, organizace a vyhledávání informací, šetření času a nákladů. Pojmem „znalost“ jako takovým v prostředí informačních technologií rozumíme souhrn informací v souvislostech shromážděných za určitým účelem. Znalostní báze často bývá součástí expertních systémů zabývajících se získáváním, uchováváním a využitím znalostí s cílem usnadnit řešení problémů.

Mezi nejvýznamnější z existujících bází znalostí pracujících s informacemi z Wikipedie patří DBpedie (DBpedia) [3] a Airpedie (Airpedia) [1], která je na DBpedii z velké části založena. DBpedie je veřejně financovaný projekt zaměřující se na získávání strukturovaných dat z projektů nadace Wikimedia, která následně poskytuje k dispozici veřejnosti ve formátu RDF fungujícího na bázi vztahu dvou pojmenovaných entit (trojice subjekt–vlastnost–odkaz, například „Mona Lisa“–„autor“–„Leonardo da Vinci“) [21]. Airpedie je rozšířením DBpedie a jejím cílem je extrahovat informace i z článků, ve kterých se nenachází infobox, a tím bází znalostí DBpedie rozšiřovat.

Výzkumná skupina znalostních technologií (KNOT) Fakulty informačních technologií Vysokého učení technického v Brně se extrakcí informací z Wikipedie zabývá dlouhodobě. Jako výchozí bod pro tuto práci posloužila klasifikační práce Petra Rusiňáka [13], která se zabývala určováním příznaků článků na anglické a české Wikipedii. Cenným zdrojem informací byla také práce Michala Planičky [5], jejímž cílem byla obdobná báze znalostí, avšak jazykově závislá na češtině a s několika nedostatky. V rámci KNOT byly dále vytvořeny skripty pro převod skript pro extrakci prvních vět. Tento skript byl původně v práci použit jako dílčí část, později však jeho výsledky již nedostačovaly a byl nahrazen.

2.4 Jazyková analýza vět v češtině a slovenštině

K úspěšnému určení typu entity práce využívá především tzv. definičních frází a definičních slov. Definiční slovo je podstatné jméno definující obsah stránky (například „město“, „spisovatel“ nebo „album“). Definiční fráze je toto slovo společně se všemi jeho přívlastky (například „okresní město v západních Čechách“, „český spisovatel“ nebo „hudební album“). Těchto definic článku může být libovolný počet a vždy se nacházejí v první větě prvního odstavce daného článku.

Český a slovenský jazyk jsou si velice podobné, lze proto k jejich analýze přistupovat stejným způsobem. Jazyková analýza má několik úrovní. Na úrovni nejnižší můžeme pracovat s jednotlivými slovy, každé z nich lze v češtině a slovenštině rozdělit na předpony, kořen a přípony a následně dle těchto parametrů určit jeho vlastnosti. Kupříkladu některé předpony

zcela mění význam nejen slova, ale i celé věty (předpony jako `ne-`, `nej-`, `od-` a podobně). Získáním kořene slova je zase možné určit význam slova na nejzákladnější úrovni. Přípona potom určuje gramatický význam slova a koncovka (zadní část poslední z přípon) rozlišuje jednotlivé jeho tvary. Tím se dostáváme na vyšší úroveň.

Po rozpoznání slovního druhu a mluvnických kategorií jako pádu, čísla nebo rodu, u sloves způsobu, osoby, či času můžeme pomocí syntaktické analýzy sestavit syntaktický strom aktuální věty. To může být, jak je zmíněno výše, velmi náročné, protože přirozený jazyk nemusí mít vždy pevně danou strukturu. Pokud ovšem budeme cílit na encyklopedické, odborné články, lze předpokládat, že výskyt jevů bránících úspěšné syntaktické analýze by měl být minimální. [16]

K úspěšné extrakci definičních frází a slov na Wikipedii je třeba provést jazykovou analýzu první věty článku. Nejdříve je tedy nezbytné určit druhy a mluvnické kategorie jednotlivých slov. K tomuto účelu existuje řada nástrojů, jedním z nich je nástroj MorphoDiTa (celým názvem Morphological Dictionary and Tagger), vyvinut v rámci Ústavu formální a aplikované lingvistiky (ÚFAL) Matematicko-fyzikální fakulty (MFF) Univerzity Karlovy v Praze [17]. Jedná se o analyzátor přirozeného jazyka, jehož prostřednictvím lze získat atributy slov ve větě a jejich základní tvar. Je možné jej použít přímo z příkazové řádky nebo jako knihovnu v jazycích jako C++ či Python.

MorphoDiTa umožňuje „tokenizaci“ (rozklad textu na jednotlivá slova), jazykovou analýzu slov (v češtině a slovenštině například slovní druh a již zmíněné mluvnické kategorie), jejich konverzi na základní tvar (takzvané „lemma“, tedy infinitivy u sloves, první pády u jmen) a případné návrhy slov v případě, že zadané slovo v přesném znění ve slovníku neexistuje. Pokud může určité slovo být zástupce více slovních druhů nebo je ve tvaru, v němž jsou jeho atributy nejednoznačné, vrátí MorphoDiTa všechny možné eventuality.

Většina úvodních vět na Wikipedii začíná názvem článku (v některých případech následovaným dovysvětlující závorkou), pokračuje tvarem slovesa „být“ a definiční frází nebo frázemi. Příkladem může být věta „Karel Čapek (9. ledna 1890 Malé Svatoňovice – 25. prosince 1938 Praha) byl český spisovatel, intelektuál, novinář, dramatik, překladatel a amatérský fotograf.“, ze které lze extrahovat celkem 6 definičních frází: český spisovatel, intelektuál, novinář, dramatik, překladatel a amatérský fotograf. Definiční slova jsou potom spisovatel, intelektuál, novinář, dramatik, překladatel a fotograf. Extrakce z takto formulovaných vět je poměrně snadná, stačí prohledávat první větu článku, dokud nenajdeme sloveso, jehož základním tvarem je „být“. Za tímto slovem se velmi pravděpodobně nachází definiční fráze oddělené čárkami nebo spojkami.

Základem definiční fráze je (jak již bylo zmíněno) podstatné jméno, definiční slovo. To může (ale nemusí) být předcházeno jmény přídavnými (shodnými přívlastky), a následováno zbytkem fráze (řada slov libovolných slovních druhů, které dohromady tvoří neshodný přívlastek), jenž končí začátkem další fráze nebo koncem věty. Tato fráze pak velmi pravděpodobně obsahuje údaje přímo se vztahující k entitě.

V některých případech může být sktruktura úvodní věty odlišná. Například ve větě „Obec Rapotín (německy Reitendorf) se nachází v okrese Šumperk v Olomouckém kraji.“ se tvar slovesa „být“ vůbec nevyskytuje a definiční slovo se nachází před názvem článku. Ve větě „Tunis (někdy také Túnis) je hlavním městem Tuniska.“ je zase definiční slovo v sedmém pádě a nikoli v prvním, jak bývá běžně zvykem. Je ovšem třeba odlišit definiční slovo od běžného podstatného jména v sedmém pádě, za nímž následuje klasické definiční slovo

v pádě prvním, například ve větě „Brno je počtem obyvatel i rozlohou druhé největší město v České republice.“ by definičním slovem nebylo slovo „počet“, ale „město“. Je tedy nutné vyhledávat definiční slova v sedmém pádě až poté, když žádná v prvním pádě nenajdeme.

Jindy může být první věta formulována nepřímo, příkladem může být věta „Balkánské tažení je souhrnný název pro boje na Balkáně za druhé světové války.“, kde je definičním slovem podstatné jméno „boj“, nikoli „název“. Z toho důvodu je nutné monitorovat výskyt slov jako název, jméno, pojmenování nebo označení bezprostředně po tvaru slovesa „být“ a považovat v takovém případě za definiční frázi podstatné jméno ve druhém pádě právě za tímto slovem, případě v pádě čtvrtém za předložkou „pro“.

Některé články ovšem nesplňují žádný z výše uvedených stereotypů a jsou formulovány naprosto atypicky. Jelikož je v těchto případech velice těžké první větu strojově přečíst, je téměř nemožné definiční fráze extrahovat. Dalším typickým problémem při extrakci definic je víceznačnost tvarů některých slov. Například ve větě „Abuja (vyslov abudža) či Abudža je hlavním městem Nigérie v západní Africe.“, je definičním slovem „město“, jelikož je ovšem v sedmém pádě, při strojovém čtení je třeba se nejprve ujistit, že ve větě neexistuje definiční slovo v pádě prvním. To ale najdeme, je jím vlastní jméno „Nigérie“, které, ačkoliv je ve větě použito v druhém pádě, jedná se o stejný tvar jako v prvním pádě, následkem čehož je chybně určeno jako definiční slovo.

Jelikož je Wikipedie veřejnou službou, do níž může přispívat každý, jsou některé články navíc poznamenány pravopisnými, gramatickými či typografickými chybami. Mezi nejtypičtější chyby patří špatné psaní závorek nebo interpunkčních znamének. S těmito překlepy pochopitelně strojové čtení nemůže počítat.

Kapitola 3

Návrh a implementace

Výsledkem práce je počítačový program implementován v programovacím jazyce Python 3 za využití nástroje MorphoDiTa a morfologického slovníku v externím souboru. Program se skládá z hlavního, spouštěcího skriptu, několika dalších skriptů a množství pomocných souborů.

3.1 Postup práce

Prvním krokem, zadaným předem vedoucím práce, byla tvorba skriptu pro extrakci definičních frází a slov článků z dumpu Wikipedie, tedy souboru ve formátu XML obsahujícího všechny stránky dané jazykové verze této encyklopedie. Jako jazyk pro počáteční pokusy byla z důvodu malého počtu článků a jejich krátkého rozsahu zvolena slovenská verze. Pro implementaci skriptu byl vybrán jazyk Python 3 [8]. Intuitivně podporuje nástroj MorphoDiTa, zvolený pro jazykovou analýzu slov, umožňuje snadnou práci s řetězci a byly v něm implementovány i předchozí projekty v rámci KNOT.

Nejprve bylo třeba zpracovat XML strukturu v dumpu Wikipedie. K tomu byla užit nástroj ElementTree ve standardní knihovně jazyka Python, která, jak prozrazuje její název, strukturu uloží jako strom, se kterým lze dále pracovat.

Jak bylo popsáno výše, ne všechny stránky jsou články a ne všechny články mohou pojednávat o entitě. Je tedy nutné stránky filtrovat, abychom předešli zbytečnému plýtvání zdrojů. Základním filtrem je jmenný prostor – má smysl dále zpracovat pouze články, tedy stránky patřící do jmenného prostoru 0. Nemá rovněž smysl zabývat se rozcestníky, seznamy a dalšími typy článků, které ze své podstaty entitou nikdy nemohou být.

K extrakci definic bylo nezbytné získat první věty článků. Jelikož v rámci KNOT byl stejný problém již řešen, byl k tomuto účelu nejprve přejat starší skript [4], který v prvé řadě zavolá skript třetí strany pro převod wikitextu na prostý text a následně z textu článku prostého formátovacích značek extrahuje první větu nebo odstavec. Z počátku se skript využíval nezměněn včetně všech pomocných souborů, brzy ovšem bylo třeba nahradit seznam českých zkratek seznamem zkratek slovenských sestaveným pomocí mnoha webových zdrojů a následně doplněným o mnohé další zkratky. Nástrojem MorphoDiTa byla tato věta poté rozložena na jednotlivá slova pomocí vestavěného „tokenizéru“ a analyzována.

Za tímto účelem byl vedoucím práce poskytnut morfologický slovník pro slovenský jazyk, neboť ten se na webových stránkách nástroje nenachází.

V této chvíli bylo možné přistoupit ke klasifikaci entit, a to za použití extrahovaných definic a seznamu typických klíčových slov pro osoby, místa, organizace a události. Později byl přidán také pátý typ, dílo. Výsledky takovéto identifikace byly značně neúplné a nepřesné, proto se přešlo k podrobnější analýze vlastního wikitextu. Především k vyhledávání infoboxů a odkazů na kategorie, pomocí čehož bylo možné šanci na úspěšnou identifikaci entity velmi zvýšit. K tomu bylo nutné vytvořit také seznamy infoboxů typických pro daný typ [9] a prefixů kategorií. Jelikož ale samotná shoda definičních slov, názvů infoboxů a názvů kategorií nemusí automaticky znamenat příslušnost k určitému typu entity (definiční slovo může být chybně určeno, infobox autorem článku chybně vybrán, kategorie nejednoznačná), je nutné alespoň dvou shod pro konečné prohlášení článku za entitu. V případě kategorií je navíc za shodu považováno pouze úspěšné nalezení alespoň dvou odpovídajících odkazů.

Úspěšnost identifikace byla v tomto bodě kolem šedesáti procent, stále více však přibývalo problémů s externím extrakčním skriptem, který často chybně ukončoval věty nebo ponechával ve výstupním textu formátovací značky. Uzavřenost skriptu a nemožnost výrazněji ovlivnit jeho funkci vyústila ve tvorbu skriptu vlastního pro získávání prvních vět a později i konvertoru wikitextu na prostý text, jelikož v knihovně jazyka Python ani repositáři pro software třetích stran nebyl nalezen žádný nástroj, jenž by dosahoval uspokojivých výsledků. Implementace vlastních algoritmů navíc zrychlila funkci skriptu a lehce snížila paměťové nároky.

Dalším bodem se stala extrakce atributů entit. Největší množství atributů lze určit u osob. Datum a místo narození (a v případě již nežijící osoby i úmrtí) lze získat věma způsoby. Z první věty, kde tyto informace bývají uvedeny v závorce za jménem osoby, nebo z infoboxu. V naprosté většině těchto šablon se buď nachází klíče „datum narození“ a „místo narození“ (případně i „datum úmrtí“ a „místo úmrtí“), nebo jsou tyto informaci spojeny v jediný klíč „datum a místo narození“ (případně i „datum a místo úmrtí“). Existencí nebo naopak absencí dat souvisejících s úmrtím osoby lze také určit, zda je daný člověk stále naživu. Jedním z nejdůležitějších atributů osoby je její národnost. Tu lze nejjednodušeji najít mezi přídavnými jmény v definičních frázích (například „český spisovatel“ či „americký herec“). Aby bylo možné přesné určení národnosti, byly vytvořeny seznamy států a národnostních přídavných jmen. Jejich počet a pořadí je shodný, díky čemuž je možné přesně spojit osobu se zemí původu. Navíc v adresáři práce existuje také anglický seznam adjektiv, opět se stejným počtem řádků a stejným pořádím. To dává možnost ve výstupu vypsat národnost osoby v anglickém jazyce a docílit jazykové nezávislosti výsledné báze znalostí. Pohlaví osoby je možné s poměrně vysokou přesností identifikovat určením rodu definičních slov (například „spisovatel“ nebo „spisovatelka“).

U míst lze určit stát, ve kterém se místo nachází (pochopitelně v případě, že se jedná o místo nižší úrovně než státní). Ve většině článků lze postupovat obdobně jako u národností osob („slovenské město“) nebo hledáním odkazů na státy v první větě článku (například „Praha je hlavní město Česka“), k čemuž je nutná první věta se zachováním formátovacích značek pro odkazy. Extrakce podrobnějších atributů jako počtu obyvatel obcí je prakticky znemožněna faktem, že se většina geografických infoboxů odkazuje na portál Wikidata a informace se tedy fyzicky nenacházejí v dump souboru. Atributy organizací má smysl extrahovat pouze z infoboxů. Významnými údaji jsou jména zakladatele (nebo zakladatelů) a jejího ředitele či předsedy, dále potom datum založení organizace. U událostí má smysl extrahovat datum

(nebo jejich rozsah), kdy se událost odehrála. V případě děl se nabízí rovněž datum a následně autor díla.

V důsledku extrakce atributů se chod skriptu výrazně zpomalil. To si společně s neúnosnou paměťovou náročností (skript v paměti po celou dobu běhu udržoval kompletní vstupní soubor) vyžádalo výrazné změny ve způsobu otevírání a prohledávání dump souboru. Zpracovávání celého souboru nástrojem ElementTree bylo nahrazeno sekvenčním čtením s vyhledáváním XML tagů `<page>` a `</page>`. Nástroj ElementTree nově zpracovává pouze text těmito tagy ohraničený, a to pro každou stránku zvlášť. Kořenovým elementem tedy není `<mediawiki>`, ale právě `<page>`. Program byl reorganizován do více skriptů a vybaven ukazatelem procentuálního vyjádření postupu s automatickým výpočtem doby běhu. Doba zpracovávání dumpu slovenské Wikipedie se snížila z téměř 30 minut na minut 7, paměťová náročnost z více než 1 GB na cca 30 MB.

V této fázi práce program stále pracoval pouze se slovenskou verzí Wikipedie. Stal se tak vůbec prvním produktem v rámci KNOT zabývajícím se tímto jazykem, cílem práce však měl být také český jazyk. Díky podobné struktuře obou jazyků byla nutná pouze drobná modifikace algoritmu pro extrakci definičních slov a frází, v identifikačních skriptech a dalších částech programu pracujících se specifickými řetězci ve slovenštině se musely klíčové části nahradit proměnnými čtenými ze souboru, do něhož byly takové řetězce uloženy. Byly vytvořeny adresáře pro slovenský a český jazyk, v nichž se kromě slovenských a českých seznamů zkratek (přičemž český seznam byl částečně přejat z původního skriptu pro extrakci prvních vět) vyskytují též právě seznamy řetězců použitých v jednotlivých skriptech. Jazyk dump souboru je automaticky rozpoznáván podle hodnoty atributu `xml:lang` kořenového elementu `<mediawiki>`. Rovněž bylo třeba ošetřit eventuality, které se nacházejí na české Wikipedii, zatímco na slovenské nikoliv.

Následovaly testování a průběžná oprava chyb. Především docházelo k vylepšování extraktoru definičních frází (obohacení o extrakci z nepřímo formulovaných vět, extrakci definic v sedmém pádě a tak dále) a zvyšování efektivity identifikace typů. Po domluvě s vedoucím práce došlo ke změně formátu výstupu a přidání extrahovaných seznamů infoboxů a kategorií příslušících článku. V poslední fázi bylo třeba orientačně ověřit úspěšnost programu. K tomu se využíval seznam osob extrahovaných z anglické Wikipedie jiným nástrojem vyvinutým v rámci KNOT. Testovací skript vyhledává články nacházející se jak v seznamu, tak mezi zpracovanými články, a následně ověří, které z nich byly korektně určeny jako osoby. Na české Wikipedii program dosahuje úspěšnosti 98 %, na slovenské verzi 96 %.

Výstup identifikace typů a extrakce atributů byl naformátován dle konvence užívané v rámci KNOT, díky čemuž jeho činností vzniká plnohodnotná báze znalostí. Nakonec byl implementován porovnávač dvou verzí zmíněné báze. Lze tak odhalit určité nedostatky identifikace a extrakce atributů a případně pokračovat v další činnosti. Díky snadné rozšiřitelnosti programu je možné jednoduše přidat identifikaci dalších typů, stejně jako extrakci dalších atributů. V případě implementace extrakce definičních slov z anglických vět je možné též bez větších problémů rozšířit funkci programu i na anglický jazyk.

3.2 Struktura programu

Adresář projektu obsahuje následující:

- **main.py** – Hlavní, spouštěcí skript. Používá funkce z extrakčního skriptu a funkce `init` a `identify` v identifikačních skriptech.
- **extract.py** – Extrakční skript. Obsahuje extrakční funkce využívané napříč ostatními skripty:
  - `init` – Načítá seznam lokalizovaných řetězců používaných v ostatních funkcích do globální proměnné na začátku běhu programu.
  - `extract` – Extrahuje definiční fráze a slova z první věty článku.
  - `get_sentence` – Extrahuje první větu z prostého textu.
  - `convert` – Převádí wikitext na prostý text.
  - `clean` – Odstranění formátovacích značek z wikitextu.
  - `clean_lite` – Odstranění formátovacích značek s výjimkou odkazů a únikových znaků (escape characters) z wikitextu.
  - `rm_links` – Odstranění odkazů z wikitextu.
  - `category_extract` – Extrahuje seznam kategorií článku.
  - `infobox_extract` – Extrahuje seznam infoboxů v článku.
  - `category_search` – Ověřuje, kolik kategorií odkazovaných v článku se nachází na seznamu kategorií typických pro daný typ entity.
  - `infobox_search` – Ověřuje, zda se některý z infoboxů v článku nachází na seznamu infoboxů typických pro daný typ entity.
- **person.py, place.py, work.py, organization.py, event.py** – Identifikační skripty. Obsahují funkci `init` sloužící k inicializaci globálních proměnných a načtení seznamů a funkci `identify` provádějící proces identifikace
- **Adresář en** – Adresář obsahující anglický seznam národnostních adjektiv.
- **Adresáře cs a sk** – Adresáře jazyků. Obsahují:
  - **Adresáře person, place, work, organization, event** – Adresáře typů. Obsahují:
    - **boxes.txt** – Seznam infoboxů, jejichž přítomnost v článku naznačuje příslušnost k danému typu entity.
    - **categories.txt** – Seznam prefixů kategorií, které odpovídají danému typu entity.
    - **expressions.txt** – Lokalizované řetězce pro použití v příslušném identifikačním skriptu.
    - **words.txt** – Definiční slova typická pro daný typ entity.
  - **abbreviations.txt** – Seznam zkratek v příslušném jazyce. Slouží k extrakci první věty.
  - **countries.txt** – Seznam názvů států v příslušném jazyce. Slouží k lokalizaci míst.
  - **nations.txt** – Seznam národnostních adjektiv v příslušném jazyce. Slouží k lokalizaci míst a určování národnosti osob.

- **eliminate.txt** – Lokalizované řetězce pro použití v hlavním skriptu k filtraci stránek.
- **extract.txt** – Lokalizované řetězce pro použití v extrakčním skriptu.
- **months.txt** – Seznam názvů měsíců v daném jazyce.
- **morpho_model.dic** – Morfologický slovník pro nástroj MorphoDiTa.

3.3 Vstupy a výstupy

Vstupem programu je dump české nebo slovenské jazykové verze Wikipedie, konkrétně soubor „pages-articles“. jedná se o XML soubor, jenž tvoří kořenový element `<mediawiki>` s vnořenými elementy `<page>`. Tyto elementy reprezentují jednotlivé stránky Wikipedie a obsahují další elementy: `<ns>`, obsahující číslo jmenného prostoru, `<id>`, obsahující identifikační číslo stránky, `<title>`, obsahující název článku, a `<revision>`, který je reprezentací konkrétní verze (revize) stránky. Jelikož používaný dump soubor „pages-articles“ obsahuje pouze nejnovější verze stránek, je tento element pouze jeden pro každou stránku (v případě souboru „pages-meta-history“ by jich potenciálně mohlo být více, jeden pro každou verzi článku). Tento element obsahuje další elementy, z nichž jediný validní pro tuto práci je element `<text>`, jehož obsahem je samotný text článku ve značkovacím jazyce MediaWiki.

Všechny výstupy obsahují ve svém jméně celý název vstupního souboru a jsou ve formátu TSV (Tab-Separated Values, tabulátorem oddělené hodnoty) [10], ačkoliv většina z nich nemá odpovídající příponu. Výstupy program ukládá do adresáře „results“, pokud tento adresář neexistuje, je na začátku činnosti skriptu vytvořen. Po skončení běhu obsahuje:

- **Soubor s příponou `.tsv`** – Výstupní soubor úvodní extrakce, mezitvar pro identifikaci a extrakci atributů. Řídí se strukturou smluvenou s vedoucím práce (sjednocený výstup více projektů v rámci KNOT), jeden řádek pro každý zpracovaný článek ve formátu:

  ```
  titulek článku » první věta článku » název prvního infoboxu »
  seznam kategorií » » seznam definičních slov »
  seznam definičních frází » » text prvního infoboxu »
  první věta včetně odkazů » první odstavec »
  první odstavec včetně odkazů
  ```

- **Soubor s příponou `.kb`** – „Knowledge base“. Výsledná báze znalostí. Její struktura je inspirována staršími projekty v rámci KNOT, jeden řádek pro každou identifikovanou entitu ve formátu:

  pro žijící osoby

  ```
  „person:alive" » jméno » popis » pohlaví »
  datum narození » místo narození » definiční slova » národnost » tituly
  ```

  pro nežijící osoby

  ```
  „person:dead" » jméno » popis » pohlaví »
  datum narození » místo narození » datum úmrtí » místo úmrtí
  » definiční slova » národnost » tituly
  ```

  pro fiktivní osoby (postavy)

„person:fictional" » jméno » popis » pohlaví »
datum narození » místo narození » datum úmrtí » místo úmrtí
» definiční slova » národnost

pro sídla (města a obce)

„place:settlement" » jméno » popis » typ » stát

pro vrcholy

„place:mountain" » jméno » popis » stát

pro organizace

„organization" » jméno » popis » hlavní představitel nebo představitelé
» zakladatel nebo zakladatelé

pro události

„organization" » jméno » popis » rok

pro písně nebo skladby

„work:song" » jméno » popis » umělec » rok

pro hudební alba

„work:album" » jméno » popis » umělec » rok

pro malby

„work:painting" » jméno » popis » autor

pro knihy

„work:book" » jméno » popis » autor » rok

- **Soubor s příponou `.ael`** – „Article entity list“. Seznam všech zpracovaných článků a typu entity, byl-li článek jako entita identifikován. Řádky ve formátu:

  `název článku » typ entity (pokud byl článek identifikován jako entita)`

- **Podadresář lists** – Obsahuje soubory seznamů s příponou `.lst` („list“):
  - **Soubor se suffixem „persons“** – Seznam článků identifikovaných jako osoby. Má formát:

    `jméno » reálná/fiktivní » pohlaví » národnost » živá/zemřelá »`
    `datum narození » místo narození » datum úmrtí »`
    `místo úmrtí » tituly`
  - **Soubor se suffixem „characters“** – Seznam článků identifikovaných jako fiktivní osoby (postavy). Formát je následující:

    `jméno » pohlaví » národnost » datum narození » místo narození »`
    `datum úmrtí » místo úmrtí`
  - **Soubor se suffixem „places“** – Seznam článků identifikovaných jako místa ve formátu:

    `název » stát`
  - **Soubor se suffixem „organizations“** – Seznam článků identifikovaných jako organizace. Jeho formát je:

    `název » datum založení v původní podobě »`
    `zakladatel nebo zakladatelé`

  - **Soubor se suffixem „events“** – Seznam článků identifikovaných jako události ve formátu:

    `název » datum události v původní podobě`
  - **Soubor se suffixem „works“** – Seznam článků identifikovaných jako díla. Má formát:

    `název » autor » datum vytvoření díla v původní podobě`

- **Podadresář links** – Obsahuje soubory vazebních seznamů, rovněž s příponou `.lst` („list“), v nichž se nachází pouze entity se všemi pro seznam relevantními atributy úspěšně extrahovanými a vazbami na jiné entity:
  - **Soubor se suffixem „alive“** – Vazební seznam článků identifikovaných jako živé osoby. Má formát:

    `jméno » pohlaví » národnost » datum narození » místo narození`
  - **Soubor se suffixem „dead“** – Vazební seznam článků identifikovaných jako zemřelé osoby. Formát je následující:

    `jméno » pohlaví » národnost » datum narození » místo narození » datum úmrtí » místo úmrtí`
  - **Soubor se suffixem „settlements“** – Vazební seznam článků identifikovaných jako sídla ve formátu:

    `název » typ (město nebo obec) » stát`
  - **Soubor se suffixem „artists“** – Vazební seznam článků identifikovaných jako díla s úspěšně extrahovaným atributem autora. Jeho formát je:

    `název » jméno autora`
  - **Soubor se suffixem „founders“** – Vazební seznam článků identifikovaných jako organizace s úspěšně extrahovaným atributem zakladatele ve formátu:

    `název » jméno zakladatele nebo zakladatelů`
  - **Soubor se suffixem „leaders“** – Vazební seznam článků identifikovaných jako organizace s úspěšně extrahovaným atributem předsedy. Má formát:

    `název » jméno vůdčí osoby nebo osob`

V testovacím režimu se standardní soubory negenerují, namísto nich jsou vytvořeny dva soubory v podadresáři test:

- `matches.txt` – Prostý seznam entit, které prošly testem (tedy takových, které existují v testovacím seznamu i seznamu identifikovaných entit), každá na jednom řádku.
- `errors.txt` – Prostý seznam entit, které existují pouze v testovacím seznamu. Tedy entity, které systém neidentifikoval.

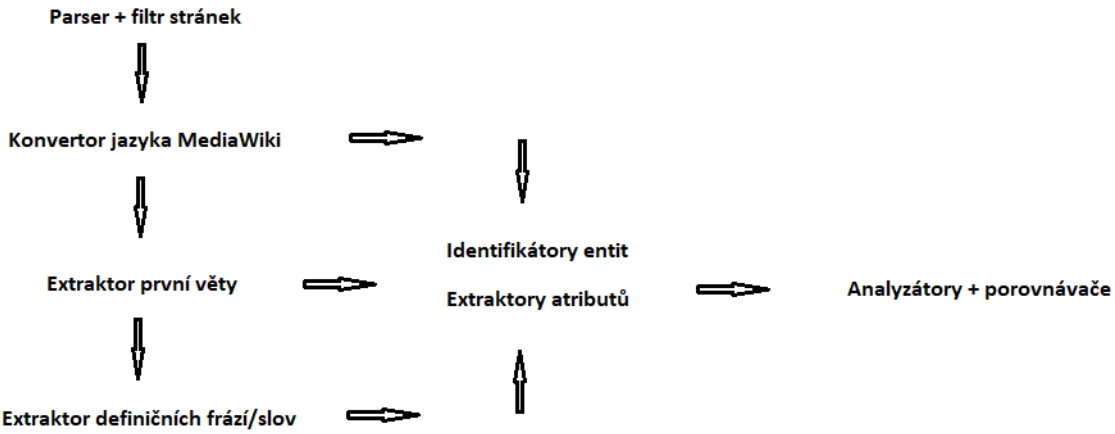
porovnávacím režimu se vytváří soubory v podadresáři compare:

- `matches.txt` – Prostý seznam entit, které prošly testem (tedy takových, které existují ve staré i nové znalostní bázi), každá na jednom řádku.
- `only_old.txt` – Prostý seznam entit, které existují pouze ve staré bázi znalostí. Tedy entity, v jejichž identifikaci systém selhává, ačkoliv předchozí verze byla úspěšná.

- `only_new.txt` – Prostý seznam entit, které existují pouze v nové bázi znalostí. Tedy nové entity, které systém identifikuje.

Pokud adresář „results“ neexistuje, bude automaticky vytvořen včetně potřebných podadresářů. Po skončení běhu programu jsou všechny prázdné (a tedy nevyužité) soubory v podsložkách smazány.

3.4 Popis funkce



Obrázek 3.1: Schéma systému.

Činnost programu začíná otevřením vstupního souboru – dumpu Wikipedie. Je ověřeno, že se jedná o validní XML strukturu, z atributu `xml:lang` je získán jazyk dumpu a v případě, že se jedná o češtinu (`cs`) nebo slovenštinu (`sk`), program spočítá počet řádků souboru a soubor zavře. Dále:

1. Jsou inicializovány identifikační skripty a extrakční skript (uložení zjištěného jazyka dumpu do globálních proměnných a následné otevření pomocných souborů v patřičné složce).
2. Je otevřen morfologický slovník pro nástroj MorphoDiTa, seznam zkratek a seznam řetězců pro filtraci stránek.
3. Jsou vytvořeny adresáře pro výstupy (pokud neexistují) a v nich otevřeny výstupní soubory, přičemž ukazatele na ně uloženy do polí `files` (obecné výstupní soubory) a `links` (vazební seznamy).

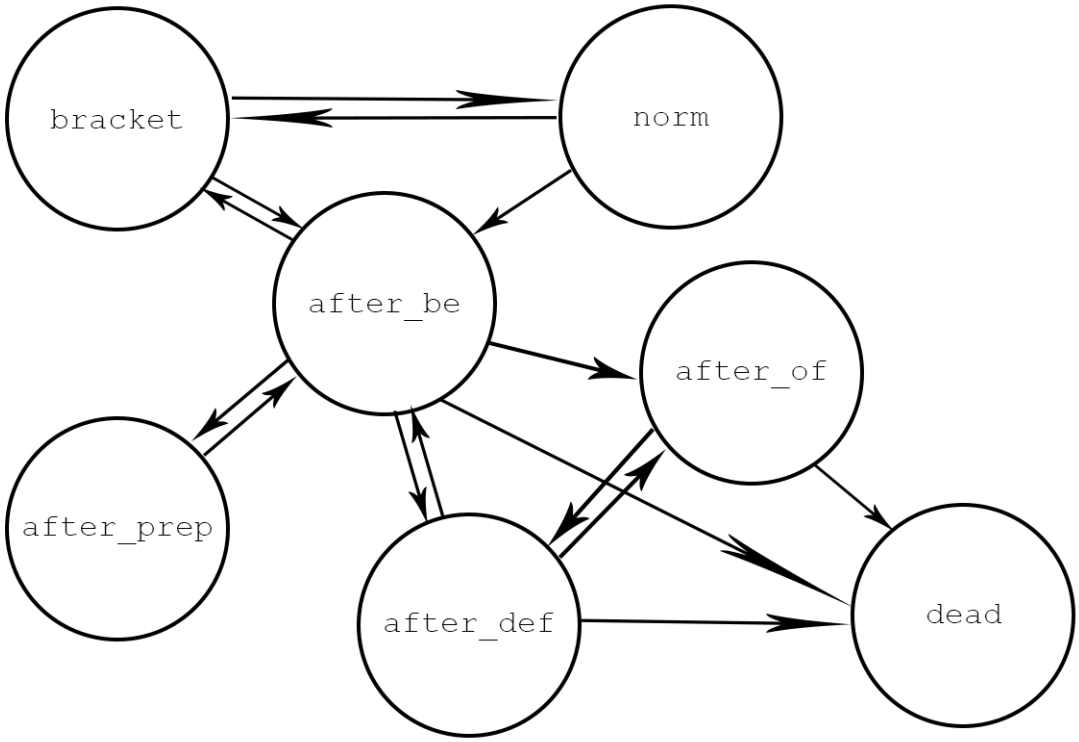
Následně je dump otevřen znovu a čten řádek po řádku. Pokud algoritmus nalezne tag `<page>`, uloží jej a všechny následující řádky po tag `</page>` do proměnné, kterou poté zpracuje nástroj ElementTree. Tím získáváme obsah jednotlivých elementů. Pokud se stránka

nachází ve jmenném prostoru 0 a její název nebo obsah neobsahují předem dané řetězce (z načteného seznamu), je text článku zbaven značek. Tento proces probíhá v několika fázích pomocí nahrazování regulárních výrazů. Text je zbaven vložených šablon, odkazů, znaků, které by mohly při dalším zpracování způsobit potíže, formátování (tučný text, kurzíva, nadpisy) či HTML tagů. Všechny bílé znaky (s výjimkou nového řádku) jsou převedeny na jedinou mezeru, jsou odstraněny prázdné závorky a přebývající interpunkce.

Z tohoto prostého textu jsou extrahovány první věta a první odstavec. Pro extrakci první věty je článek nejdříve ukončen v místě prvního odřádkování a následně jsou čtena všechny slova následovaná tečkou. Pokud se slovo nenachází v seznamu zkratek a splňuje i další kritéria, je prohlášeno za slovo ukončující větu. Z původního, neupraveného textu jsou získány názvy infoboxů a kategorie, do kterých stránka patří.

Z první věty jsou následně extrahovány definiční fráze a definiční slova. Věta je nástrojem MorphoDiTa rozložena na jednotlivé tokeny (jazyková jednotka samostatně analyzovatelná nástrojem MorphoDiTa, slova nebo interpunkční znaménka) a každé je podrobeno morfologické analýze. Algoritmus pro extrakci definic je stavový automat. Prvotním stavem je stav `norm`, tedy normální režim, dalšími jsou:

- `after_be` – Vyhledávací režim, stav následující po detekování tvaru slovesa „být“ v normálním režimu. Je základním vstupním bodem do většiny ostatních stavů. Po detekci přídavného jména je nastaven příznak a toto adjektivum je přidáno do dočasné proměnné. Pokud dalším detekovaným tokenem není přídavné jméno, podstatné jméno, čárka nebo spojka, je tato dočasná proměnná vyprázdněna.
- `after_def` – Stav následující po detekování podstatného jména ve vyhledávacím režimu. Detekce čárky nebo spojky znamená přechod zpět do vyhledávacího režimu (konec definiční fráze), jakýkoli jiný token je považován za součást volného přívlastku a je přidán do dočasné proměnné. Čtení definiční fráze je ukončeno čárkou, spojkou nebo koncem věty. Tehdy je obsah dočasné proměnné zařazen do výstupního řetězce. Podstatné jméno, tedy definiční slovo, je uloženo ještě jednou, samostatně.
- `after_of` – Stav následující po detekování příznaku nepřímo formulované věty (slova jako „název“, „jméno“ nebo „označení“) ve vyhledávacím režimu. Systém vyhledává definiční slova ve druhém pádě, v případě detekce slova „pro“ v pádě čtvrtém.
- `after_prep` – Mezistav následující po detekování předložky ve vyhledávacím režimu zabraňující detekci slova po předložce jako definičního slova (jelikož je jisté, že by tato identifikace byla chybná, neboť definiční slovo po předložce nikdy následovat nemůže). Systém se hned v další iteraci vrací do předchozího režimu.
- `bracket` – Stav následující po detekování otevírací závorky v normálním nebo vyhledávacím režimu, ukončením závorky se systém vrací zpět do předchozího stavu. Umožňuje vícenásobné zanořování závorek.
- `dead` – Stav znamenající ukončení vyhledávání.



Obrázek 3.2: Schéma možností přechodů mezi stavy konečného automatu pro extrakci definičních frází a slov.

Dále program volá funkce `identify` v jednotlivých identifikačních skriptech, které postupně ověřují příslušnost právě zpracovávaného článku k danému typu entity. Pro každou entitu existují čtyři pomocné soubory, uložené v jejím adresáři, v nichž jsou uloženy seznamy definičních slov, infoboxů a kategorií, jejichž přítomnost v článku indikuje příslušnost k danému typu, a seznam lokalizovaných řetězců sloužících především pro vyhledávání informací v infoboxech.

Identifikace osoby a extrakce jejích atributů probíhají následujícím způsobem:

1. V první větě článku je vyhledána závorka s informacemi o narození (a případně i úmrtí) osoby. Pokud existuje, stane se zdrojem těchto informací.

2. V textu článku proběhne vyhledávání infoboxů. Pokud je vyhledávání úspěšné a systém nezískal všechny informace ze závorky, získá data odsud, jsou-li dostupná. Nejprve je vyhledán klíč „datum narození“, pokud existuje a obsahuje validní hodnotu, systém vyhledá „místo narození“. Pokud neexistuje, systém se pokusí vyhledat klíč „datum a místo narození“ a zjistit informace odsud. Dále se vyhledává „datum úmrtí“. Pokud obsahuje hodnotu, osoba je uznána za nežijící, hodnota je uložena a proběhne vyhledání klíče „místo úmrtí“. Pokud „datum úmrtí“ neexistuje, je ještě vyhledán klíč „datum a místo úmrtí“. Neexistuje-li ani ten, nebo neobsahuje hodnotu, je osoba prohlášena za žijící.

3. Pro úspěšnou identifikaci musí článek splnit alespoň dvě identifikační kritéria. Nebyloli tedy úspěšné vyhledání informační závorky nebo infoboxu, prohledá systém kate-

gorie, aby potvrdil, že článek pojednává o osobě. V opačném případě je identifikace již ukončena.

4. Nebylo-li úspěšné ani prohledání kategorií, jsou prohledána definiční slova. V případě, že neuspěly alespoň dvě identifikační metody, článek pravděpodobně nepojednává o osobě a funkce `identify` vrací `False`.

5. Systém ověří, zda se v první větě před názvem článku (jménem osoby) nevyskytují akademické tituly. Pokud ano, extrahuje je.

6. Systém se pokusí zjistit národnost osoby. Prohledá přídavná jména v definiční frázích. Pokud se některé nachází v seznamu národnostních adjektiv, první z nich velmi pravděpodobně určuje národnost osoby.

7. Je určeno pohlaví osoby. Pokud název článku končí koncovkou „ová“, je osoba automaticky vyhodnocena jako žena. Jinak je určen rod definičních slov, převládající rod pravděpodobně určuje pohlaví osoby. Slova jako „osoba“, „postava“ nebo „člověk“ jsou ignorována, stejně jako slova ve středním rodě. V případě rovnosti mužského a ženského rodu je extrakce neúspěšná.

8. Systém určí, zda je osoba reálná, nebo fiktivní. Osoba je označena jako fiktivní, pokud se mezi jejími definičními frázemi vyskytují předem dané řetězce.

9. Data jsou upravena do standardního formátu `RRRR-MM-DD`, z názvů míst jsou extrahovány odkazy, čímž je vytvořena vazba mezi osobou a sídlem.

10. Systém zapíše záznam do výstupního seznamu osob, báze znalostí, a pokud se zdařila extrakce všech atributů, i do patřičného vazebního souboru.

Pokud článek není identifikován jako osoba, systém přistoupí k pokusu o identifikaci článku jako místo. Identifikace a případná extrakce atributů probíhají takto:

1. Systém prohledá definiční slova. Každá shoda znamená jedno splněné kritérium.

2. Nedošlo-li zatím k úspěšné identifikaci, dojde k vyhledávání infoboxů.

3. Dále, nedošlo-li stále k identifikaci, hledají se další definiční slova před názvem článku (typické například v článcích o českých obcích). Pokud ani toto nedostačuje, dalším krokem je prohledání kategorií. Nejsou-li ani poté splněna alespoň dvě kritéria, funkce `identify` vrací `False`.

4. V první větě jsou vyhledány odkazy, pokud některé odkazují na stát (v seznamu států), první z nich pravděpodobně určuje stát, ve kterém se místo nachází. Pokud ne, systém vyhledává klíč „country“ nebo jeho lokalizovaný ekvivalent v infoboxu, je-li přítomen.

5. Systém zapíše záznam do výstupního seznamu míst a báze znalostí. Je-li mezi definičními slovy jedno ze slov identifikující město nebo obec, je proveden zápis i do vazebního seznamu sídel. Případně do seznamu vrcholů, pakliže jedno z definičních slov tomuto nasvědčuje.

Identifikace a extrakce atributů díla se odehrává následovně:

1. Jsou prohledána definiční slova, každá shoda znamená jedno splněné kritérium.
2. Vyhledávají se infoboxy. V nich jsou zároveň vyhledávány klíče nejčastěji nesoucí informace o autorovi díla a datu publikování v daném jazyce.
3. Pokud článek není stále identifikován jako dílo, dojde k prohledání kategorií. V případě, že ani tak nedojde k úspěchu, funkce `identify` vrací `False`.
4. Jména jsou extrahována jako odkazy, následkem čehož jsou vytvořeny vazby.
5. Systém zapíše záznam do výstupního seznamu děl, a pokud některé z definičních slov identifikuje album, píseň, malbu nebo knihu, i do báze znalostí. Byl-li určen autor, je vložen záznam i do vazebního seznamu umělců.

Organizace je identifikována a její atributy extrahovány takto:

1. Systém prohledá definiční slova, každá shoda znamená jedno splněné kritérium.
2. Jsou vyhledány infoboxy, které jsou následně prohledány na klíče nejčastěji nesoucí informace o datu založení, jménu zakladatele nebo zakladatelů a jménu aktuálního ředitele či předsedy.
3. Pokud nedošlo k úspěšné identifikaci, následuje prohledání kategorií. V případě, že ani tak článek není identifikován jako organizace, funkce `identify` vrací `False`.
4. Jména jsou extrahována jako odkazy. Tak se vytvoří vazby.
5. Systém zapíše záznam do výstupního seznamu organizací a báze znalostí. Zároveň, došlo-li k úspěšné extrakci jména zakladatele nebo předsedy, je vložen záznam do odpovídajícího vazebního seznamu.

Nebyl-li článek identifikován, probíhá takto jeho identifikace jako události (a případné určení jejího atributu):

1. Prohledávají se definiční slova, každá shoda znamená jedno splněné kritérium.
2. Jsou prohledány infoboxy, kde jsou zároveň vyhledávány klíče nejčastěji nesoucí informace o datu události.
3. Nedošlo-li dosud k úspěšné identifikaci, následuje prohledání kategorií. Pokud ani tak článek není identifikován, funkce `identify` vrací `False`. Článek nepojednává o žádném ze zkoumaných typů entit.
4. Systém zapíše záznam do výstupního seznamu událostí a báze znalostí.

Po skončení identifikace a extrakce systém zapíše vlastnosti článku do výstupního souboru a pokračuje zpracování dalšího článku. Po zpracování celého dumpu Wikipedie jsou zavřeny všechny výstupní soubory, vypočítána doba běhu a činnost programu končí.

3.5 Použití programu

Jak bylo již několikrát zmíněno, program implementován v rámci této práce je sada skriptů v jazyce Python 3. Pro svou činnost využívá nástroj MorphoDiTa, který je součástí repositáře The Python Package Index (PyPI) [7], oficiálního úložiště pro software třetích stran vyvíjený pro použití v Jazyce Python. Software z tohoto repositáře lze instalovat vícero nástroji, z nichž nejpoužívanějším a preferovaným je nástroj PIP. Ten lze v linuxových systémech ve verzi pro Python 3 nainstalovat jedním z následujících příkazů:

- `sudo apt install python3-pip` pro systémy z rodiny Debian.
- `sudo dnf install python3-pip` pro systémy z rodiny Red Hat.

Následně lze nainstalovat nástroj MorphoDiTa příkazem `pip3 install ufal.morphodita`.

Program je spustitelný z linuxového terminálu zavoláním hlavního skriptu main.py v kořenové složce práce. Spuštěním s parametrem `help` je možné vyvolat nápovědu. Aplikace má jeden povinný parametr `source`, jehož atribut specifikuje zdrojový soubor, ve standardním režimu dump české nebo slovenské Wikipedie. V případě přeskočení úvodní extrakční fáze (parametr `identify`) je zdrojovým souborem TSV reprezentace dumpu, při zadání parametru `convert` textový soubor obsahující jeden článek ve formě wikitextu. V testovacím režimu je jako zdroj očekáván předgenerovaný soubor s příponou `.ael`, v porovnávacím režimu hotová báze znalostí (soubor s příponou `.kb`).

Obecně je aplikaci tedy možné spustit ve čtyřech hlavních režimech:

- **Standardní režim** – Program načte články z dumpu Wikipedie, převede jej do standardizované podoby (kterou uloží do souboru ve formátu TSV) a provede identifikaci entit s extrakcí jejich atributů.
- **Extrakční režim** – Při použití parametru `extract`. Program načte články z dumpu Wikipedie a převede jej do standardizované podoby, kterou uloží do souboru ve formátu TSV. Identifikace entit se neprovádí.
- **Identifikační režim** – Při použití parametru `identify`. Program načítá články z předgenerované standardizované podoby Wikipedie ze souboru ve formátu TSV a provede identifikaci entit s extrakcí jejich atributů.
- **Konverzní režim** – Při použití parametru `convert`. Program načte jeden článek ve formě wikitextu z textového souboru a extrahuje z něj první odstavec ve formátu prostého textu.

Další režimy slouží ke spouštění testování nebo porovnávání:

- **Testovací režim** – Při použití parametru `test`, vyžaduje současné použití parametru `entity` ke specifikaci typu entity k otestování. Je otevřen textový soubor (v cestě zadané pomocí atributu) se seznamem entit identifikovaných jiným, referenčním nástrojem. Testovací skript tyto entity vyhledá mezi články a vypíše, které z nichy byly systémem úspěšně identifikovány jako zadaný typ entity. Jako parametr `source` je očekáván dříve exportovaný seznam zpracovaných článků (`.ael`).

- **Porovnávací režim** – Při použití parametru `compare`, vyžaduje současné použití parametru `entity` ke specifikaci typu entity k porovnání (je možné zadat jeden, společný název entity, nebo dva názvy oddělené lomítkem). Je otevřena jiná verze báze znalostí (v cestě zadané pomocí atributu). Porovnávací skript zjistí, které entity byly identifikovány v obou verzích a které pouze v jedné z nich.

Kapitola 4

Experimenty a vyhodnocení

Testování a vyhodnocování úspěšnosti implementovaného systému bylo prováděno především pomocí imaginárních článků a jejich experimentálními změnami, dále potom porovnáváním výsledné báze znalostí s referenčními výsledky vyprodukovanými jinými systémy.

4.1 Testování

Testování identifikace a extrakce prováděné programem probíhalo prostřednictvím uměle vytvořených dump souborů obsahujících jediný článek a změnou rozličných údajů v tomto článku. V pozdější fázi byla aplikace testována na celém dumpu slovenské a později i české Wikipedie. Tímto způsobem byl především ověřován počet identifikovaných entit a extrahovaných atributů.

Takto byly postupně odhalovány a odstraňovány chyby a nedostatky. Mezi nejčastější příčiny chybné identifikace patří:

- **Zavádějící kategorie** – Slova jako „město“ nebo „stát“ na začátku názvu kategorie implikují místo, slovo ale může mít v názvu úplně jiný význam či daná entita může být například osobou narozenou v určitém městě. Z tohoto důvodu bylo po řadě problémů rozhodnuto, že musí dojít ke shodě minimálně dvou kategorií, aby bylo toto kritérium považováno za splněné.
- **Chybná identifikace definičních slov** – Chybná extrakce definic může zapříčinit naprostou misinterpretaci článku. Viz kapitola o implementaci.
- **Jiný význam definičního slova** – Ačkoliv definiční slovo implikuje jistý typ entity, realita může být jiná (například slovo „obraz“ může kromě malby mít řadu dalších významů). Toto nebezpečí je z velké části eliminováno přidáním potřeby splnění alespoň dvou kritérií, aby mohl být článek prohlášen za entitu určitého typu.

Přestože se většina článků na Wikipedii drží jistého základního vzoru pro první odstavec, existuje řada takových, které tyto nepsané zásady wikipedistů porušují. Příkladem mohou být již zmíněné články o českých obcích, jejichž první věty obsahují definiční slovo „obec“ na samém začátku a tvar slovesa „být“ se zde vůbec nevyskytuje (například „Obec Nový Malín

se nachází v okrese Šumperk v Olomouckém kraji.“). Dále mohou být úvodní věty formulovány zcela atypicky, například „Jako mezikruží označujeme geometrický útvar rozlišený v rovině dvěma soustřednými kružnicemi.“ nebo „Pythagorova věta popisuje vztah, který platí mezi délkami stran pravoúhlých trojúhelníků v euklidovské rovině.“, což zcela znemožňuje extrakci definičních frází a slov. Většina takových článků ale nepojednává o žádné z entit, jimiž se tato práce zabývá, a proto toto nemá žádné dalekosáhlejší důsledky.

K testování úspěšnosti identifikace byly implementovány dva skripty. První z nich, testovací, vyžaduje pro svůj chod prostý seznam entit identifikovaných referenčním nástrojem. Předpokládá se, že proběhla standardní činnost programu a je k dispozici seznam článků (soubor s příponou `.ael`). Skript prochází zpracované články z české nebo slovenské Wikipedie a každý z nich vyhledá v testovacím seznamu. Pokud se zde vyskytuje, zkontroluje, zda byl článek systémem identifikován jako testovaný typ entity. Pokud ano, je zapsán do souboru `matches.txt`, v opačném případě do souboru `errors.txt`.

Úspěšnost lze také sledovat pomocí porovnáváním buď jednotlivých verzí znalostní báze nebo srovnáním s jinou bází podobné struktury. K tomu slouží porovnávací skript. Otevře obě báze (předpokládá se opět, že proběhla činnost programu a báze je vygenerovaná), vyjme z nich entity určitého typu a pokusí se je najít v bázi druhé. Následně naopak. Výsledkem jsou tři soubory: `matches.txt` s entitami nacházejícími se v obou bázích, `only_old.txt` s entitami, které jsou pouze ve starší verzi, a `only_new.txt`, kde najdeme nově identifikované entity.

Program byl po dlouhou dobu postupně optimalizován, aby jeho činnost trvala co nejkratší dobu. Systém po každé znatelné změně znovu zpracovával celý dump soubor, čímž nejen exportoval výsledky k porovnání, ale zároveň vypsal dobu běhu. Nakonec se časová náročnost zpracování české Wikipedie dostala na úroveň, na níž se před optimalizací nacházela verze slovenská.

4.2 Vyhodnocení výstupních dat

V dumpu slovenské Wikipedie ze dne 1. dubna 2019 program identifikuje:

- **25 546 osob** – Z toho 20 901 mužů a 3 655 žen, u 990 osob se nepodařilo určit pohlaví. U 8 489 živých a 7 710 zemřelých osob byly úspěšně extrahovány všechny atributy, v případě 63 % osob se tudíž zdařila kompletní extrakce (toto číslo ovšem plně nevystihuje úspěšnost extrakce atributů, neboť velké množství článků o osobách některé z údajů vůbec neobsahuje). 37 osob bylo určeno jako fiktivní postavy.
- **56 762 míst** – Z toho 55 578 měst a obcí, přičemž u 55 262 z nich byl určen stát, ve kterém se nacházejí. Vrcholů bylo identifikováno 23.
- **1 190 organizací**
- **2 879 událostí**
- **3 246 děl** – Z toho 903 písní a skladeb, 1 530 alb, 87 obrazů a 249 knih. U 2 705 děl byl určen jejich autor.

V dumpu české Wikipedie ze stejného data bylo identifikováno:

- **110 070 osob** – To je výrazně více, než v případě slovenské Wikipedie (česká verze obsahuje cca o 75 % článků více, ale o více než 330 % větší množství z nich jsou osoby). Je mezi nimi 81 634 mužů a 18 227 žen, u 10 209 osob se určit pohlaví nepodařilo. U 29 089 živých a 16 197 zemřelých osob byly úspěšně extrahovány všechny atributy, v případě 41 % osob se tudíž zdařila kompletní extrakce (toto číslo ale rovněž nevystihuje úspěšnost extrakce atributů, jelikož velké množství článků o osobách, při jejich množství na české Wikipedii zvláště, některé z údajů vůbec neobsahuje). Jako fiktivní postavy systém určil 310 osob.
- **37 598 míst** – Česká verze naopak obsahuje nižší množství článků o místech, než slovenská. A to nejen relativně, ale dokonce absolutně. Z tohoto počtu je 30 363 článků o městech a obcích, přičemž u 55 262 z nich byl určen stát, ve kterém se nacházejí. Vrcholů bylo identifikováno 1 977.
- **8 166 organizací**
- **7 105 událostí**
- **11 588 děl** – Z toho 1 410 písní a skladeb, 6 402 alb, 146 obrazů a 1 685 knih. Autor byl určen u 8 091 děl.

Pomocí testovacího skriptu byla otestována úspěšnost identifikace osob. 98 % článků na české Wikipedii, které se vyskytují také v referenčním seznamu, bylo správně určeno jako osoby. V případě slovenské verze se tímto způsobem odhadnutá úspěšnost pohybuje kolem 96 %.

Porovnávací skript výslednou znalostní bázi z výstupu programu srovnal s bází vygenerovanou nástrojem vyvíjeným v rámci KNOT v minulosti. Tato báze znalostí byla za tímto účelem poskytnuta vedoucím práce. Pro ilustraci byla srovnána identifikace měst a obcí. Zatímco nový systém identifikoval (jak již bylo zmíněno) 30 363 sídel, ve starší bázi znalostí se nachází sídel 36 221. Nový systém tedy identifikoval 83 % z počtu původních entit. Avšak pouze 51 % sídel je shodných. Novému systému se nepodařilo identifikovat celkem 13 266 entit typu sídlo, které identifikoval starý. Je ovšem nutné zmínit, že řada těchto entit nepatří mezi města a obce (původní systém chybně identifikoval jako sídlo například některé ostrovy nebo vojenské újezdy). 8 459 sídel naopak nový systém identifikoval oproti staršímu navíc.

Obdobná situace nastává při porovnání identifikace fiktivních osob. Starý systém identifikoval 1 424 imaginárních postav, velké množství z nich je ale určeno naprosto chybně (entity jako Sebevražedný oddíl, Přemysl Otakar II. nebo Saladin). Nový systém má výrazně menší chybovost, zároveň ale nalezne pouze 310 fiktivních postav. 114 z nich ovšem ve starší znalostní bázi chybí.

Ve standardním režimu je doba běhu programu 6–7 minut pro slovenskou verzi a 15–16 minut pro verzi českou. Do tohoto času není započítáno porovnávání a testování. Systém má navíc velice nízkou paměťovou náročnost, protože díky postupnému načítání vstupních souborů nejsou nikdy v paměti uloženy velké objemy dat.

Kapitola 5

Závěr

Cílem této práce byla implementace systému schopného extrakce informací z české a slovenské Wikipedie a jejich uložení ve formě strukturované báze znalostí. Byly popsány základní techniky extrakce informací z textu, struktura článků na Wikipedii a způsob identifikace typů entit v těchto článcích. V rámci projektu byly také vyvinuty algoritmy pro extrakci prvních odstavců a vět v prostém textu z článků psaných ve značkovacím jazyce vyvinutém nadací MediaWiki, jež Wikipedii provozuje. Rovněž byly sepsány seznam sloveských zkratek a seznamy slovenských a českých názvů států a národnostních adjektiv.

Systém převádí wikitext na prostý text s vysokou úspěšností, poměr korektně identifikovaných článků o osobách je více než 95 %. U cca poloviny z nich jsou navíc úspěšně extrahovány všechny jejich atributy. Pro účely ověření úspěšnosti konkrétních činností je program vybaven možností testování. Je také možné spustit pouze jednu ze dvou dílčích částí systému, není-li druhé z jakéhokoli důvodu třeba.

Kód programu přímo neobsahuje žádné jazykově závislé řetězce, všechny lokalizované texty jsou uloženy v externích souborech. Celý systém je tak jednoduše rozšiřitelný, je tedy možné na zde popisovanou činnost navázat. Jako ideální cíle se jeví přidání nových typů entit, přidání nových atributů k extrakci a implementace dalších jazyků, především anglického. V takovém případě by ovšem bylo nutné navrhnout zcela odlišný algoritmus pro extrakci definičních slov anglických článků.

Literatura

[1] Aprosio, A. P.; Giuliano, C.: *Airpedia*. 2019, [Online; navštíveno 14. 05. 2019].
URL http://www.airpedia.org/

[2] Chernykh, N.: *Analýza textu (text mining) pomocí vybraného softwaru*. Bakalářská práce, Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, Praha, 2012, vedoucí práce Ing. Stanislava Hrušková, Ph.D.

[3] DBpedia Association: *DBpedia*. 2019, [Online; navštíveno 28. 04. 2019].
URL https://wiki.dbpedia.org

[4] KNOT@FIT: *Decipher wikipedia — KNOT Wiki*. 2018, [Online; navštíveno 03. 05. 2019].
URL https://knot.fit.vutbr.cz/wiki/index.php?title=Decipher_wikipedia&oldid=35260

[5] KNOT@FIT: *Entity kb czech3 — KNOT Wiki*. 2019, [Online; navštíveno 14. 05. 2019].
URL https://knot.fit.vutbr.cz/wiki/index.php/Entity_kb_czech3

[6] Manning, C. D.; Schütze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press, 1999, ISBN 0-262-13360-1.

[7] Python Software Foundation: *PyPI — The Python Package Index*. 2019, [Online; navštíveno 11. 05. 2019].
URL https://pypi.org

[8] Python Software Foundation: *Python documentation*. 2019, [Online; navštíveno 03. 05. 2019].
URL https://docs.python.org

[9] Přispěvatelé Wikipedie: *Kategorie:Šablony:Infoboxy — Wikipedie: Otevřená encyklopedie*. 2018, [Online; navštíveno 14. 05. 2019].
URL https://cs.wikipedia.org/w/index.php?title=Kategorie:%C5%A0ablony:Infoboxy&oldid=16254672

[10] Přispěvatelé Wikipedie: *TSV (datový formát) — Wikipedie: Otevřená encyklopedie*. 2018, [Online; navštíveno 14. 05. 2019].
URL https://cs.wikipedia.org/w/index.php?title=TSV_(datov%C3%BD_form%C3%A1t)&oldid=16420783

[11] Přispěvatelé Wikipedie: *Nápověda:Infoboxy — Wikipedie: Otevřená encyklopedie*. 2019, [Online; navštíveno 14. 05. 2019].
URL https://cs.wikipedia.org/w/index.php?title=N%C3%A1pov%C4%9Bda:Infoboxy&oldid=16939957

[12] Přispěvatelé Wikipedie: *Wikipedie — Wikipedie: Otevřená encyklopedie*. 2019, [Online; navštíveno 16. 04. 2019].
URL https://cs.wikipedia.org/w/index.php?title=Wikipedie&oldid=17067916

[13] Rusiňák, P.: *Určování typů entit na základě extrakce informací z Wikipedie*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, Brno, 2018, vedoucí práce Doc. RNDr. Pavel Smrž, Ph.D.

[14] Sedláček, P.: *Text mining a jeho možnosti (aplikace)*. 2003, [Online; navštíveno 12. 05. 2019].
URL https://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

[15] Sedláček, P.: *Extrakce informací z textu*. Diplomová práce, Masarykova univerzita, Fakulta informatiky, Brno, 2006, vedoucí práce doc. RNDr. Lubomír Popelínský, Ph.D.

[16] Sedláček, R.: *Morfologický analyzátor češtiny*. Diplomová práce, Masarykova univerzita, Fakulta informatiky, Brno, 1999.

[17] Straka, M.; Straková, J.: *MorphoDiTa*. 2019, [Online; navštíveno 01. 05. 2019].
URL http://ufal.mff.cuni.cz/morphodita

[18] Wikimedia foundation: *Markup spec — MediaWiki, The Free Wiki Engine*. 2019, [Online; navštíveno 27. 04. 2019].
URL
https://www.mediawiki.org/w/index.php?title=Markup_spec&oldid=3130201

[19] Wikimedia foundation: *Wikimedia Downloads*. 2019, [Online; navštíveno 27. 04. 2019].
URL https://dumps.wikimedia.org

[20] Wikipedia contributors: *Wikipedia:Namespace — Wikipedia, The Free Encyclopedia*. 2019, [Online; navštíveno 27. 04. 2019].
URL https://en.wikipedia.org/w/index.php?title=Wikipedia:Namespace&oldid=890934535

[21] World Wide Web Consortium: *RDF — Semantic Web Standards*. 2014, [Online; navštíveno 14. 05. 2019].
URL https://www.w3.org/RDF/

[22] Ševčíková, M.; Žabokrtský, Z.; Krůza, O.: *Zpracování pojmenovaných entit v českých textech*. Technická zpráva, Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Praha, 2007.

Příloha A

Příklad výstupního seznamu

Antoni Grabowski	male	polish	1857-06-11	Nowe Dobre	1921-06-04	Varšava	
Julio Baghy	male	hungarian	1891-01-13	Segedín	1967-03-18	Budapešť	
Nikolaj Vladimirovič Někrasov	male	russian	1900-12-18	Moskva	1938-10-04	popraven	
Eugen Michalski	male	russian	1897-01-21	Letyčiv	1937-10-14	Kyjev	
Ivo Lapenna	female	croatian	1909-11-05	Split	1987-12-15	Kodaň	
František Lorenc	male	czech	1872-12-24	Zbyslav (Vrdy)	1957-05-24	Porto Alegre	
Stanislav Kamarýt	male	czech	1883-11-10	Velešín	1956-06-30	Havlíčkův Brod	
Miloš Lukáš	male	czech	1897-09-15	Jilemnice	1976-04-26	Hradec Králové	
Albert Škarvan	male	slovakian	1869-01-31	Tvrdošín	1926-03-29	Liptovský Hrádok	
Martin Benka	male	slovakian	1888-09-21	Kostolište	1971-06-28	Malacky	
Eduard Vladimír Tvarožek	male	slovakian	1920-05-13	Horný Ďur	1999-08-09	Trenčín	
Jaroslav Foglar	male	czech	1907-06-06	Praha	1999-01-23	Praha	
Arthur Troop	male	english	1914-12-15	Lincoln	2000-11-30	Peterborough	
Aloys Skoumal	male	czech	1904-06-19	Pačlavice	1988-06-04	Praha	
Nelson Mandela	male	south african	1918-07-18	Mvezo	2013-12-05	Johannesburg	
Gregory Peck	male	american	1916-04-05	La Jolla	2003-06-12	Los Angeles	
Denis Diderot	male	french	1713-10-05	Langres	1784-06-31	Paříž	
Věra Chytilová	female	czech	1929-02-02	Ostrava	2014-03-12	Praha	
Miloš Kopecký	male	czech	1922-08-22	Praha	1996-02-16	Praha	
Jiří Šust	male	czech	1919-08-29	Praha	1995-04-30	Praha	
François Truffaut	male	french	1932-02-06	Paříž	1984-10-21	Neuilly-sur-Seine	
Claude Jade	female	french	1948-10-08	Dijon	2006-12-01	Boulogne-Billancourt	
Karel Kachyňa	male	czech	1924-05-01	Vyškov	2004-03-12	Říčany	
James Watt	male	scottish	1736-01-19	Greenock	1819-08-25	Handsworth	
Charles Perrault	male	french	1628-01-12	Paříž	1703-05-16	Paříž	
Jana Rybářová	female	czech	1936-03-31	Praha	1957-02-11	Praha	
Pavel Juráček	male	czech	1935-08-02	Příbram	1989-05-20	Praha	
Jaroslav Ježek	male	czech	1906-09-25	Žižkov	1942-01-01	New York	
Edgar Allan Poe	male	american	1809-01-19	Boston	1849-10-07	Baltimore	
František Josef Čečetka	male	czech	1871-04-23	v Nových Hradech u Vysokého Mýta	1942-06-03	Praha	
Alan Marshall	male	australian	1902-05-02	Noorat	1984-01-21	East Brighton	
Zdeněk Liška	male	czech	1922-03-16	Smečno	1983-06-13	Praha	
Pio z Pietrelciny	male	italian	1887-05-25	Pietrelcina	1968-09-23	San Giovanni Rotondo	
Vladimír Bejval	male	czech	1942-12-31	Praha	2011-09-19	Praha	
Petr Skoumal	male	czech	1938-03-07	Praha	2014-09-28	Praha	
Jarmila Fastrová	female	czech	1899-06-01	Praha	1968-11-28	Praha	
Václav Bahník	male	czech	1918-04-13	Hrdlořezy u Mladé Boleslavi	2003-03-03	Liberec	
Mirko Eliáš	male	czech	1899-05-13	v Chocni	1938-05-29	v Chvojenci u Holic	
Karel Kořistka	male	czech	1825-02-07	Březová nad Svitavou	1906-01-18	Praha	
Lubor Tokoš	male	czech	1923-02-07	Šternberk	2003-09-29	Zlín	
Gabriel Blažek	male	czech	1842-09-20	Borovnice u Chocně	1910-12-06	Praha	
František Josef Studnička	male	czech	1836-06-27	v Janově u Soběslavi	1903-02-21	v Praze	
Erik Adolf Saudek	male	czech	1904-10-18	Vídeň	1963-06-16	Sozopol	
František Josef Gerstner	male	czech	1756-02-22	Chomutov	1832-06-25	Mladějov	
Zdeněk Mézl	male	czech	1934-10-31	Praha	2016-05-23	Praha	
Robert Hliněnský	male	czech	1908-11-03	v Brně	1979-01-08	v Brně	
Bedřich Hrozný	male	czech	1879-05-06	Lysá nad Labem	1952-12-12	Praha	
Franz Anton Leonard Herget	male	czech	1741-11-06	Andělská Hora	1800-10-01	Praha	
Václav Hlavatý	male	czech	1894-01-27	Louny	1969-01-11	Bloomington	
František Kadeřávek	male	czech	1885-06-26	Praha	1961-02-09	Praha	
Johannes Kepler	male	german	1571-12-27	Weil der Stadt	1630-11-15	Řezno	

Obrázek A.1: Část vazebního seznamu zemřelých osob z české Wikipedie.

Příloha B

Příklady pomocných souborů

```
česká obec
důl
geografický region
hora
chráněné území
jeskyně
kontinent
les
letiště
mikroregion
moře
obora
ostrov
park
průplav
průsmyk
region
rozhledna
sídlo
sídlo světa
souostroví
statutární město
stát
ulice
vodní plocha
vodní tok
vodopád
zaniklý stát
železniční stanice
```

Obrázek B.1: Seznam infoboxů implikujících místo.

```
datum[ _]narození
místo[ _]narození
datum a místo narození
datum[ _]úmrtí
místo[ _]úmrtí
datum a místo úmrtí
fiktivní|imaginární|mytologick(ý|á)|mýtick(ý|
á)|bájn(ý|á)|(literární|filmová|seriálová)
postava|postava .*(románu|povídky|filmu|
seriálu|mytologie)|bůh|božstvo
```

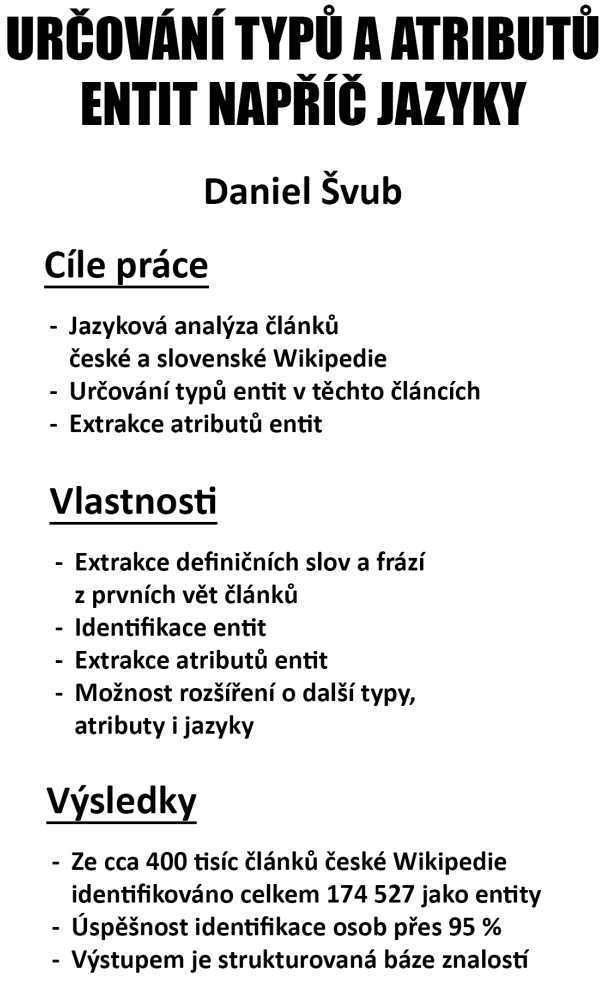
Obrázek B.2: Seznam lokalizovaných řetězců pro extrakci osob.

```
afghan
albanian
algerian
american
andorran
angolan
antiguan
argentinean
armenian
australian
austrian
azerbaijani
bahamian
bahraini
bangladeshi
barbadian
batswana
belarusian
belgian
belizean
bhutanese
bolivian
```

Obrázek B.3: část anglického seznamu národnostních adjektiv.

Příloha C

Plakát



Obrázek C.1: Plakát prezentující práci, její cíle a výsledky.

Příloha D

Obsah přiloženého média

- **Adresář latex** – Zdrojové soubory této technické zprávy
- **Adresář results** – Výstupy systému pro dump soubory české a slovenské Wikipedie z 1. dubna 2019.
- **Adresář source** – Zdrojové soubory systému.
- `README` – Informace o obsahu média.
- `xsvubd00-Urcovani-typu-a-atributu-entit.pdf` – Elektronická verze této technické zprávy.