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 ENTIT NA ZÁKLADĚ EXTRAKCE INFORMACÍ Z WIKIPEDIE

IDENTIFYING ENTITY TYPES BASED ON INFORMATION EXTRACTION FROM WIKIPEDIA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **PETR RUSIŇÁK**
AUTHOR

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

BRNO 2018

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2017/2018

Zadání bakalářské práce

Řešitel: **Rusiňák Petr**
Obor: Informační technologie
Téma: **Určování typů entit na základě extrakce informací z Wikipedie**
Identifying Entity Types Based on Information Extraction from Wikipedia
Kategorie: Databáze

Pokyny:

1. Prostudujte formáty zápisu jednotlivých částí článků Wikipedie - informačních rámečků, kategorií, odkazů na ekvivalentní stránky v jiných jazycích, definičních odstavců, seznamů a zjednoznačňovacích stránek.
2. Seznamte se s metodami převodu nestrukturovaných textových informací do strukturované podoby a s relevantními projekty souvisejícími bezprostředně s Wikipedií, např. DBpedia, AIRpedia apod.
3. Navrhněte a implementujte systém pro automatickou extrakci typů entit na základě kombinace výše uvedených částí článků Wikipedie.
4. Vyhodnoťte výsledky systému na reprezentativním vzorku dat.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle domluvy s vedoucím

Pro udělení zápočtu za první semestr je požadováno:

- Funkční prototyp řešení

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2017
Datum odevzdání: 16. května 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
L.S.612 66 Brno, Božetěchova 2

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Cílem této práce je identifikovat typy článků na Wikipedii (např. rozlišit články o osobách od článků o sportovních utkáních), přičemž tento systém by měl být použitelný pro libovolný typ extrahované entity. Vstupem pro tento systém je seznam několika vzorových článků patřících do hledané entity a seznam několika článků nepatřících do této entity. Na základě těchto seznamů budou vygenerovány příznaky, které lze použít k nalezení všech článků patřící do této entity. Tyto příznaky mohou být založeny jak na základě strukturovaných informací na Wikipedii (např. šablony, kategorie), tak i na základě analýzy přirozeného textu v první větě článku, kde bude nalezeno definiční podstatné jméno vystihující entitu daného článku. Tento systém podporuje extrakci česky a anglicky psaných článků a je rozšířitelný pro podporu dalších jazyků.

Abstract

This paper presents a system for identifying entity types of articles on Wikipedia (e.g. people or sports events) that can be used for identifaction of any arbitrary entity. The input files for this system are a list of several pages that belong to this entity and a list of several pages that do not belong to this entity. These lists will be used to generate features that can be used for generation of the list of all pages belonging to this entity. The fatures can be based on both structured information on Wikipedia such as templates and categories and non-structured informations found by the analysis of natural text in the first sentence of the article where a defining noun that represents what the article is about will be found. This system support pages written in Czech and English and can be extended to support other languages.

Klíčová slova

Wikipedie, určování typů entit, asociační analýza

Keywords

Wikipedia, identifying entity types, association analysis

Citace

RUSIŇÁK, Petr. *Určování typů entit na základě extrakce informací z Wikipedie*. Brno, 2018. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Doc. RNDr. Pavel Smrž, Ph.D.

Určování typů entit na základě extrakce informací z Wikipedie

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. Smrže. Další informace mi poskytl pan Otrusina, který mě vedl v předmětech Projektová praxe 1 a Projektová praxe 2. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Petr Rusiňák
16. května 2018

Poděkování

Rád bych poděkoval svému vedoucímu Pavelu Smržovi, doc. RNDr., Ph.D. za poskytnuté informace a rady při tvorbě této práce. Také bych rád poděkoval ing. Lubomíru Otrusinovi, který mě vedl v předcházejících předmětech Projektová praxe 1 a 2 a taktéž mi poskytl velmi cenné rady a informace.

Obsah

Kapitola 1

Úvod

V současnosti je Wikipedie jednou z největších internetových encyklopedií. Aktuálně má anglická verze této encyklopedie cca 137 000 aktivních editorů, kteří za poslední měsíc editovali alespoň jeden článek a obsahuje přes 5,6 milionů článků. [25] Česká verze této encyklopedie v současnosti má 2 100 aktivních editorů a obsahuje cca 400 000 článků. [29] Celkově je doména wikipedia.org pátou nejnavštěvovanější stránkou na internetu. [2]

Články této encyklopedie jsou volně ke stažení, a to buď jednotlivě pomocí aplikačního programového rozhraní (API), nebo všechny najednou ve formátu XML. Možnosti, jak lze tyto články stáhnout, budou popsány v kapitole 2.2. U článků je však kromě některých metadat, jako je čas poslední editace, uložen pouze jejich text ve **speciálním formátovacím jazyce MediaWiki**, který bude podrobněji popsán v kapitole 2.3. Wikipedie nenabízí žádné rozhraní, pomocí kterého by bylo možné přistupovat ke konkrétním informacím uloženým v daném článku, a to ani v případě, kdy se při běžném přístupu pomocí webového prohlížeče zdá, že obsah článku je uložen ve strukturované podobě (například seznam epizod seriálu Simpsonovi v článku List of The Simpsons episodes). Některé formátovacích značky[1] MediaWiki jsou však dostatečně strukturovány na to, aby z nich bylo možné extrahovat konkrétní informace[2]. [16] [30]

Aby však bylo možné z nějakého článku spolehlivě strojově extrahovat určité informace, je nejprve potřeba přesně vědět, o čem daný článek je. Jedná se o sportovní událost, člověka, město, řeku, film, druh měny, psí rasu, nebo počítačovou hru? Nebo o něco úplně jiného? U člověka asi má smysl extrahovat jeho datum narození, ale snaha o vyhledání data narození u města by nejspíše moc logická nebyla. Stejně tak by to byl nesmysl i u filmu – a to i v případě, že by některý z extrakčních nástrojů skutečně nějaké datum našel: nejspíše by v tomto případě bylo nalezeno datum napsání scénáře, datum natočení, datum uvedení do kin či televizního vysílání nebo podobné datum a omylem by bylo považováno za datum narození, ale o datum narození by se zcela jistě nejednalo. Ten ani nemělo smysl u tohoto článku chtít hledat.

Tato práce se zabývá problematikou **určení typů entit článků na Wikipedii**, tedy jednoznačným určením, které články pojednávají o jedné skupině věcí (např. o řekách) a které nikoliv. Wikipedie sice obsahuje **mechanismus kategorií**, nicméně jak si podrobně ukážeme v kapitole 5.2, možnosti jeho využití k tomuto účelu jsou velmi omezené.

[1] Především šablony, infoboxy a kategorie (budou popsány v kapitole 2.3)

[2] Například data narození a úmrtní u osob, počet odvysílaných epizod u seriálů, rok natočení u filmů, zemi původu u psích ras, výrobce u modelů aut, stát, kde se dané město nachází, u měst, počet obyvatel u měst, krajů, států a kontinentů apod.

K identifikaci entit článků budou využity příznaky [12], což jsou ověřitelná tvrzení o daném článku (např. test na přítomnost šablony či kategorie), které budou popsány v kapitole 5. Z těchto příznaků budou vytvořeny asociační pravidla [1] ve tvaru *příznak* → *výsledek identifikace* postupem, který bude uveden v kapitole 7.4, kde *výsledek identifikace* je buď skutečnost, že daný článek do hledané entity patří, nebo skutečnost, že daný článek do této entity nepatří. Na základě asociačních pravidel pak může být vygenerován seznam článků, které patří do hledané entity a seznam článků, které do této entity nepatří. Tyto seznamy jsou výstupními soubory tohoto systému.

Cílem tohoto systému je nalézt **bázi znalostí** obsahující informace, které články do hledané entity patří a které nikoliv, přičemž je tento systém **znovupoužitelný pro různé typy entit** a může být rozšířen i podporu jiných jazyků. Úspěšnost tohoto systému bude vyhodnocena v kapitole 8 výběrem náhodných stránek ze seznamů zařazených a nezařazených stránek do hledané entity a jejich manuálním ověřením.

Kapitola 2

Struktura dat na Wikipedii

Tato kapitola se zabývá tím, jakým způsobem jsou data na Wikipedii uložena a jak je lze stáhnout k pozdějšímu off-line použití. Nejprve budou definovány základní pojmy *stránka* a *článek*. Poté bude popsáno, jakým způsobem lze načíst zdrojový kód jednotlivé stránky, jaké další údaje lze o jednotlivých stránkách získat a jak si stáhnout všechny stránky Wikipedie najednou. V poslední podkapitole 2.3 bude vysvětlen formátovací jazyk, který se na Wikipedii používá k zapsání zdrojového kódu jednotlivých stránek.

2.1 Základní pojmy

Základními prvky, které Wikipedie používá pro uložení informací, jsou stránka a článek. Význam pojmu **stránka** je stejný jako u jiných webových serverů a představuje nějaký vzdálený objekt, ke kterému lze přistoupit pomocí jedinečné URL [3] adresy. Příklady stránek na Wikipedii jsou Richard Moser, Category:People from New York City či User:Steve2011.

Článek je taková stránka, která obsahuje encyklopedické informace [27], tj. jedná se o stránky s užitečnými informacemi, kvůli nimž většina návštěvníků Wikipedii navštěvuje (např. Richard Moser nebo Inamuragasaki). Mimo článků se na Wikipedii nacházejí i jiné typy stránek, jako jsou stránky uživatelských účtů, stránky kategorií, stránky s multimediálními přílohami článků, stránky šablon, stránky s nápovědnou pro editory Wikipedie a další speciální stránky. Z formálního hlediska jsou články takové stránky, které jsou zařazeny v jmenném prostoru `0`. V této práci budeme určovat typy entit pouze u článků, ostatní stránky se při identifikaci budou ignorovat.

Nicméně také stojí za zmínku uvést, že tzv. **rozcestníky** (anglicky disambiguation pages; příklad rozcestníku: FBI (disambiguation)) **nejsou zařazeny do samostatného jmenného prostoru**, a tedy se jedná o články. Stejně tak do jmenného prostoru `0` patří **přesměrování** na jinou stránku, a tedy jsou považovány za články. Tyto typy článků budou rovněž přeskakovány, jelikož neobsahují žádné užitečné informace.

2.2 Přístup ke stránkám na Wikipedii

Základním způsobem, jak mohou čtenáři přistupovat ke článkům Wikipedie, je s použitím webového prohlížeče. Tímto způsobem lze zobrazit vykreslenou stránku v grafické podobě, zobrazit zdrojový kód článku ve formátu HTML a po kliknutí na odkaz *Zobrazit zdroj* nebo *Editovat zdroj* v horní části každé stránky i zobrazit zdrojový kód článku ve formátu MediaWiki.

Tento přístup však není vhodný ke strojovému načítání obsahu článků: zdrojový kód v HTML obsahuje velké množství režijních informacích, jejichž struktura (například názvy CSS tříd) se může kdykoliv změnit. Proto je lepší vždy pracovat s původním zdrojovým formátem MediaWiki, který je přesně definován a bude popsán v podkapitole 2.3. Stejně tak je dlouhodobě nespolehlivé i strojové získávání zdrojového kódu ve formátu MediaWiki pomocí speciální stránky přístupné pod odkazem *Zobrazit zdroj* umístěného nad nadpisem každého článku, který je obalen přesně nedefinovanou HTML strukturou. Kromě tohoto by hromadné stahování článků tímto způsobem bylo negativně vnímáno ze strany Wikipedie.[1]

Wikipedie však nabízí i způsoby, jak k jejím stránkám přistupovat automatizovaně. K tomuto lze využít buď **aplikační programové rozhraní** (API), které nabízí on-line přístup k aktuálním revizím jednotlivých článků, nebo tzv. **dump soubor**, který obsahuje všechny stránky Wikipedie ve formátu XML pro jejichž pozdější zpracování off-line. Tyto způsoby budou popsány v následujících podkapitolách.

2.2.1 Aplikační programové rozhraní Wikipedie

Aplikační rozhraní Wikipedie umožňuje přístup k jednotlivým stránkám Wikipedie a obsahuje jejich aktuální podobu. Toto rozhraní je zřízeno pro každou jazykovou mutaci encyklopedie zvlášť, API anglické mutace lze nalézt na adrese https://en.wikipedia.org/w/api.php (u ostatních jazyků se tato adresa liší v doménovém jméně). Toto rozhraní je poskytováno systémem MediaWiki [17], na kterém je tato encyklopedie provozována a kromě čtení dat umožňuje i operace zápisu, jako jsou editace textu článků, hromadné mazání článků a podobně. Nicméně režimy umožňující provádění změn nebudou v rámci této práce nijak využity. Za účelem identifikace článků s využitím API by bylo možné použít režim `query` [18], který vypíše obsah zadané stránky. Například pro vypsání článku anglické Wikipedie s názvem *Dog* lze na server dané jazykové mutace Wikipedie (v tomto případě https://en.wikipedia.org/) odeslat následující dotaz pomocí metody HTTP GET:

```
/w/api.php?action=query&format=json&prop=revisions
        &titles=Dog&rvprop=content
```

Odpovědí na tento dotaz je následující zpráva (zkráceno):

```
{
    "batchcomplete":"",
    "query":{"pages":{"4269567":{
        "pageid":4269567,
        "ns":0,
        "title":"Dog",
        "revisions":[{
            "contentformat":"text/x-wiki",
            "contentmodel":"wikitext",
            "*":"<...> The '''domestic dog''' (''Canis lupus
                familiaris'' or ''Canis familiaris'') <...>"
        }]
    }}}
}
```

[1] Viz také: Why not just retrieve data from wikipedia.org at runtime? na nápovědě Wikipedie

Jak je z výše uvedené ukázky patrné, pomocí aplikačního rozhraní lze o jedné stránce vyčíst hned několik údajů. Mezi ně patří **číselný identifikátor** stránky *(v tomto případě 4269567, lze použít při další práci s API, např. k editaci stránky, pro náš účel toto číslo nebude mít žádné využití)*, **jmenný prostor** stránky (jak již bylo uvedeno v kapitole 2.1, číslo nula představuje typ stránky *článek*) a nijak nezpracovaný **zdrojový text** článku ve formátu používaném MediaWiki nepřevedeném do jazyka HTML (bude popsán v následující podkapitole 2.3). Přidáním parametru `&rvprop=timestamp|user|comment` k původnímu dotazu by bylo navíc možné zjistit datum poslední editace tohoto článku, uživatelské jméno editora Wikipedie, který poslední změnu stránky provedl a komentář, který tento uživatel při odesílání revize článku uvedl. S použitím dalších parametrů lze dále získat předchozí revize či kompletní historii změn daného článku, ovšem tyto údaje nejsou pro účely identifikace entit článků relevantní.

Možnosti aplikačního rozhraní Wikipedie byly v této práci uvedeny především kvůli tomu, aby bylo názorně ukázáno, jakým způsobem jsou články (a ostatní stránky) na Wikipedii uloženy a aby byl čtenáři ukázán způsob, jak si i bez stahování dump souboru celé Wikipedie vyhledat veškerá data uložená o libovolném článku. Nicméně při zpracování většího počtu článků má tento přístup zásadní nevýhodu, a tou je nutnost být po celou dobu zpracování připojen k internetu a na každý článek se dotazovat serveru encyklopedie. I když ze strany Wikipedie není stanoven žádný pevný limit počtu požadavků za jednotku času, pokud nedochází paralelně k vícenásobným přístupům a API umožňuje přístup k více článkům v rámci jednoho dotazu, nejspíše by tento přístup nebyl vnímán pozitivně a jelikož Wikipedie obsahuje miliony článků, byl by časově neefektivní ve srovnání s použitím dump souboru, který bude popsán v následující kapitole.

2.2.2 XML soubor s daty Wikipedie (tzv. dump soubor)

Alternativou k on-line přístupu pomocí API je stažení tzv. **dump souboru**. Dump soubor je soubor, který je cca jednou měsíčně vygenerován Wikipedií a obsahuje všechna data z této encyklopedie v jediném XML souboru. Podobně jako v případě API je tento soubor generován pro každou jazykovou mutaci Wikipedie zvlášť. Vzhledem k tomu, že tyto soubory jsou relativně rozsáhlé a ve většině případů není nutné pracovat se všemi daty, které jsou na Wikipedii uloženy (např. s minulými revizemi článků nebo s binárními daty, jako jsou obrázky), nabízí Wikipedie více variant tohoto souboru s různým rozsahem uložených informací [24] (velikosti souborů uvedené v závorkách jsou orientační a představují aktuální velikost daného souboru bez komprese pro anglickou Wikipedii):

- `pages-articles.xml` – obsahuje pouze aktuální revizi každé stránky (články i ostatní typy stránek), neobsahuje stránky uživatelských účtů ani tzv. diskuzní stránky (58 GB).
- `pages-meta-current.xml` – obsahuje pouze aktuální revize všech stránek včetně stránek uživatelských účtů i diskuzních stránek (cca 100 GB).
- `all-titles-in-ns0` – seznam názvů všech článků (bez ostatních typů stránek, 300 MB)
- `pages-meta-history` – všechny stránky včetně kompletní historie změn (několik terabajtů, soubor je rozdělen do více částí a musí být specifickým způsobem spojen)

Výše uvedený seznam souborů není úplný. Pro účely extrahování informací z článků a identifikaci typů jejich entit bude třeba pracovat se seznamem všech článků, který bude obsahovat jejich názvy, texty a jmenné prostory. Proto bude tato práce pracovat se souborem

`pages-articles.xml`, který tyto informace obsahuje. Soubor je uložen ve formátu XML s pevně danou strukturou, která je definována pomocí následujícího XSD [9] dokumentu:

https://www.mediawiki.org/xml/export-0.10.xsd

Kořenovým elementem tohoto XML souboru je prvek `<mediawiki>`, který obsahuje jeden element `<siteinfo>` a několik *(nula a více)* elementů `<page>`. Prvek `<siteinfo>` uchovává informace o encyklopedii, jejíž stránky jsou v daném dumpu obsaženy, jako je její název, jazyková mutace a seznam jmenných prostorů stránek s jejich popisem. Prvek `<page>` pak obsahuje informace o jedné konkrétní stránce z této encyklopedie v níže uvedeném formátu (jeden dump soubor obsahuje tolik prvků `<page>`, kolik daná encyklopedie obsahuje stránek vyhovujícím filtrům pro daný typ dump souboru):

```
<page>
  <title>Dog</title><ns>0</ns><id>4269567</id>
  <revision>
    <id>821387961</id><parentid>820506738</parentid>
    <timestamp>2018-01-20T04:44:39Z</timestamp>
    <contributor>
      <username>TAnthony</username><id>1808194</id>
    </contributor>
    <comment>Update unknown, deprecated or missing [...]</comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text xml:space="preserve">{{about|the domestic dog [...]</text>
    <sha1>3jcp6i6y8d3633vcbddpze0wrisys7p</sha1>
  </revision>
</page>
```

Jak je z výše uvedené ukázky patrné, z dump souboru je možné o jednotlivých článcích získat podobné informace, jako pomocí API (název stránky *Dog*, jmenný prostor *0*, ID stránky *821387961*, zdrojový text stránky ve formátu MediaWiki atd.). Příklad úplné struktury dump souboru včetně elementu `<siteinfo>` lze nalézt v příloze A.

2.3 Formátování článků pomocí MediaWiki

Za účelem formátování textu MediaWiki (a Wikipedie) používá vlastní značkovací jazyk. Tento jazyk do jisté míry připomíná značkovací jazyk Markdown [23] a rovněž umožňuje konverzi do HTML, avšak nabízí více funkcí, jako jsou podpora tabulek, kategorií či šablon. V této kapitole budou jednak popsány základní formátovací značky tohoto jazyka, a také ty značky, které budou mít pro identifikaci stránek nějaký význam.

Stejně jako u podobných formátovacích jazyků i zde platí, že text, který je zadán přímo bez uvedení jakýchkoliv formátovacích značek, bude vypsán obyčejným písmem běžné velikosti. Formát neobsahuje žádnou hlavičku ani patičku, formátovací značky se používají pouze ke změně vlastností části textu. Text uzavřený do dvou apostrofů bude vypsán *kurzívou*, text ve třech apostrofech bude vypsán **tučně**. Například věta:

```
Příliš ''žluťoučký'' kůň úpěl ''ďábelské '''ódy'''''.
```

bude vypsána následovně: Příliš *žluťoučký* kůň úpěl *ďábelské* ***ódy***. **Nadpisy** jednotlivých kapitol jsou tvořeny napsáním `= Text nadpisu =`, nebo `== Text podnadpisu ==`, či `=== Text nadpisu 3. úrovně ===` apod. na samostatném řádku, kde počet rovnítek určuje

úroveň nadpisu. Nadpisy první úrovně jsou vyhrazeny pro název článku, hlavní kapitoly textu jsou tvořeny s dvěma rovnítky, pro vytvoření její podkapitoly (např. *1.2*) se použijí tři rovnítka atd. Pro vytvoření **odkazu na jinou stránku** v rámci stejné jazykové mutace wiki lze použít `[[Název cílové stránky]]`.

Článek může být také zařazen do kategorie. Mechanismus kategorií bude podrobněji popsán v kapitole 5.2. Pro přidání článku do **kategorie** s názvem *Název kategorie* stačí kamkoliv do článku napsat text `[[Category:Název kategorie]]`[2]. Jeden článek může být umístěn ve více kategoriích. Tento text většina editorů umisťuje na samostatný řádek na konec článku, viz níže uvedená část textu článku Dog (tento přístup lze vypozorovat u naprosté většiny článků na Wikipedii). Ukázka zobrazení seznamu kategorií tohoto článku na Wikipedii je zobrazena na obrázku 2.1. Podobný seznam lze nalézt na konci téměř každého článku na Wikipedii.

```
... text článku ...
[[Category:Dogs]]
[[Category:Cosmopolitan vertebrates]]
[[Category:Extant Late Pleistocene first appearances]]
[[Category:Mammals described in 1758]]
```

Obrázek 2.1: Ukázka kategorií na Wikipedii

Posledním mechanismem značkovacího jazyka používaného na Wikipedii, který bude v této práci popsán, je mechanismus šablon. **Šablony** slouží k vložení stejných úseků textů či kódů do více článků. Příklad použití šablony lze vidět na obrázku 2.2 získaného ze článku Homer Simpson. Zde je na konci všech článků o postavách ze seriálu Simpsonovi uveden totožný seznam s odkazy na ostatní postavy z tohoto seriálu.

Obrázek 2.2: Ukázka šablony na Wikipedii

Pro vytvoření šablony je nutné nejprve vytvořit samostatnou stránku s touto šablonou, která poté bude vložena do jednotlivých článků. Tato stránka by měla být až na výjimky uložena v jmenném prostoru číslo 10, který je pro vytváření šablon určen (nicméně je možné načíst šablonu i z jiného jmenného prostoru [26]). Pro vložení šablony bez parametrů lze do textu článku vložit text `{{Název šablony}}` nebo `{{Jmenný prostor:Název šablony}}`,

[2] V případě jiných než anglických jazykových mutacích Wikipedie (např. v české) lze místo `[[Category:Název kategorie]]` rovněž použít `[[Kategorie:Název kategorie]]`. Nicméně editoři české Wikipedie používá oba způsoby zápisu, proto skript popsaný v této práci akceptuje obě varianty bez ohledu na použitou jazykovou mutaci Wikipedie.

kde *Název šablony* je jméno stránky, kde je tato šablona uložena a *Jmenný prostor* je název jmenného prostoru, odkud bude tato stránka načtena (pokud není uveden, bude použit právě jmenný prostor *10 – Template*).

Dále tento systém umožňuje i vytváření složitějších **šablon s parametry**. Výhodou šablon s parametry je, že vkládaný text nemusí být ve všech článcích totožný, ale lze do každého článku vložit částečně modifikovaný text. Příkladem může být šablona Infobox dog breed, která slouží k vytvoření tabulky s informacemi o psí rase. Na obrázku 2.3 je znázorněno použití této šablony na stránkách Golden Retriever, Siberian Husky a Bernese Mountain Dog (zleva) – tato šablona se vždy nachází v pravém horním rohu stránky.

Obrázek 2.3: Ukázky šablony s parametry (zobrazeny jsou pouze horní části šablon)

Ve všech třech případech šablon zobrazených na obrázku 2.3 byla použita stejná šablona Infobox dog breed, rozdíly v použití této šablony mezi jednotlivými články se liší pouze v parametrech této šablony. Jak je z výše uvedených obrázku patrné, tato šablona vždy tučným písmem vypíše název dané rasy, vloží obrázek odpovídajícího zvířete, vypíše přezdívky dané rasy, zemi jejího původu, obvyklou hmotnost dospělého psa této rasy a některé další údaje (na obrázcích není zobrazena celá šablona). Některé údaje mohou být také nepovinné a v případě jejich nezadání se určitá část kódu šablony při jejím vložení do článku nepoužije: například u článku Golden Retriever nebyly zadány jiné názvy (přezdívky) této rasy, a tak se řádky *Other names* a *Common nicknames* v tomto případě vůbec nepoužily a vykreslená tabulka začala řádkem *Origin*.

Pro použití šablony s parametry je třeba do článku místo `{{Název šablony}}` vložit `{{Název šablony|Parametr1|Parametr2}}`, kde `Parametr1` a `Parametr2` představují tyto parametry (počet parametrů může být libovolný). Parametry mohou být buď nepojmenované, nebo pojmenované. U **nepojmenovaných parametrů** se na místo `Parametr1` vloží hodnota daného parametru. Jednotlivé parametry se pak od sebe odlišují na základě jejich pořadí. V případě **pojmenovaných parametrů** se na místo `Parametr1` vloží `název=Hodnota`, kde `název` je název daného parametru a `Hodnota` je jeho hodnota. Názvy parametrů jsou citlivé na velikost písmen. Ovšem u obou typů parametrů může být před i za znaky `|`, `=`, `{{` a `}}` libovolný počet bílých znaků (mezer, tabulátorů a odřádkování). Jedna šablona může současně používat pojmenované i nepojmenované parametry. Například šablona Infobox dog breed používá jen pojmenované parametry. Její použití u článku Siberian Husky vypadá následovně (parametry, jejichž význam není na obrázku 2.3 vidět, byly vynechány):

```
{{Infobox Dogbreed
| altname = Chukcha, Chuksha
| country = [[Siberia]]
| image = Black-Magic-Big-Boy.jpg
| image_caption =
| name = Siberian Husky
| nickname = Husky<br>Sibe
| maleweight = {{convert|45|-|60|lb|kg}}
| femaleweight = {{convert|35|-|50|lb|kg}}
}}
```

Názvy zde použitých parametrů (`altname`, `country`,...) jsou definovány při vytváření příslušné šablony. Hodnoty parametrů jsou de facto libovolné a mohou využívat i další formátovací značky MediaWiki. Například hodnotou parametru `altname` je odkaz na článek Siberia, hodnota `nickname` obsahuje ruční zalomení řádku a hodnota parametru `maleweight` této šablony dokonce využívá další šablonu *Convert*, která slouží k zobrazení váhy v kilogramech i librách (výsledkem `{{convert|45|-|60|lb|kg}}` je text *45–60 pounds (20–27 kg)*). Šablona *Convert* naopak využívá pouze nepojmenované parametry (v tomto případě jich bylo uvedeno pět, a to *45*, -, *60*, *lb* a *kg*).

Kapitola 3

Existují zdroje

V této kapitole bude popsáno, jaké v současnosti existují projekty, které tuto nebo podobnou problematiku řeší. Nejprve budou zmíněny některé skripty vytvořené v rámci skupiny znalostních technologií (KNOT) na FIT VUT v Brně. Poté bude popsán projekt DBpedie, který se zabývá extrakcí informací z Wikipedie. Při popisu DBpedie bude zmíněn i projekt YAGO. Tyto systémy dosahují relativně vysoké úspěšnosti, ale jak bude ukázáno v podkapitole 3.2.3, obvykle nenaleznou zcela všechny články patřící do dané entity a nepodporují dynamickou definici entit. Tato práce se snaží tyto nedostatky odstranit.

3.1 Řešení vzniklá v rámci skupiny KNOT na FIT VUT

V rámci skupiny znalostních technologií [14] bylo vytvořeno několik projektů, jejichž cílem je identifikace typů článků na Wikipedii nebo extrakce informací z určitých typů článků. Projekty zabývající se určováním typů entit většinou podporují identifikaci několika určitých typů entit pomocí vyhledávání staticky nastavených vzorů v textu článků.

V této práci bylo zvoleno řešení, které tyto vzory vyhledává automaticky, a může být tedy použito k identifikaci libovolné entity. Návrh řešení tohoto systému z jiných systémů vytvořených v rámci výzkumné skupiny nevycházel, nicméně výsledky některých skriptů vytvořených v této skupině byly použity k porovnání správnosti vygenerovaných dat oběma systémy.

Také byl v této skupině vytvořen skript na extrakci prvních vět z článků Wikipedie v lidsky čitelné podobě (bez formátovacích značek) [13]. Jelikož tento skript dosahuje spolehlivých výsledků a získání první věty článku je pro řešení této úlohy pouze dílčí částí, byl tento skript pro účely tohoto systému převzat. První věta článků je využita především při vyhledávání definičního slova, které bude popsáno v kapitole 6.

3.2 DBpedie

Dalším systémem, který se zabývá extrakcí informací ze článků Wikipedie a určování jejich typů, je DBpedie [15]. Tento projekt se především zabývá extrakcí strukturovaných informací z Wikipedie. Nalezené informace z těchto článků se ukládají ve formátu RDF [21], který k uložení dat používá množinu trojic. Každá trojice se skládá z podmětu, vlastnosti a předmětu. Příkladem této trojice může být (dbr:Alban_Lafont, dbo:birthDate, 1999-01-23). První prvek této trojice obvykle odkazuje na konkrétní článek, druhý prvek obsahuje informaci o typu extrahované entity a poslední prvek obsahuje extrahovanou hodnotu.

Za účelem určení typů článků jsou zajímavé trojice, které jako druhý prvek obsahují **rdf:type**, např. (dbr:Alban_Lafont, rdf:type, dbo:Person) nebo (dbr:German_Shepherd, rdf:type, dbo:Mammal).

K těmto trojicím lze přistupovat různými způsoby. Jednak je možné z webových stránek[1] DBpedie stáhnout dump soubor, který obsahuje všechny nalezené trojice, podobně jako dump soubor Wikipedie obsahuje všechny její články. Tento soubor lze procházet běžnými nástroji pro práci s textovými soubory, nebo jej lze naimportovat do speciálního databázového systému. Kromě tohoto DBpedia nabízí on-line přístup k databázovému serveru, kde jsou uloženy totožné trojice. Tento server je dostupný z http://dbpedia.org/sparql, nicméně protože se jedná o veřejný server, je zde omezen přístup ke sdíleným prostředkům maximální délkou vykonávání dotazu a maximálním počtem vrácených položek.

3.2.1 SPARQL

Pro komunikaci s výše uvedeným databázovým serverem DBpedie se používá dotazovací jazyk SPARQL [10]. Jedná se o obdobu jazyka SQL určenou pro DBMS (databázové systémy), které pracují s daty uloženými ve formátu RDF (množině trojic). V této podkapitole budou ukázány základní příkazy tohoto jazyka. Všechny zde uvedené dotazy budou vytvořeny tak, aby vracely maximálně 100 výsledků (nikoliv prvních 100 výsledků, neboť pořadí trojic v množině není definováno). Pro změnu počtu vrácených výsledků je třeba upravit číslo v klauzuli `LIMIT 100`.

- Pro vrácení všech trojic v systému lze použít následující dotaz:

```
SELECT ?a ?b ?c WHERE {
    ?a ?b ?c .
}
LIMIT 100
```

- K vypsání trojic, které na určité pozici obsahují konkrétní hodnotu je třeba danou hodnotu uvést místo proměnné `?a`, `?b`, nebo `?c` v klauzuli `WHERE` a odebrat tuto proměnnou v seznamu za příkazem `SELECT`. Například pro vypsání všech osob (trojic, které na druhém místě obsahují `rdf:type` a na třetím místě `dbo:Person`) lze použít tento dotaz:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?person WHERE {
    ?person rdf:type dbo:Person .
}
LIMIT 100
```

- Dále tento jazyk podporuje agregační výrazy. Následující dotaz lze použít ke zjištění, kolik je na DBpedii nalezeno videoher:

```
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT count(distinct ?p) as ?cnt where {
    ?p rdf:type dbo:VideoGame .
}
```

[1] http://wiki.dbpedia.org/develop/datasets

3.2.2 Dbpedia live

Nevýhodou dump souboru, jehož možnost stažení byla výše uvedena, je jeho stáří. Poslední verze tohoto souboru je z října 2016 a od této doby se mnoho článků na Wikipedii změnilo. Nicméně byl vytvořen projekt DBpedie Live, který pracuje s posledními verzemi článků na Wikipedii. Tento systém přistupuje k Wikipedii pomocí jejího aplikačního rozhraní (API; viz kapitola 2.2.1), ze kterého si v pravidelných intervalech zjišťuje seznam naposledy změněných článků. Pokud je některý článek změněn, tak DBpedie Live:

1. zjistí, jaké trojice, které se nacházejí v systému, byly vygenerovány na základě původní verze článku.
2. Pomocí API stáhne aktuální verzi článku a vygeneruje novou množinu trojic.
3. Na základě staré a nové množiny trojic vygeneruje množiny přidaných, upravených a smazaných trojic.
4. Na základě množin vytvořených v předchozím bodě aplikuje změny v databázovém systému tak, aby data zůstala aktuální (přidá nové záznamy a upraví nebo smaže staré).
5. Informace o přidaných, upravených a smazaných trojicích uloží do logovacího souboru.

Podobně jako DBpedia nabízí DBpedia live on-line přístup k databázovému systému, kde jsou uloženy aktuální data tohoto systému. Tento systém je přístupný z následujících URL adres: `http://live.dbpedia.org/sparql` a `http://dbpedia-live.openlinksw.com/sparql` (obě adresy nabízí přístup k totožným trojicím).

Pokud jsou omezení tohoto serveru na maximální počet vrácených záznamů problém (což v případě, že budeme chtít vypsat seznam všech osob nebo jiné rozsáhlé entity, problém bude), je možné si lokálně stáhnout dump soubor DBpedie Live, který lze nalézt na `http://live.dbpedia.org/dumps/`. Tento soubor však z pochopitelných důvodů nemůže být generován při každé změně článku (každou minutu). Poslední verze tohoto dumpu pochází z února 2017. K získání aktuálních dat DBpedie Live je však možné si stáhnout jednotlivé záznamy s přidanými, upravenými a smazanými trojicemi[2] a provést tyto změny nad staženým dumpem (DBpedie Live stále generuje nové soubory se změněnými trojicemi, poslední soubor je několik minut až hodin starý). DBpedie Live nabízí synchronizační nástroj [4], který dokáže automaticky stahovat nové soubory se změnami a provádět je nad lokální kopií dat z DBpedie. Tento nástroj předpokládá, že stažené záznamy jsou uloženy v databázovém systému Virtuoso [19].

3.2.3 Zhodnocení

DBpedie extrahuje širokou sadu dat a dokáže identifikovat 753 typů entit článků, jejichž seznam lze nalézt v [5]. V tabulce 3.1 jsou uvedeny počty článků některých z těchto entit, které byly DBpedií nalezeny. V některých případech lze paradoxně zpozorovat, že DBpedie Live, která obsahuje novější data, než jsou data uložená na serveru DBpedie, najde méně článků, než se jich nachází na serveru DBpedie se staršími daty. Toto může mít několik příčin. Jednak se ve starší verzi extraktoru dat mohla nacházet chyba, která byla v novější verzi opravena. Toto je například případem u osob, kde provedením SPARQL dotazu na

[2] Dostupné z: `http://live.dbpedia.org/changesets/`

vypsání všech osob uvedeným v kapitole 3.2.1 na serveru DBpedie lze nalézt mnoho článků, které osobami nejsou, například 23rd century BC, 10BASE5 a CMYK color model. Stejně tak může být problém s novou verzí extraktoru DBpedie, který již nemusí být schopen nalézt všechny články, které našla jeho stará verze. Toto se například stalo u článků Douglas Smith (broadcaster), Domenico Cimarosa a Emperor Sutoku. Také mohly být některé články z Wikipedie odstraněny.

Tabulka 3.1: Počty nalezených článků některých entit na DBpedii a DBpedii Live

Entita	**DBpedie**	**DBpedie Live**
Osoby (dbo:Person)	1 818 074	983 101
Umělci (dbo:Artist)	82 757	92 862
Playboy playmates (dbo:PlayboyPlaymate)	274	275
Sériový vrazi (dbo:SerialKiller)	0	0
Zvířata (dbo:Mammal)	15 109	7 478
Psi (dbo:Dog)	0	0
Společnosti (dbo:Company)	109 629	77 744
Přístroje (dbo:Device)	27 457	36 673
Videohry (dbo:VideoGame)	28 869	21 459
Letiště (dbo:Airport)	16 684	12 882
Události (dbo:Event)	77 583	74 735

Dále databázový server s daty DBpedie obsahuje trojice extrahované z jiného systému YAGO [11], který dokáže extrahovat data z Wikipedie, GeoNames a WordNetu. Tento systém rovněž dokáže identifikovat typy entit článků na DBpedii. Pro nalezení všech osob identifikované tímto systémem lze použít níže uvedený SPARQL dotaz. Tabulka 3.2 obsahuje počty článků některých entit nalezené tímto systémem.

```
PREFIX yago: <http://dbpedia.org/class/yago/>
SELECT ?p where {
    ?p rdf:type yago:Person100007846 .
}
LIMIT 100
```

Tabulka 3.2: Počty nalezených článků některých entit systémem YAGO

Entita	**Počet článků (YAGO)**
Osoby (Person100007846)	1 010 473
Umělci (Artist109812338)	128 800
Playboy playmates (Playmate110441037)	63
Zvířata (AnimalGroup107993929)	2 057
Společnosti (Company108058098)	71 407
Přístroje (Device103183080)	22 041
Videohry (ComputerGame100458890)	12 270
Letiště (Airport102692232)	12 013
Události (Event100029378)	163 469

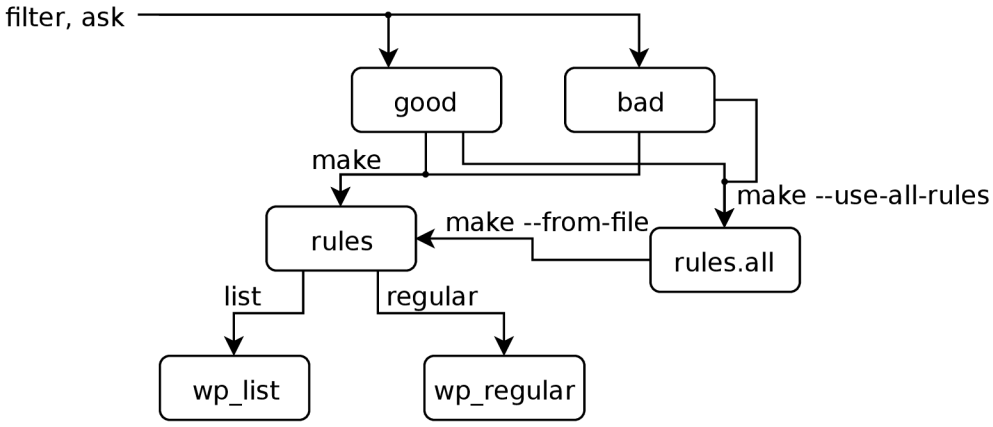
Jak lze z výše uvedených tabulek vidět, úspěšnost jednotlivých systémů i jejich verzí je různá. DBpedie rovněž obsahuje entity, které ještě neumí zpracovat a obsahují 0 článků.

Tyto systémy také umí zpracovat pouze konečný počet typů entit, pro podporu identifikace nové entity je nutné vytvořit nové mapovaní mezi strukturovanými značkami na Wikipedii (obvykle šablonami) a typy entit. Systém vytvořený v této práci naopak umožňuje definovat nové typy entit dynamicky pomocí sad vzorových článků patřících a nepatřících do dané entity.

Kapitola 4

Návrh systému

Tento systém byl navržen tak, aby se při jednom spuštění systému vykonávala vždy jen jedna činnost v závislosti na tom, v jakém režimu byl spuštěn. Výhodou tohoto přístupu je to, že není nutné při každém spuštění skriptu znovu provádět všechny akce, jako je vygenerování seznamu příznaků na základě vzorových článků. Hlavním vstupem pro tento systém je dump soubor Wikipedie, který je vyžadován ve všech režimech. Pro definici a identifikaci jednotlivých entit systém používá soubory uvedené v obrázku 4.1, kde je také naznačeno, jaké režimy skriptu převedou jeden soubor na druhý.



Obrázek 4.1: Základní soubory potřebné pro identifikaci jedné entity článků

Mimo dump soubor Wikipedie jsou hlavními soubory, které definují jednotlivou entitu (skupinu) článků, soubory *good* a *bad*, kde soubor *good* obsahuje názvy vzorových článků, které patří do hledané entity a soubor *bad* obsahuje názvy vzorových článků, které do hledané entity nepatří. Tyto soubory je nutné vytvořit ručně. Pro vypsání názvů všech článků, které se nacházejí na Wikipedii (a tedy mohou být uvedeny v jednom z těchto dvou souborů) lze použít režim systému *filter*. Nebo je možné použít režim *ask*, který bude interaktivně nabízet náhodné stránky z dumpu Wikipedie, které se ještě nenacházejí v souboru *good* nebo *bad* a po načtení odpovědí ze standardního vstupu v příkazové řádce tyto články přidá do souborů *good* či *bad*.

Jakmile jsou vytvořeny vstupní soubory s názvy vzorových článků, je potřeba vygenerovat příznaky resp. asociační pravidla, které budou články v této skupině charakterizovat. K vytvoření těchto příznaků slouží režim skriptu *make*. Při jeho běžném spuštění (bez dalších parametrů) jsou vygenerovány jen ty příznaky, které lze použít k identifikaci článků

v této entitě. Při použití parametru `--use-all-rules` budou vygenerovány všechny nalezené příznaky, i ty, které mohou při identifikaci článků přinášet nepravdivé informace. Tento soubor může být převeden do souboru, který by byl vygenerován bez tohoto příznaku, spuštěním skriptu v tomto režimu s parametrem `--from-file`. Vytvoření souboru se všemi pravidly může být výhodné při úpravách nastavení systému, který nalezené pravidla filtruje a který bude popsán v kapitole 7.4, poněvadž při tomto převodu není nutné procházet všechny články v dumpu.

Po vytvoření příznaků resp. asociačních pravidel, které budou danou entitu identifikovat, již je možné vypsat seznam článků spadajících do této entity pomocí režimu *list* nebo seznam článků, které do této entity nespadají, pomocí režimu *regular*.

Způsob implementace jednotlivých režimů bude popsána v kapitole 7.

Kapitola 5

Příznaky článků

Jak bylo v kapitole 2 naznačeno, Wikipedie nikde u jednotlivých stránek přímo neuchovává, o jakém typu entity daná stránka je. Například u článku *Inspector Rex* nikde není přímo uvedeno, že se jedná o televizní seriál. Přesto se však na této stránce nacházejí jisté indície, díky kterým lze určit, že se s nejvyšší pravděpodobností jedná o seriál. Mezi tyto indície patří například to, že tento článek obsahuje vloženou šablonu *Infobox television* a je zařazen do kategorií jako *Television shows about dogs* či *1990s Austrian television series*. Nicméně jak bude v následujících podkapitolách ukázáno, tyto informace nejsou stoprocentně spolehlivé. Například výše uvedený článek obsahuje i kategorii *Fictional dogs*, která by spíše naznačovala, že tento článek pojednává o psovi než o televizním seriálu. Navíc, jak bude ukázáno v podkapitole 5.2, je těchto kategorií na Wikipedii velké množství a nelézt seznam všech kategorií, které pojednávají o rozsáhlejší skupině stránek (např. o všech lidech, o všech rasách psů, o všech filmech apod.) není triviální.

Z těchto indícií lze vytvořit tzv. **příznaky**, což jsou tvrzení o jednotlivých článcích, jejichž pravdivost lze pro konkrétní článek deterministicky ověřit. Příkladem příznaku může být například tvrzení, že daný článek je zařazen do kategorie *Television shows about dogs*. Pro libovolný článek na Wikipedii lze toto tvrzení jednoznačně ověřit (pro článek *Inspector Rex* je toto tvrzení splněno, protože obsahuje uvedenou kategorii a pro článek *The Beatles (No. 1)* toto tvrzení splněno není, protože neobsahuje uvedenou kategorii).

Informace o tom, které příznaky (tvrzení) jsou u daného článku splněny, lze využít k rozhodnutí, zda daný článek patří do hledané entity. Každý příznak může mít jeden ze tří významů v závislosti na tom, co se o daném článku dozvíme, pokud je tento příznak splněn:

- buď skutečnost, že byl daný příznak splněn, naznačuje, že daný článek **patří** do hledané skupiny stránek (entity) – tzv. **pozitivní příznak** – *např. to, že článek obsahuje kategorii Dog breeds při hledání psích ras*, nebo
- skutečnost, že byl daný příznak splněn, naznačuje, že daný článek **nepatří** do hledané skupiny stránek (entity) – tzv. **negativní příznak** – *např. to, že článek obsahuje kategorii Musicians při hledání psích ras*, nebo
- skutečnost, že byl daný příznak splněn, nijak nenaznačuje odpověď na otázku, zda daný článek patří, nebo nepatří do hledané skupiny stránek (entity) – *např. to, že článek obsahuje kapitolu Further reading při hledání psích ras*. Tyto příznaky se při identifikaci entit snažíme nepoužívat.

V následujících podkapitolách bude popsáno, s jakými typy příznaků bude tato práce dále pracovat. Další kapitoly (zejména kapitola 7.4) se budou zabývat tím, jak vybrat ty příznaky, které budou k identifikaci typů entit co nejvíce užitečné.

5.1 Šablony

Jak bylo uvedeno v kapitole 2.3, jednotlivé články ve svém textu mohou obsahovat šablony. Tyto šablony mohou být použity buď bez parametrů, nebo s pojmenovanými nebo nepojmenovanými parametry. Tato práce umožňuje na základě šablon použitých v článku vytvořit dva typy příznaků.

Prvním typem podporovaných příznaků je test přítomnosti šablony pevně daného jména. Příkladem příznaku tohoto typu může být *„Článek obsahuje šablonu Infobox television“*.

Druhým typem příznaků vytvářených z šablon Wikipedie je test přítomnosti pojmenovaného parametru s pevně daným názvem v libovolné šabloně. Například může jít o příznak *„Článek obsahuje šablonu libovolného jména s parametrem birth_date.“* Výhodou tohoto typu příznaku je, že není nutné znát přesný název šablony v případě, že existuje více podobných šablon (v tomto případě např. Infobox writer, Infobox musical artist a Infobox person).

5.2 Kategorie

Obdobným způsobem, jakým je možné ověřit přítomnost dané šablony v konkrétním článku, lze taktéž zkontrolovat, zda článek obsahuje kategorii určitého názvu. Nicméně použití kategorií k ověření typu entity článků je o něco složitější, jelikož Wikipedie obvykle obsahuje podstatně více kategorií, než článků. Například anglická Wikipedie obsahuje 340 tisíc šablon a 961 tisíc kategorií, které obsahují alespoň dva články[1].

Dále rovněž existuje mnoho podobných kategorií, které se liší jen v číslovce, jako jsou kategorie *1880 births*, *1988 births* či *2005 births*. Přitom lze předpokládat, že všechny tyto kategorie budou pojednávat o osobách. Proto zde popisovaný program umí interně tyto kategorie sloučit do jediné kategorie `<číslo> births`, přičemž původní kategorie zůstanou zachovány pro případ, že by uživatel programu chtěl definovat entitu, která by byla podmnožinou stránek v této sloučené kategorii, tj. v tomto případě například skupinu osob narozenou v konkrétním roce.

V této práci se především porovnává shoda celého názvu kategorie, shoda prvního slova v názvu kategorie a shoda posledního slova v jejím názvu.

Nicméně také je vhodné upozornit na to, že samostatné použití kategorií na Wikipedii k identifikaci článků není příliš spolehlivé – například článek Afroasiatic languages obsahuje kategorii *Afroasiatic peoples*, která by naznačovala, že se jedná o osobu, i když ve skutečnosti se o osobu nejedná. Nebo článek Arthur William à Beckett neobsahuje kategorii, která by přímo obsahovala text *people*, avšak obsahuje kategorie *English male journalists* a *English humorists*, ze kterých lze odvodit, že se jedná o člověka.

[1] Wikipedie rovněž obsahuje 226 207 kategorií a 2 314 704 šablon, které jsou použity pouze v jediném článku. Tyto kategorie a šablony však při identifikaci typů entit článků zcela jistě nebude možné jakýmkoliv způsobem využít. Zdroj: spuštění systému v režimu `counts`.

5.3 Kapitoly

Dále mohou být při identifikaci entit článků využity názvy jeho kapitol. Například článek *Inspector Rex* obsahuje následující kapitoly nejvyšší úrovně: *Synopsis*, *Production*, *Characters*, *Episodes*, *Broadcasters*, *Dubbing*, *DVD releases*, *References* a *External links*. Z toho kombinace názvů *Production*, *Characters*, *Episodes*, *Broadcasters* a *Dubbing* může být silnou indícií, že se jedná o televizní seriál.

Tato práce generuje příznaky testující shodu názvu libovolné kapitoly nejvyšší úrovně s předem daným řetězcem. Tyto příznaky jsou velmi často využívány i v negativním smyslu – například při vyhledávání osob je přítomnost kapitoly s názvem *Episodes* příznakem toho, že daný článek nepojednává o osobě.

5.4 Název článku

U některých článků může i jejich název ukrývat indícii, o čem daný článek je. Konkrétně jsou to články, na konci jejichž názvu se nachází závorka s jejich upřesněním za účelem rozlišení několika článků, které by jinak měly totožný název. Příkladem mohou být články *Ada (programming language)*, *Ada (name)* a *Ada (food)*. U těchto článků lze i bez čtení jejich textu určit, že první článek s vysokou pravděpodobností pojednává o programovacím jazyce, druhý o křestním jméně a poslední o jídle.

U anglicky psaných článků lze obdobným způsobem využít i první dvě slova názvů článku u článků, jejichž názvy obsahují alespoň tři slova a druhým slovem je slovo *of*. Jde například o články *Economy of India*, *Geography of Canada* a *Demographics of the world*.

Tato práce bude využívat dva typy příznaků pracující s názvem článku, které vycházejí z výše uvedených poznatků: test textu v závorce na konci názvu článku a test prvních dvou slov víceslovných názvů článků.

5.5 Definiční slovo

Také však existují články, které neobsahují žádnou masově užívanou kategorii ani šablonu a ani názvy jejich kapitol či název článku neposkytují žádné smysluplné indície o tom, o čem daný článek je. Mezi takové články patří například článek Actaeon, který obsahuje šablonu *Infobox deity*, kterou používá jen 653 článků, a kategorie *Artemis*, *Deaths due to dog attacks*, *Metamorphoses into animals in Greek mythology*, *Mythological Greek archers* a *Dogs in art*. S výjimkou kategorie *Mythological Greek archers*, která je celkově použita v jedenácti článcích z celé Wikipedie, žádná kategorie nenaznačuje, že tento článek pojednává o osobě (zatímco přítomnost kategorie *Dogs in art* mylně naznačuje, že by daný mohl pojednávat o psovi). Podobnou situaci lze vypozorovat například i u osob Nereus (v tomto článku nelze dokonce ani najít žádnou vhodnou šablonu), Isaac či Artemis.

Z tohoto důvodu byl implementován mechanismus vyhledávání definičního slova, který se analýzou přirozeného textu snaží v první větě článku vyhledat druhové podstatné jméno (tzv. definiční slovo) podobně jako je to popsáno v první kapitole článku [6], které přináší informaci o tom, o čem daný článek je. Tento systém byl inspirován tím, že většina prvních vět článků na Wikipedii si je podobná a většinou obsahuje stručný popis, o čem daný článek je, viz například *(v závorkách jsou uvedeny názvy článků, odkud byla první věta převzata, delší věty bylo zkráceny, přidáno zvýraznění definičního slova a slov is a)*:

- Ankara (...) **is the capital** of the Republic of Turkey. *(Ankara)*

- Aries **is** one of **the constellations** of the zodiac. *(Aries (constellation))*
- Audi AG **is a** German automobile **manufacturer** that designs, engineers,... *(Audi)*
- Asteroids **are** minor **planets**, especially those of the inner Solar System. *(Asteroid)*

Cílem tohoto příznaku je v první větě článku nalézt definiční podstatné jméno, které definuje, o čem daný článek pojednává. Ve výše uvedených příkladech jsou to slova *capital*, *constellations*, *manufacturer* a *planets*. Tyto slova mohou být poté porovnána s předem definovanými slovy, obdobně jako při porovnávání názvů kategorií, šablon, nebo kapitol.

Implementace tohoto systému je z pochopitelných důvodů závislá na konkrétním (přirozeném) jazyce, v této práci jsou podporovány věty psané v češtině nebo v angličtině. Způsob, jakým je systém pro vyhledávání definičního slova implementován, bude podrobně popsán v kapitole 6, kde budou také popsány výsledky tohoto systému.

5.6 Typ členu v první větě

Poslední indícií, která bude v této práci použita, je typ použitého členu před výskytem názvu článku v první větě tohoto článku. Tento typ příznaku byl implementován pouze pro anglicky psané věty a dále by mohl být implementován při identifikaci článků psaných v jazycích, které jmény používají členy před podstatnými (v případě angličtiny *a*, *an*, *the*) – tj. nelze ho použít u česky psaných článků.

Úkolem tohoto příznaku je nalézt v první větě článku jeho název a podívat se, jaký člen je uveden před ním. Například článek Football player s první větou *„**A** football player is a sportsperson who plays one of the different types of football.“* obsahuje před výskytem slovního spojení *football player* neurčitý člen *a*. Tento typ příznaku se používá výhradně k vytváření negativních příznaků, tj. k vyloučení možnosti, že daný článek patří do hledané entity. Například lze tímto způsobem spolehlivě oddělit články o lidech od článků o profesích – před názvem člověka se v angličtině člen nepíše, zatímco při prvním zmínění názvu profese se skoro vždy píše neurčitý člen (*a* nebo *an*). Příklady mohou být následující první věty ze článků o profesích, které před názvem článku obsahují neurčitý člen:

- **A chef** is a trained professional cook who... (Chef),
- **A politician** is a person active in party politics... (Politician)

a následující věty ze článků o osobách, kde se před jejich jmény žádný člen nenachází:

- **Donald John Trump** is the 45th and current President of the United States... (Donald Trump)
- **Rolando Jorge Pires** da Fonseca, known simply as Rolando, is a Portuguese footballer... (Rolando (footballer))

Vyhledávání se omezuje na první větu článku především proto, že se v angličtině používá jiný typ členu v případě, že se o dané věci autor zmiňuje poprvé (určitý, pokud se jedná obecně známou unikátní věc, např. *the sun*, jinak neurčitý), než při použití dříve zmíněné věci (určitý nebo žádný) – zde by tedy docházelo ke ztrátě informací.

Kapitola 6

Definiční slova článků

Jak již bylo uvedeno v kapitole 5.5, jedním typem příznaků, kterou lze využít při určování typů entit článků, je porovnání definičního slova článku s dříve nalezenými definičními slovy podobných článků. Definiční slovo je takové slovo (obvykle podstatné jméno) nebo slovní spojení, které obecně definuje, o čem daný článek je. Například definiční slovo článku Richard Billows může být *professor* nebo *person*, pro článek 1129 jde o slovo *year* a pro článek Christ in Majesty o slovo *image* nebo případně o slovní spojení *image of Christ*. Toto definiční slovo lze u většiny článků na Wikipedii nalézt v jeho první větě. Například v první větě z článku Dragon 32/64 *„The Dragon 32 and Dragon 64 are home* ***computers*** *that were built in the 1980s.“* se nachází definiční slovo *computers* a ve větě *„Listopad je jedenáctý* ***měsíc*** *gregoriánského kalendáře v roce.“* z článku Listopad lze nalézt slovo *měsíc*.

V tomto systému je implementováno pouze vyhledávání definičního slova v první větě článku, jelikož u velké části článků jej lze v této větě spolehlivě nalézt. V této kapitole bude popsáno, jakým způsobem je implementováno vyhledávání tohoto slova v česky a anglicky psaných větách článků, kolik definičních slov skript dokázal najít a která definiční slova jsou nejčastější. Nakonec bude v podkapitole 6.4 popsáno, jakým způsobem by bylo možné rozšířit podporu získávání definičních slov ze článků psaných v jiných jazycích.

6.1 Nalezení první věty článku a její rozdělení na tokeny

K tomu, aby bylo možné v první větě článku nalézt definiční slovo, je nejprve potřeba **získat** tuto **první větu článku**. Nicméně, jak bylo uvedeno v kapitolách 2.2.1 a 2.2.2, API Wikipedie i dump soubor s daty z Wikipedie umožňuje pouze přístup k celým článkům včetně formátovacích značek MediaWiki (které byly popsány v kapitole 2.3), nikoliv k jednotlivým větám. Nejprve je tedy nutné tyto formátovací značky zpracovat a vhodně odstranit (názvy kapitol je třeba odstranit úplně, tučný text musí být zachován (převeden na obyčejný), u odkazů je třeba odstranit URL adresu, ale zanechat jejich popisek apod.) a poté z takto získaných vět nalézt tu první.

K tomuto úkolu byl využit skript [13] vytvořený v rámci výzkumné skupiny znalostních technologií (KNOT) na FIT VUT Brno, který jako svůj vstup přijímá text článku Wikipedie včetně formátovacích značek a jako výstup vrací seznam všech jeho vět bez formátovacích značek. Tento skript svoji úlohu zvládá na relativně dobré úrovni, avšak dopouští se několika drobných chyb, které jsou v této práci řešeny dodatečným zpracováním vráceného textu z externího skriptu. Zejména se jedná o opravu situace, kdy skript považuje tečku ve spojení

„5. dubna 2015“ za konec věty a o neúplné odstranění některých složitějších formátovacích značek.

Po úspěšném nalezení první věty článku z ní ještě budou **odstraněny všechny závorky** včetně textu, který se v nich nacházel, neboť bylo zjištěno, že hledaná definiční slova se ze zásady v závorkách nenacházejí. Toto je dáno tím, že editoři Wikipedie do závorek v první větě obvykle vkládají doplňující informace, jako jsou název daného článku v jiných jazycích nebo datum narození či úmrtí osob. Příkladem prvních vět článků obsahující závorky mohou být *„Labrador (anglicky Labrador Peninsula, francouzsky Péninsule du Labrador) je poloostrov na severovýchodě Kanady…“* a *„15. prosinec je 349. den roku podle gregoriánského kalendáře (350. v přestupném roce).“* Hledanými definičními slovy v těchto ukázkách jsou slova *poloostrov* a *den*, která se uvnitř závorek nenacházejí.

K tomu, aby bylo možné s touto větou dále pracovat, je nutné ji **rozdělit na** jednotlivá slova, resp. **tokeny**. K tomuto rozdělení byla využita knihovna pro zpracování přirozeného jazyka spaCy [7], která se bude používat i v dalších fázích hledání definičního slova. Každé slovo nebo interpunkční znaménko v původní větě bude reprezentováno samostatným tokenem[1], bílé znaky (mezery) budou zahozeny. Každý z těchto tokenů vytvořených pomocí knihovny spaCy bude dále zapouzdřen do instance vlastní třídy `Word`, která umožní k jednotlivým slovům ukládat další údaje potřebné k nalezení definičního slova, které budou popsány v dalších podkapitolách. Například z věty *„Albania, (…) officially the Republic of Albania (Albanian: Republika e Shqipërisë,), is a country in Southeastern Europe.“* bude vygenerováno následujících 15 tokenů (tokeny se generují až po odstranění textu v závorkách):

Albania	,	officially	the	Republic	of	Albania	,
is	a	country	in	Southeastern	Europe	.	

6.2 Základní algoritmus pro nalezení definičního slova

Poté, co je první věta článku úspěšně nalezena a rozdělena na tokeny, je možné mezi těmito tokeny začít hledat definiční slovo článku. Vzhledem k tomu, že formulace prvních vět článků na Wikipedii si jsou velice podobné, byl algoritmus pro vyhledání definičního slova navržen tak, aby spolehlivě pokryl věty typu, které se podobají velké části prvních vět na Wikipedii. Rozhodně není cílem tohoto algoritmu nalézt *nějakým způsobem důležité podstatné jméno* v libovolné anglicky nebo česky psané větě ani pokrýt 100 % prvních vět článků na Wikipedii. Například tento algoritmus nenajde žádné definiční slovo v první větě článku History of the Comoros *„The history of the Comoros goes back some 1,500 years.“*, protože tato věta žádné smysluplné definiční slovo neobsahuje. Stejně tak nebude nalezeno žádné definiční slovo v první větě článku Transport in Ivory Coast *„Ivory Coast invested remarkably in its transport system.“*, i když by při použití sofistikovanějšího systému bylo možné najít slovo *transport* či slova *trasport system*, poněvadž by tento systém musel pochopit význam celé této věty a při tom se velmi pravděpodobně dopouštěl většího počtu chyb (a navíc by v tomto případě stejně nebylo nalezeno více informací něž lze získat analýzou prvních dvou slov názvu článku).

Základní algoritmus použitý k nalezení definičního slova bude vysvětlen na první větě článku Animalia (book):

„Animalia **is an** illustrated children's **book** by Graeme Base.“

[1] Slovo s interpunkčním znaménkem, které se nacházejí bezprostředně za sebou, vytvoří celkem dva tokeny.

Prvním krokem tohoto algoritmu je v této větě **najít tvar slovesa *být***, v tomto případě se jedná o dvojici slov *is an* (obecný postup bude popsán v podkapitole 6.2.1). Na anglické Wikipedii se toto sloveso nachází v 87,6 % prvních vět článků. Cca 40 % článků, kde se toto sloveso nenachází, navíc jsou rozcestníky. U vět, které toto sloveso neobsahují, bude algoritmus ukončen bez nalezení definičního slova. Před tímto slovem lze očekávat název pojmu, který bude za slovesem *být* definován, obvykle se jedná o název článku, stejně jako tomu je v tomto případě (slovo *Animalia*).

Hledané definiční slovo se vždy nachází za slovesem *být*, nicméně mezi tímto slovesem a hledaným definičním slovem se mohou nacházet i jiná slova (v tomto případě jsou to slova *illustrated* a *children's*). V základní variantě tohoto algoritmu se očekává tvar první věty odpovídající tomuto regulárnímu výrazu:

```
(cokoliv)+ (sloveso být) (přídavné jméno)* (podstatné jméno) (cokoliv)*
```

tj. předpokládá se, že mezi slovesem *být* a hledaným podstatným jménem se může nacházet libovolný počet (i nula) přídavných jmen. Za určitých okolností na tomto místě mohou být tolerovány i jiné slovní druhy, například číslovka, pokud se za ní nachází podstatné nebo přídavné jméno nebo číslovka (*two dogs* je v pořádku, *one of the dogs* nikoliv).

6.2.1 Vyhledání slovesa být

Vyhledání slovesa *být* v první větě článků je velmi snadný úkol. V případě česky[2] psaných vět stačí postupně projít seznamu tokenů a najít ten, jehož text je jedno ze slov *je*, *byl*, *byla*, *bylo*, *bude*. V případě anglicky psaných vět skript prochází dvojice po sobě jdoucích tokenů a hledá takovou dvojici tokenů (w_1, w_2), kde w_1, w_2 představují texty odpovídajících tokenů, $w_1 \in \{is, are, was, were\}$ a $w_2 \in \{a, an, the\}$. Pokud není odpovídající dvojice nalezena, pokusí se skript alespoň najít token, jehož text je *is*, *are*, *was* nebo *were* (bez toho, aby se za ním nacházel určitý nebo neurčitý člen).

Důvodem, proč se systém v případě anglicky psaných vět snaží ke slovesu zahrnout i člen před dalším podstatným jménem je, že bezprostředně za slovesem *být* se tento člen nacházet může, ale mezi dalšími slovy (přídavnými jmény) a hledaným definičním jménem nikoliv. Při opačném přístupu by hrozilo riziko, že skript bude ve větách jako *„Looe Island (...)* ***is a*** *small island* ***a*** *mile from the mainland town of Looe off Cornwall, England."* (z článku Looe Island) považovat slova *small*, *island* a *a* za přídavná jména a jako definiční jméno bude označeno slovo *miles* (správně by měl systém za přídavné jméno považovat jen slovo *small* a slovo *island* vrátit jako nalezené definiční podstatné jméno).

6.2.2 Určení slovních druhů

K tomu, aby systém na vyhledávání definičních slov mohl fungovat, je naprosto nezbytné správně určit slovní druhy jednotlivých slov (tokenů) v procházené větě, přičemž nejvíce je tento systém citlivý na chyby udělané při rozlišení podstatných a přídavných jmen (přídavná jména se přeskakují, zatímco podstatná jména jsou výsledky vyhledávání definičních slov). Přitom zejména v angličtině je běžné, že jedno slovo může v závislosti na kontextu při jeho použití mít více slovních druhů – např. slovo *island* může znamenat jak *ostrov* (podstatné jméno), tak i *ostrovní* (přídavné jméno, např. ve spojení *island country*).

[2] Velká část systému na vyhledávání definičních slov je implementována různě v závislosti na tom, v jakém jazyce byl příslušný článek zapsán. Získání informace o tom, v kterém jazyku je daný článek napsán, není problém, neboť jak bylo uvedeno v kapitole 2.2.2 a je patrné v příloze A, dump soubor Wikipedie obsahuje kód jazykové mutace daného souboru.

Určení slovních druhů v anglicky psaných větách

Proto je v případě anglicky psaných vět k detekci slovních druhů použita knihovna pro zpracování přirozeného textu spaCy [7], která s využitím statistických modelů, které jsou taktéž součástí této knihovny, dokáže tuto úlohu velmi dobře splnit. Přesto je však nutné před reálným použitím slovních druhů získaných z této knihovny provést několik úprav, které opraví nedostatky, kterých se tato knihovna dopouští:

- Ze složených slov, která jsou od sebe oddělena pomlčkou (např. *high-precision*), spaCy vytvoří tři tokeny, u kterých jsou slovní druhy jednotlivých slov vyhodnoceny zvlášť, například:

Token:	high	–	precision
Slovní druh:	Adjective	Punctuation	Noun

  toto je však při vyhledávání definičních slov nevýhodné, neboť by mohlo dojít k situaci, že by se výsledným definičním slovem stala jen tohoto půlka slova (ta, která je podstatným jménem, v tomto případě *precision*).

  Tento problém byl vyřešen zavedením interního slovního druhu `BEFORE_DASH`, který je automaticky přiřazen slovu nacházejícímu se před pomlčkou a tokenu, který reprezentuje pomlčkou mezi těmito slovy (tokeny *high* a –). U slova za pomlčkou zůstane slovní druh zachován, ale interně se k tomuto tokenu uloží informace o tom, že je součástí slovního spojení. Při vyhledávání definičního slova jsou slova se slovním druhem `BEFORE_DASH` přeskakovány. Pokud se však výsledkem vyhledávání definičního slova stane slovo, které se nachází za pomlčkou, bude toto slovo spojeno s předcházejícími tokeny tak, aby výsledkem byla celá trojice tokenů.

- Podstatná jména, která obsahují přivlastňovací apostrof (např. *children's*), spaCy rozdělí do dvou tokenů, u kterých opět bude vyhodnocen slovní druh zvlášť:

Token:	children	's
Slovní druh:	Noun	Part

  toto je pro vyhledávání definičních slov obzvlášť nevhodné, neboť se tento typ slova může nacházet před hledaným slovem a poté by bylo místo skutečného definičního slova vráceno toto slovo. Tento problém je řešen změnou slovního druhu obou tokenů na přídavné jméno, což povede k přeskočení těchto slov.

- Některá přídavná jména spaCy označí za podstatná jména. Například v první větě článku Documentary film *„Donald Ervin Knuth is an American computer scientist...“* spaCy určí slovní druhy slov *American*, *computer* a *scientist* následovně:

Token:	American	computer	scientist
Slovní druh:	Adjective	Noun	Noun

  správně by však slovo *computer* mělo být přídavné jméno. Tato situace je řešena tak, že pokud spaCy nalezne dvě nebo více podstatných jmen za sebou (aniž by mezi nimi byl jiný token, např. čárka nebo spojka *and*), bude slovní druh všech těchto podstatných jmen kromě posledního změněn na přídavné jméno. Obdobně se postupuje, pokud je nalezeno podstatné jméno, za kterým následuje přídavné jméno nebo číslovka.

Určení slovních druhů v česky psaných větách

Aktuální verze spaCy však neobsahuje přímou podporu zpracování vět psaných v češtině. Tato knihovna sice nabízí možnost vytvoření a natrénování nového statistického modelu [8],

nicméně jelikož se ze všech funkcí, které tato knihovna nabízí, využívá především detekce slovních druhů a již v anglicky psaných větách knihovna občas nesprávně rozliší podstatné jméno od přídavného, bude k detekci slovních druhů v češtině použit jiný přístup.

Detekce slovních druhů v češtině je proti angličtině o něco jednodušší, neboť čeština používá různé tvary slov pro jednotlivé slovní druhy, zejména podstatná a přídavná jména, jejichž odlišení je pro nalezení definičního slova stěžejní, lze rozlišit na základě pouze na základě daného slova bez dalšího kontextu (např. *jaro* vs. *jarní*). Existují výjimky, např. slovo *vedoucí* může být jak podstatné jméno, tak i přídavné jméno, nicméně tento problém lze vyřešit obdobně jako v angličtině detekcí velkého počtu podstatných jmen za sebou.

K detekci slovních druhů psaných v češtině lze tedy použít libovolný **slovník**, který u jednotlivých slov obsahuje jejich slovní druhy. Pro tuto práci jsem zvolil slovník **Wiktionary (česky Wikislovník)** [28], což je projekt provozovaný Wikipedia Foundation a jedná se podobnou encyklopedii jako je Wikipedia (obě encyklopedie jsou provozovány na stejném počítačovém systému), kde každý článek obsahuje definici jednoho slova. Stejně jako v případě Wikipedie je hlavní předností tohoto slovníku jeho otevřená licence, možnost jeho editace kýmkoliv na světě a široká základna editorů (a na rozdíl od slovníků, které používají textové editory ke kontrole pravopisu, obsahuje slovní druhy jednotlivých slov).

Stejně jako je tomu v případě Wikipedie je však obsah tohoto slovníku zapsán ve speciálním formátovacím jazyce MediaWiki (viz kapitola 2.3), který nebyl původně navržen ke strojové extrakci informací. Nicméně proti Wikipedii je zde situace jednodušší, neboť způsob uložení informací je v článcích v rámci projektu jednotný. Extrakce dat z tohoto slovníku probíhá následovně:

- Název definovaného slova odpovídá názvu článku (např. článek *kočka* definuje význam slova *kočka*).
- V hodnotách parametrů šablon *adjektivum*, *stupňování*, *substantivum*, *sloveso* a *zájmeno* (akceptují se i šablony s podobnými názvy, např. *Adjektivum (cs)*) lze nalézt další tvary definovaného slova (*kočky*, *kočce*, *kočku*, *kočkami* apod.).
- Možné slovní druhy tohoto slova lze nalézt v kapitolách 3. úrovně příslušného článku (např. pokud se v článku nachází kapitola *podstatné jméno*, popisované slovo může být podstatné jméno).

Výsledkem této extrakce informací z Wikislovníku je soubor v následujícím formátu:

```
znechutil       5
znechutím       5
znechutili      5
z               7
tryskem         1, 6
reklamní        2
...
```

kde každý řádek tohoto souboru obsahuje informace o jednom slově. První sloupec tohoto seznamu obsahuje název slova a druhý sloupec jeho slovní druh[3]. Pokud některé slovo může mít v závislosti na kontextu více slovních druhů (např. slovo *tryskem* může být

[3] 1 = podstatné jméno, 2 = přídavné jméno, 3 = zájmeno, 4 = číslovka, 5 = sloveso, 6 = příslovce, 7 = předložka, 8 = spojka, 9 = částice, 10 = citoslovce

jak podstatné jméno, tak i příslovce), budou ve druhém sloupci uvedeny všechny vyhovující slovní druhy. Jednotlivé sloupce jsou od sebe odděleny tabulátorem.

Aktuální verze[4] tohoto slovníku obsahuje 99 264 článků, ze kterých bylo výše popsaným způsobem extrahováno celkem 556 tisíc tvarů slov. Cca třetina těchto slov jsou překlady slov do jiných jazyků. Tyto překlady nejspíše nebudou při vyhledávání definičního slova příliš užitečné, nicméně jejich přítomnost v seznamu známých slov žádné komplikace nezpůsobí. Celkem tedy bylo z Wikislovníku extrahováno cca $556380 \cdot 0,63 \doteq 350000$ tvarů českých slov.

Takto vytvořený slovník pokryl 76,6 % slov, které se nacházejí v prvních větách článků. Slova, která nejsou v tomto slovníku nalezena, zčásti tvoří zkratky, názvy míst či římské číslice (např. *km*, *m*, *př*, *III*, *sv* a *Budějovice*). Zbytek jsou slova, která zatím žádný přispěvatel do Wikislovníku neuložil (např. *reprezentant*, *studiové*, *provincii* a *sídlící*). Seznam slov, které se v prvních větách článků vyskytují nejčastěji, s informací, zda tyto slova byla nalezena ve slovníku Wiktionary, je uveden v příloze B (v seznamu je uvedeno cca 30 nejčastějších slov, které ve slovníku nalezeny byly a cca 30 nejčastějších slov, které ve slovníku nalezeny nebyly).

Pro případ, že se nebude ve slovníku Wiktionary nacházet některé slovo, které se v prvních větách článků vyskytuje často, byl vytvořen uživatelský slovník, pomocí kterého lze význam těchto slov ručně dodefinovat. Formát souboru s tímto slovníkem je stejný, jako je výše definovaný soubor s extrahovanými slovy ze slovníku Wiktionary. Pokud ani v tomto slovníku nebude slovo nalezeno, bude slovní druh slova odhadnut na základě slov ve slovníku s totožnou koncovkou (např. slova končící *-ní* nejspíše budou přídavná jména).

6.3 Rozšíření algoritmu pro vyhledání definičních slov

Výše popsaná část systému je dostatečná k tomu, aby u článků, které mezi slovesem *být* a hledaným definičním slovem obsahují pouze přídavná jména, definiční slovo spolehlivě našla. V této podkapitole budou popsány některá implementovaná rozšíření tohoto systému, která zvyšují jeho úspěšnost i u dalších článků. Tato rozšíření byla implementována na základě analýzy dosavadních výsledků tak, aby byl nalezen vyšší počet definičních slov a došlo k odstranění chybně nalezených slov.

6.3.1 Vyhledání skupin slov

U některých článků lze nalézt více než jedno definiční slovo. Například ve větě z článku A:

> „A is the first **letter** and the first **vowel** of the ISO basic Latin alphabet.“

lze nalézt definiční slova *letter* a *vowel*. Za tímto účelem jsou v první větě vyhledávány tzv. **skupiny slov**, které obsahují několikanásobné větné členy v této větě. Skupina slov je posloupnost několika tokenů v první větě, která se skládá pouze z podstatných jmen, přídavných jmen, spojek (včetně čárky) a členů (určitého nebo neurčitého). K nalezení těchto skupin slov lze použít zásobníkový automat definovaný gramatikou s následujícími pravidly:

```
[start] -> [part] [commas] [conjunctions]
[part] -> noun
```

[4] Jedná se o verzi z dubna 2018

```
[part] -> noun [non-noun] [part]
[part] -> [non-noun] [part]
[part] -> ε
[non-noun] -> adjective
[non-noun] -> article
[commas] -> , [part] [commas]
[commas] -> ε
[conjunctions] -> conjunction [part] [conjunctions]
[conjunctions] -> ε
```

kde `noun` je podstatné jméno, `adjective` je přídavné jméno, `article` je člen (jedno ze slov *a*, *an*, *the*; jen v případě anglicky psaných vět) a `conjunction` je spojka (nezahrnuje čárku, jedno ze slov *and*, *or* (v případě anglicky psaných vět) nebo *a*, *i*, *ani*, *nebo*, *či* (u česky psaných vět)). Mezi základní rysy tohoto automatu patří především:

- Pokud se v rámci skupiny nachází spojka (např. *and*), nesmí se za ní nacházet čárka (,) – např. skupina *„dogs, cats and mice“* je v pořádku, zatímco skupina *„dogs and cats, because ...“* nikoliv.
- V rámci skupiny se nemohou nacházet dvě podstatná jména za sebou. V případě anglicky psaných vět tato situace nastat nemůže (slovní druh prvního ze slov by byl změněn na přídavné jméno). V češtině tato situace nastat může, ale poté se nejedná o skupinu slov – například ve větě *„Labe je jednou z* ***největších řek a vodních cest*** *Evropy.“* se za sebou nachází dvě podstatná jména *cest* a *Evropy*. Slovo *Evropy* ovšem do hledané skupiny několikanásobných větných členů nepatří.

Kromě toho, aby nalezená skupina vyhovovala výše definovanému jazyku, je ještě vyžadováno, aby výsledná skupina obsahovala alespoň jednu spojku nebo čárku mezi slovy. Pokud je i tato podmínka splněna, bude daná skupina slov vytvořena.

Jakmile bude nalezeno definiční slovo obvyklým způsobem, bude zkontrolováno, zda toto slovo patří do některé skupiny slov. Pokud ano, budou jako definiční slova vrácena všechna podstatná jména, která se v této skupině nachází.

6.3.2 Skupina slov obsahující of nebo pro

Další situací, kterou je třeba při vyhledávání definičního slova řešit, jsou věty, které za nalezeným definičním slovem obsahují buď slovo *of* nebo *for* v případě anglicky psané věty, nebo slovo *pro* v případě česky psané věty, například:

1. An abbey is a **complex of buildings** used by...
2. The Austrian School is a **school of** economic **thought** that is...
3. Old Glory is a **nickname for** the **flag** of the United States.
4. The European Central Bank is the central **bank for** the **euro** and administers monetary policy of the eurozone...
5. Postupimská dohoda je **název pro** Závěrečný **protokol** postupimské konference...

V některých z těchto případů se hledané definiční slovo nachází před slovem *of* / *for* / *pro* (věty 2 a 4), zatímco v jiných případech se hledané slovo nachází až za tímto slovem (věty 1, 2 a 5). Pokaždé, kdy se za nalezeným definičním slovem nachází jedno z této trojice slov, pokusí se systém za tímto slovem nalézt druhé definiční slovo. K vyhledání tohoto druhého slova bude použit stejný algoritmus, který byl použit k nalezení prvního definičního slova, kde slovo *of* nebo jeho ekvivalent bude považováno za tvar slovesa *být*. V případě anglicky psané věty bude za tímto slovem navíc tolerována přítomnost určitého nebo neurčitého členu, což není běžná situace při standardním vyhledávání definičního slova.

Pokud nebude za slovem *of* či jeho ekvivalentem nalezeno žádné nové definiční slovo, zůstane zachováno původně nalezené definiční slovo. V opačném případě záleží výběr definičního slova na tom, které z trojice slov se nacházelo za původně nalezeným definičním slovem. V případě anglicky psaných vět bude porovnána pravděpodobnost výskytu jednotlivých slov dle údajů získaných z knihovny spaCy [7], kdy v případě skupiny obsahující slovo **of** bude vráceno slovo s nižší pravděpodobností výskytu a u skupin obsahující slovo **for** bude vráceno slovo s vyšší pravděpodobností výskytu. Použití tohoto způsobu bylo ověřeno na několika desítkách vět obsahující tuto skupinu slov. V případě českých vět bude vždy vráceno druhé slovo. Tento přístup byl zvolen proto, že více než 75 % slov nacházejících se před slovem **pro** jsou obecné výrazy typu *označení*, *termín*, *název*, *zkratka* apod.

6.3.3 Rozšířené definiční slovo

Pokud je hledaná kategorie stránek příliš úzká (obsahuje malé množství článků), mohou být nalezená definiční slova příliš obecná na to, aby dokázaly definovat hledanou entitu stránek. Například, pokud hledanou entitou jsou antická města, nepřinese příznak, zda definiční slovo daného článku je *town* žádnou informaci o tom, zda daný článek do této entity patří, či nikoliv. Místo toho by v tomto případě bylo lepší, kdyby systém na vyhledání definičních slov vrátil jako definiční slovo spojení *ancient town* (toto by neplatilo, kdyby hledanou entitou článků byly všechny města – v tomto případě by byl k pokrytí stejné sady článků vyšší počet příznaků, přičemž lze předpokládat, že by systém nemusel tyto příznaky nalézt všechny).

Řešením této situace je vrácení více variant stejného definičního slova. Jedna varianta bude obsahovat definiční slovo i s přídavnými jmény, které se před ním nacházely a druhá varianta vrátí pouze nalezené podstatné jméno, jako tomu bylo doposud. Poté může být vygenerován takový příznak, který bude co nejpřesněji pokrývat hledanou sadu článků.

6.4 Podpora dalších jazyků

V předchozích podkapitolách bylo uvedeno, jak byl systém na nalezení definičního slova implementován pro česky a anglicky psané věty. V této podkapitole bude popsáno, jakým způsobem by mohl být tento systém rozšířen i o podporu dalších jazyků, například němčiny nebo slovenštiny.

Prvním krokem k podpoře cizího jazyka je v modulu `first_sentence_utils` vytvořit potomka třídy `NounFinder`, stejně jako jsou nyní vytvořeny třídy `CzechNounFinder` a `EnglishNounFinder`. V rámci této nově vytvořené třídy bude implementována většina úkonů souvisejících se získáním definičního slova v novém jazyce.

Nejdůležitější úlohou, kterou je při implementaci podpory dalšího jazyka třeba splnit, je podpora **detekce slovních druhů** jednotlivých slov v daném jazyce. Pokud je pro daný jazyk k dispozici statistický model pro spaCy, bude nejvýhodnějším řešením použití této

knihovny (lze vyjít z implementace pro angličtinu). V opačném případě je třeba zvážit natrénování knihovny spaCy pro nový jazyk, použití jiné knihovny pro zpracování přirozeného textu, použití slovníku jako je Wiktionary, nebo implementaci zcela jiného řešení. Jako rozhraní mezi novým systémem pro identifikaci slovních druhů a vyhledávačem definičních slov slouží metody začínající `is_` (`is_noun(word)`, `is_adjective(word)`,...) ve třídě `NounFinder`.

Poté je třeba implementovat metodu `find_is_a`, jejíž úkolem je v první větě nalézt sloveso *být* v libovolném tvaru. Tímto je základní podpora definičních slov v cizím jazyce dokončena. V případě potřeby je možné přetížit i další metody třídy `NounFinder`, které umožňují implementaci dalších funkcí, jako je možnost prohledání slov v okolí nalezeného definičního slova a případně jeho změna, úprava systému generování skupin slov, dodatečná změna slovních druhů po nalezení všech slovních v celé větě apod.

6.5 Zhodnocení nalezených definičních slov

Pro český i anglický dump Wikipedie byl vygenerován seznam všech nalezených definičních slov. Na české Wikipedii bylo nalezeno celkově 910 302 definičních slov (definiční slova, která se nacházela ve více článcích, byla započítána vícekrát), přičemž se jednalo o 181 855 různých slov, z nichž 48 561 slov bylo nalezeno alespoň dvakrát a 8 556 slov alespoň desetkrát. Nejčastějším definičním slovem je slovo politik, které bylo nalezeno 11 348krát.

Na anglické Wikipedii bylo nalezeno 12 188 204 definičních slov, přičemž se jednalo o 1 486 485 různých slov, z nichž 390 131 slov bylo nalezeno alespoň dvakrát a 72 841 slov alespoň desetkrát. Nejčastějším definičním slovem je slovo village, které bylo nalezeno 218 536krát.

Seznam nejčastější definičních slov je uveden v příloze C.

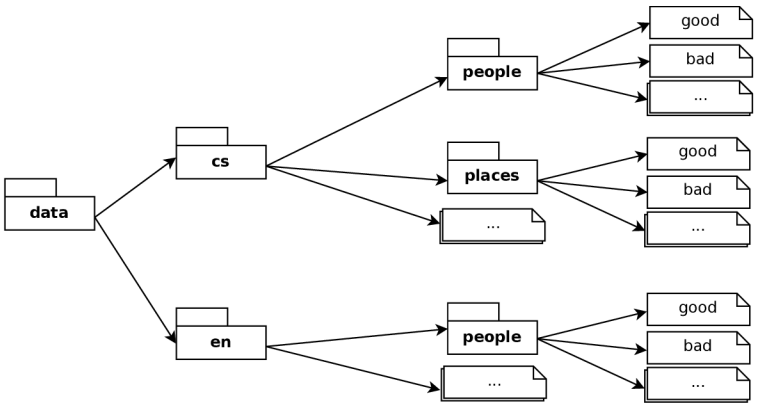
Kapitola 7

Implementace

V této kapitole budou popsány některé aspekty implementace systému pro identifikaci entit článků na Wikipedii. Nejprve bude v podkapitole 7.1 popsána adresářová struktura se vstupními a výstupními soubory. Poté bude v podkapitole 7.2 uveden účel jednotlivých modulů vytvořených v tomto systému. Podkapitola 7.3 se bude zabývat spuštěním vytvořeného programu a v podkapitole 7.4 probrán způsob, jakým se systém rozhoduje, které příznaky zvolit pro účinnou identifikaci dané entity.

7.1 Vstupní a výstupní soubory

Vstupní i výstupní soubory, se kterými bude tento systém pracovat, se sice mohou nacházet kdekoliv, ale k udržení jisté míry přehlednosti je vhodné použít následující adresářovou strukturu. Při použití této struktury nebude nutné k jednotlivým vstupním souborům nastavovat cestu pomocí přepínačů, ale bude možné využít jejich výchozí hodnoty.



Obrázek 7.1: Doporučená adresářová struktura se vstupními a výstupními soubory

Tato adresářová struktura obsahuje složku `data/`, ve které se nacházejí podsložky pro jednotlivé jazykové mutace. Složka každého jazyka pak obsahuje podsložky s vyhledávanými entitami a textové soubory `wiki-pages-articles.xml` a `test-is-a` a volitelně může obsahovat soubory `wiktionary` a `wiktionary_custom`. První z uvedených souborů obsahuje vstupní dump soubor dané jazykové mutace Wikipedie. Soubor `test-is-a` obsahuje data

k automatickému testu podsystému na vyhledávání definičních slov[1]. Zbývající dva soubory obsahují data načtená ze slovníku Wiktionary, pokud se tento slovník používá při identifikaci slovních druhů slov v první větě článků (jeden z těchto souborů obsahuje strojově zpracovaná data z Wiktionary, druhý ručně doplněné záznamy; viz kapitola 6.2.2).

Ve složce s konkrétní identifikovanou entitou se pak nachází tyto soubory:

- seznam vzorových stránek, které patří do dané entity (`good`),
- seznam vzorových stránek, které nepatří do dané entity (`bad`),
- vygenerovaný seznam příznaků pro identifikaci této entity (`rules` nebo `flags`),
- vygenerovaný seznam stránek patřící do této entity (`wp_list`),
- vygenerovaný seznam stránek nepatřící do této entity (`wp_regular`) a
- vygenerované statistiky o použití jednotlivých příznaků při identifikaci entit (`stats`).

7.2 Zdrojové soubory systému

Tento systém na určování typů článků byl implementován v interpretovaném jazyce Python 3 [20]. Vzhledem k rozsahu této úlohy byly zdrojové soubory systému rozděleny do několika modulů a tříd. Tyto zdrojové soubory jsou na přiloženém paměťovém médiu umístěny ve složce `src/`. Zde bude popsána úloha jednotlivých modulů tohoto systému:

- `wikipedia_extract.py` – hlavní modul programu, který zpracuje parametry z příkazové řádky a spustí další moduly
- `page.py` – obsahuje třídu `Page` reprezentující konkrétní článek na Wikipedii
- `modes.py` – definuje logiku systému pro jeho jednotlivé režimy
- `rules.py` – definice jednotlivých typů příznaků
- `rules_set.py` – obsahuje pomocnou třídu pracující s množinou příznaků
- `first_sentence_utils.py` – podsystém pro vyhledávání definičních slov
- `wiki_dict.py` – rozhraní pracující se záznamy ve slovníku Wiktionary
- `shared_data.py` – zajišťuje výměnu dat mezi souběžně běžícími procesy tohoto systému
- `utils.py`, `page_utils.py` a `color.py` – různé pomocné funkce
- `WikiExtractor.py` – převzatý modul pro nalezení první věty článku [13]

[1] Test modulu na vyhledávání definičních slov je založen na porovnání skriptem nalezených definičních slov a ručně zadaných očekávaných výsledků – jedná se tedy o způsob jednotkového testování tohoto modulu, kde jako vstupní data byly použity ty věty, jejichž identifikace v průběhu vývoje činila problémy (aktuální úspěšnost tohoto testu je 100 %).

7.3 Vybrané režimy skriptu a jeho spuštění

Zde popsaný systém nebude při každém jeho spuštění vykonávat totožnou operaci. Například, uživatel tohoto systému může požadovat, aby skript vypsal názvy všech stránek, které patří do hledané entity článků. V jiném případě může být požadováno, aby skript vypsal seznam definičních slov jednotlivých článků. V jiném případě může být požadováno, aby skript vygeneroval nový seznam vhodných příznaků. K odlišení, jakou operaci má skript provést, byly zavedeny tzv. **režimy skriptu**.

Každý režim skriptu je definován jednoznačným textovým identifikátorem a vlastním potomkem třídy `Mode`. Tato třída slouží jako základní rozhraní mezi jádrem skriptu a aplikační logikou odpovídajícího režimu skriptu. Nejdůležitějšími metodami této třídy jsou:

- `init()` – slouží k inicializaci proměnných před zahájením zpracování jednotlivých článků, např. k načtení seznamu příznaků, slovníku Wiktionary apod.
- `iter(page)` – zpracování konkrétního článku. Tato metoda bude zavolána tolikrát, kolik vstupní dump soubor obsahuje vyhovujících[2] článků. Tato třída může být spuštěna paralelně pro více různých článků v několika procesech.
- `finish()` – vyhodnocení provedených úkonů po zpracování všech článků. Např. výběr příznaků, které pokryly velké množství vzorových článků, při jejich generování.

Jako spouštěč tohoto systému slouží skript `src/wikipedia_extract.py`, jehož prvním parametrem bude identifikátor požadováno režimu skriptu (např. `list` – výpis článků patřící do hledané entity, `regular` – výpis článků nepatřící do hledané entity, `make` – vygenerování příznaků, `is-a` – výpis definičních slov, `wiktionary` – extrakce slov z Wiktionary, nebo `parse-stats` – zpracování statistik o využití jednotlivých příznaků). Kromě režimu lze nastavit další parametry:

- `-e entity` – název entity, se kterou se bude pracovat (výchozí: *people*)
- `--lang code` – kód jazykové mutace Wikipedie, se kterou bude systém pracovat (výchozí: *en*)
- `--dump filepath` – cesta k dump souboru Wikipedie (výchozí: *data/%lang/wikipages-articles.xml*, kde *%lang* je hodnota parametru `--lang`)
- `-p number` – počet procesů, který má být vytvořen k paralelnímu zpracování jednotlivých článků
- `title1 [title2 [...]]` – seznam článků, které mají být zpracovány (při nastavení tohoto parametru nebude zpracován celý dump soubor). Slouží především k ladícím účelům.

Výše uvedený seznam není kompletní, některé režimy skriptu akceptují další parametry, které jsou specifické jen pro daný režim, například režim `regular` definuje další přepínač `--no-redirects`, který do výsledného seznamu článků nepatřících do hledané entity nezahrne stránky, které jsou přesměrováním na jinou stránku a režim `make` generující seznam příznaků obsahuje parametry `-g` a `-b`, které definují cestu k souborům se seznamy vzorových stránek, na kterých se bude tento systém učit.

[2] standardně všechny stránky nacházející se ve jmenném prostoru `0` (články), lze překrýt

K zjednodušení zadávání těchto parametrů do příkazové řádky byl vytvořen skript `src/bash_complete.sh`. Pokud bude tento skript spuštěn během inicializace shellu (například jeho spuštěním v souboru `~/.profile`), bude v daném shellu vytvořen alias `wi`, který spustí hlavní skript tohoto systému *(předtím je nutné ve skriptu `bash_complete.sh` upravit absolutní cestu k tomuto systému)*. V tomto skriptu je navíc implementováno doplňovaní parametrů pomocí tabulátorů, tj. po napsání `wi m` a stisknutí tabulátoru bude zadaný příkaz doplněn na `wi make` apod.

7.4 Výběr vhodných příznaků pro identifikaci typů entit

Další důležitou vlastností tohoto systému, která ještě nebyla popsána, je výběr příznaků, které budou použity k určování typů entit. Vstupními daty pro tento systém je totiž seznam několika vzorových článků, které do hledané entity patří a seznam několika vzorových článků, které do této entity nepatří. Tedy například vstupem pro vyhledávání českých osob může být informace, že články *Eugen Wüster*, *Kazimierz Bein*, *Miroslav Horníček* a *Vladimír Blucha* pojednávají o osobě a články *Astronomie*, *Periodická tabulka*, *Kokpit* a *Nobelova cena za fyziku* o osobě nepojednávají. Ovšem k rozhodnutí, zda článek, který není ani na jednom z těchto dvou seznamů, pojednává o osobě, je potřeba vytvořit seznam příznaků, na jehož základě bude moct systém toto rozhodnutí učinit.

Vygenerování tohoto seznamu příznaků se skládá ze dvou fází. Nejprve je třeba projít všechny články, jejichž názvy se nachází v jednom z výše uvedených seznamů a **nalézt v nich všechny příznaky**, které mohou daný článek identifikovat. Jelikož tento systém pracuje s konečným počtem typů příznaků (podporované typy příznaků byly definovány v kapitole 5) a zdroje, ze kterých jsou tyto příznaky generovány jsou taktéž konečné (článek nemůže obsahovat nekonečno kategorií apod.), nepředstavuje vygenerování všech přijatelných příznaků pro konkrétní článek žádný problém. Například pro článek *Kazimierz Bein* bude vygenerováno celkem 39 příznaků, jejichž seznam je uveden v příloze D.

Poté, co budou vygenerovány příznaky všech článků, u nichž byl definován očekávaný výsledek identifikace, je potřeba vybrat ty příznaky, které přinesou pro budoucí identifikaci ostatních článků nějakou přidanou hodnotu. Například příznak „článek obsahuje kategorii *Úmrtí [číslo]*“, který byl mezi osmi vzorovými články nalezen celkem třikrát, a to vždy u osob (*Eugen Wüster*, *Kazimierz Bein* a *Miroslav Horníček*), bude při identifikaci osob velmi užitečný; zatímco příznak „článek obsahuje šablonu *Portály*“, příliš užitečný nebude. Sice byl tento příznak nalezen dokonce v šesti vzorových článcích (všechny články kromě *Kazimierz Bein* a *Kokpit*), ale ve třech případech je jednalo o články osob a ve třech případech se o články osob nejednalo, a tedy tento příznak nepřináší žádnou informaci o tom, zda články, které tomuto příznaku vyhovují, pojednávají o osobách, nebo nikoliv.

K rozhodnutí, které příznaky jsou užitečné a které ne, bude využita tzv. **asociační analýza** (angl. association analysis) [22], která slouží k nalezení vztahů skrytých ve velkých sadách dat. Touto sadou může být právě seznam nalezených příznaků u vzorových článků, viz příklad v tabulce 7.1.[3]

[3]Každý řádek této tabulky obsahuje informace o nalezených příznacích pro jednotlivé články. Příznaky, které daný článek obsahuje, jsou označeny znakem X. Význam příznaků: **spis.** = čl. obsahuje kategorii *Čeští spisovatelé*, **Muži** = čl. obsahuje kategorii *Muži*, **NK** = čl. obsahuje šablonu *NK ČR*, **jméno** = čl. obsahuje šablonu s pojmenovaným parametrem *jméno*, **odkazy** = čl. obsahuje nadpis *Externí odkazy*, **režisér** = čl. obsahuje definiční slovo *režisér*. **Os.** = článek pojednává o osobě, **!Os.** daný článek o osobně nepojednává (hodnoty sloupců Os. a !Os. lze zjistit na základě toho, do kterého ze vstupních souborů byl článek umístěn).

Tabulka 7.1: Příklad datové sady pro asociační analýzu (vysvětlivky sloupců v poznámce pod čarou)

Článek	**spis.**	**Muži**	**NK**	**jméno**	**odkazy**	**režisér**	**Os.**	**!Os.**
E. Wüster		X	X				X	
K. Bein		X					X	
M. Horníček	X	X	X	X		X	X	
V. Blucha	X	X	X				X	
Astronomie			X					X
Per. tabulka				X	X			X
Kokpit					X			X
Nob. cena...					X			X

Asociační analýza dokáže v těchto datech nalézt často se opakující vzory. Například se může jednat o informaci, že všechny články, které obsahují příznak *režisér* (článek obsahuje definiční slovo *režisér*), také vyhovují příznaku *Muži* (článek obsahuje kategorii *Muži*). Tyto vzory se nazývají **asociační pravidla** (angl. association rules). Tyto asociační pravidla se zapisují ve formě implikace, např. zde uvedené pravidlo by mohlo být zapsáno jako `{režisér} => {Muži}`.

Asociační pravidla lze definovat pro libovolnou dvojici příznaků, a to i v případě, že dané pravidlo není splněno pro všechny vzorové články. Například pravidlo `{NK} => {Muži}` je splněno u článků *E. Wüster*, *M. Horníček* a *V. Blucha* a není splněno u článku *Astronomie*. Pro další použití bude pochopitelně třeba vybrat ty asociační pravidla, která budou platit pokud možno vždy, nebo alespoň u většiny článků – a to jak u vzorových stránek, tak i u těch, které ve vzorových souborech nejsou. K jejich výběru budou využity standardní metriky *support*, *confidence* a *lift*, které budou popsány za okamžik.

Obecně mohou asociační pravidla na jedné straně implikace obsahovat i více než jeden příznak. Například pravidlo `{Muži, NK} => {spis.}` říká, že články, které současně vyhovují příznakům *Muži* a *NK*, budou (pravděpodobně) také vyhovovat příznaku *spis*. Nicméně pro účely identifikace typů entit článků budou relevantní jen ty pravidla, které (na pravé straně implikace) přinášejí informaci o tom, zda daný článek patří do hledané entity článků, tj. v tomto případě se jedná o pravidla ve tvaru `{cokoliv} => {Os.}` nebo `{cokoliv} => {!Os.}`. Dále lze vyjít z toho, že jednotlivé příznaky jsou na sobě nezávislé, tedy že použití více příznaků najednou nemá jiný význam, než kdyby byly použity samostatně. Nelze například očekávat, že pokud články, které obsahují kategorii *Muži*, pojednávají o osobě a články, které obsahují kategorii *Čeští spisovatelé*, pojednávají o osobě, tak by články, které obsahují kategorie *Muži* a *Čeští spisovatelé* současně, mohly nepojednávat o osobě. Z tohoto důvodu nemá smysl generovat asociační pravidla, které by na levé straně obsahovaly více než jeden příznak.

Jak již bylo uvedeno, k ověření spolehlivosti jednotlivých pravidel budou využity standardní metriky *support*, *confidence* a *lift*, které budou níže definovány. Nechť N je celkový počet vzorových článků (jak osob, tak i ostatních), X a Y představují množinu příznaků testovaných na levé resp. pravé straně pravidla `{X} => {Y}` a funkce $\sigma(A)$ vrací počet vzorových článků vyhovující všem příznakům v množině A. Potom jsou metriky support, confidence a lift definovány následovně: [22]

- Metrika **support** $s(X \to Y) = \frac{\sigma(X \cup Y)}{N}$ *(vzorec 5.1 v [22])* definuje podíl počtu článků, u kterého bylo dané pravidlo splněno. Například pro pravidlo `{NK} => {Os.}` je tato

hodnota rovna $\frac{3}{8}$, neboť existují 3 články (*E. Wüster*, *M. Horníček* a *V. Blucha*) z celkových 8, kde bylo toto pravidlo splněno.

- Metrika **confidence** $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$ *(vzorec 5.2 v [22])* definuje přesnost daného asociačního pravidla. Hodnotou této metriky je podíl počtu článků, u kterých bylo (celé) toto pravidlo splněno, ku počtu článků, kde byla splněna levá strana implikace tohoto pravidla. Například pravidlo `{NK} => {Os.}` říká, že články, které vyhovují příznaku *NK*, by měly být osoby. Ve vzorové sadě dat však existují celkem 4 články, které příznak *NK* obsahují, z toho ve 3 případech se jedná o osobu a v 1 případě nikoliv. Hodnotou *confidence* tohoto pravidla bude $\frac{3}{3+1} = \frac{3}{4}$.

- Metrika **lift** $l(X \to Y) = \frac{s(X \cup Y)}{s(X) \cdot s(Y)} = \frac{s(X \to Y)}{s(X) \cdot s(Y)}$, *(vzorec 5.5 v [22])* kde $s(x)$ je funkce support pro dané pravidlo nebo příznak (pro příznak se hodnota support počítá pomocí $s(x) = \frac{\sigma(x)}{N}$). Například pro pravidlo `{jméno} => {Os.}` je hodnotou této metriky $\frac{\frac{1}{8}}{\frac{2}{8} \cdot \frac{4}{8}} = 1$, kde $\frac{1}{8}$ je podíl článků obsahující současně příznaky *jméno* a *Os.* (tj. podíl článků vyhovující tomuto pravidlu), $\frac{2}{8}$ je podíl článků obsahující příznak *jméno* a $\frac{4}{8}$ je podíl osob mezi vzorovými články. Výhodou metriky *lift* je to, že zohledňuje skutečnost, že počet článků o osobách (nebo jiné extrahované entity) a počet jiných článků ve vzorových souborech nemusí být totožný.

Při rozhodování, zda daný příznak má být zařazen do seznamu příznaků, které budou použity při identifikaci článků, budou z daného příznaku vygenerovány asociační pravidla `{příznak} => {hledaná entita}` a `{příznak} => {není hledaná entita}`. Jelikož je množina `{není hledaná entita}` doplňkem množiny `{hledaná entia}`, bude součet hodnot *confidence* obou vygenerovaných asociačních pravidel vždy roven 1. Z těchto pravidel se bude dále pracovat pouze s tím, které má vyšší hodnotu *confidence* (v případě, že obě pravidla budou mít totožnou hodnotu *confidence* ($c_1 = c_2 = \frac{1}{2}$), tj. polovina článků obsahující tento příznak pojednává o hledané entitě a druhá polovina nikoliv, budou obě pravidla zahozeny).

Aby mohlo být dané pravidlo resp. příznak použito, musí jeho metriky *support*, *confidence* a *lift* překročit minimální stanovené hodnoty těchto metrik. Tyto hodnoty jsou závislé na konkrétním typu příznaku a byly zvoleny na povaze jednotlivých typů příznaků a na základě dosažených výsledků při generování seznamu článků patřících do dané entity. Například u příznaků porovnávající definiční slovo může být minimální vyžadovaná hodnota příznaku *support* nižší, než u příznaků porovnávající názvy použitých šablon, poněvadž článek na české Wikipedii průměrně obsahuje 1,05 definičních slov a 4,84 šablon. Zároveň také šablona, která byla použita ve více než jednom článku, je průměrně použita v 138,6 článcích, zatímco definiční slovo, které je použito ve více než jednom článku, bylo průměrně použito v 38,5 článcích[4].

7.5 Vyhodnocování příznaků při identifikaci článků

Při rozhodování, zda daný článek patří do hledané entity, se postupně prohledávají všechny příznaky, které byly nalezeny postupem uvedeným v minulé kapitole 7.4, přičemž výsledek

[4] 53,9 % definičních slov a 24,5 % šablon je použito jen v jediném článku. Tyto články nebyly započítány, poněvadž by silně zkreslily výsledný průměr. Nejčastější definiční slovo je *politik* a bylo použito v 11 349 článcích, a nejpočetnější šablona *Portály* obsahuje 225 695 článků.

identifikace je závislý jen na příznacích, které současně byly nalezeny v minulé kapitole, a které se nacházejí v prohledávaném článku. Tyto příznaky se porovnávají v pořadí dle nalezené metriky *confidence* jim odpovídajícího asociačního pravidla od nejvyšší po nejnižší, poněvadž pravidla s vyšší hodnotou *confidence* budou více spolehlivá.

Jelikož žádný příznak není stoprocentní, nebylo by vhodné rozhodovat výsledek identifikace článku na základě prvního splněného pravidla. Zároveň však jsou různé příznaky různě spolehlivé (viz standardní metriky), a tedy by rovněž nebylo vhodné vyhodnocovat všechny příznaky se stejnou vahou. V této práci bylo zvoleno vyhodnocení na základě prvních pěti nalezených příznaků, čím dojde k vyřešení obou zmíněných problémů a zároveň může být ušetřen strojový čas, protože se nemusí vždy vyhodnocovat všechny příznaky.

Pokud některý článek nebude vyhovovat žádnému příznaku, nebo bude nalezen stejný počet příznaků, které naznačují, že tento článek patří do hledané entity, jako příznaků, které naznačují, že tento článek nepatří do hledané entity, bude výsledkem jeho identifikace skutečnost, že daný článek nepatří do entity hledaných článků.

Kapitola 8

Výsledky

V této kapitole bude popsáno, na jakých typech entit byl tento systém testován, kolik tento systém nalezl článků daných entit, jaká je jeho úspěšnost a jak tato úspěšnost může být ovlivněna různými nastaveními tohoto systému. Také bude u některých typů entit porovnána úspěšnost tohoto systému s existujícími systémy popsanými v kapitole 3.

Všechny zde popsané výsledky vycházejí z posledního dumpu Wikipedie dostupného před odevzdáním práce, tj. z dumpu z května 2018.

8.1 Typy identifikovaných entit

Tento systém je schopen identifikovat libovolnou entitu na základě seznamu několika článků, které do hledané entity patří a seznamu několika článků, které do hledaného seznamu nepatří. Pro otestování výsledků tohoto systému je však nutné definovat konkrétní entity, na kterých bude tento systém testován.

Nejprve budou vyhledány všechny osoby na anglické a české Wikipedii. Tato entita byla vybrána proto, že obsahuje velké množství článků (přes jeden milion na anglické Wikipedii a sto tisíc na české Wikipedii). Poté bude tento systém otestován na entitách, které obsahují menší počet článků, konkrétně na státech a psích rasách.

8.2 Ověřování výsledků na základě náhodného výběru článků

Nejspolehlivější metodou, jak ověřit výsledky identifikace entity článků, je z vygenerovaného seznamu článků náhodně vylosovat několik článků[1] a manuálně zkontrolovat jejich správnost. Poté bude stejným způsobem vylosováno i několik článků ze seznamu článků nepatřících do hledané entity a ověřena správnost těchto dat.

Tento test bude prováděn na všech seznamech vygenerovaných v následujících kapitolách, kde budou uvedeny počty chyb, které byly tímto způsobem odhaleny. Seznamy náhodně vygenerovaných článků, na kterých byl tento test prováděn, lze nalézt v souboru `data/Seznamy.md` na přiloženém paměťovém médiu.

[1] Např. `cat wp_list | shuf | head` na unixovém systému nebo `wi run-script samples.sh wp_list` při nastaveném spouštěči `wi`

8.3 Identifikace osob

Nejprve budou identifikovány všechny osoby na české Wikipedii, kde vstupní soubor bude obsahovat 10 článků, které osobami jsou a 10 článků, které osoby nejsou.[2] V těchto článcích bylo celkem nalezeno 678 příznaků. Typy těchto příznaků jsou uvedeny v tabulce 8.1. Z toho 100 příznaků bylo nalezeno ve více než jednom článku, typy těchto příznaků jsou uvedeny v tabulce 8.2.

Tabulka 8.1: Počty příznaků nalezených v 10 + 10 článcích Wikipedie dle jejich typu

Typ příznaku:	člen	kat. (sh.)	kat (1. sl.)	kat. (posl. sl.)	nadpis
Počet článků:	1	120	63	99	60
Typ příznaku:	def. sl.	šablona	klíč šabl.	tit. (1. + 2. sl.)	tit. (závorka)
Počet článků:	21	81	227	4	2

Tabulka 8.2: Příznaky, které byly nalezeny ve více než jednom z 20 vzorových článků

Typ příznaku:	člen	kat. (sh.)	kat (1. sl.)	kat. (posl. sl.)	nadpis
Počet článků:	1	4	10	6	7
Typ příznaku:	def. sl.	šablona	klíč šabl.	tit. (1. + 2. sl.)	tit. (závorka)
Počet článků:	0	24	48	0	0

Z těchto 100 příznaků bylo vytvořeno 31 asociačních pravidel, které indikují, že daný článek pojednává o osobě a 69 asociačních pravidel, které indikují, že daný článek o osobě nepojednává. Z těchto příznaků byly vybrány ty, jejichž metrika *confidence* je alespoň 0,8 a metrika *lift* je alespoň 1 (minimální hodnota *support* byla omezena na 0,1 výběrem příznaků, které byly nalezeny alespoň u dvou článků). Těmto kritériím vyhovuje celkem 47 příznaků. Celkově bylo tímto způsobem označeno 164 528 článků za osoby, zbývajících 240 239 článků v dumpu osoby nebyly, nicméně chybovost této identifikace je velmi vysoká. Z náhodně vybraných 40 článků z vygenerovaného seznamu osob 16 článků (40 %) nepojednávalo o osobách. Ovšem z náhodně vybraných 40 článků z vygenerovaného seznamu článků, které nepojednávají o osobách, naprostá většina článků osobami skutečně nebyla, pouze jeden z těchto článků (2,5 %) byl ve skutečnosti osobou.

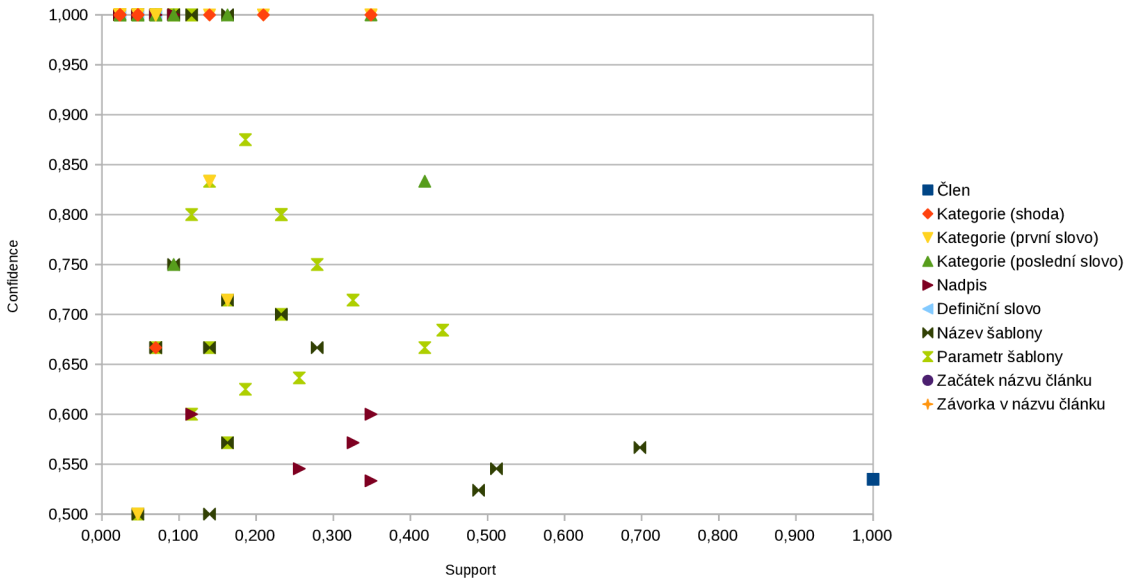
Aby byla chybovost při generování českých osob nižší, byla upravena **minimální požadovaná hodnota metriky *support*** tak, aby byla pro každý typ příznaku (shoda názvu kategorie, název šablony, definiční slovo apod.) jiná. Minimální požadovaná hodnota *support* byla pro každý typ příznaku nastavena na 70 % hodnoty příznaku stejného typu s nejvyšší hodnotou metriky *support*, a to nezávisle pro pozitivní a negativní příznaky. Požadavky na metriky *confidence* a *lift* zůstaly nezměněny. Celkem bylo tímto způsobem vygenerováno 32 pravidel a bylo nalezeno 89 358 článků osob. Z náhodně vybraných 100 článků ze seznamu osob všechny články pojednávaly o osobách. 7 ze 100 (7 %) náhodně vybraných článků ze seznamu ostatních článků (ne osob) ve skutečnosti pojednávalo o osobě. Je tedy patrné, že tato změna vedla ke zlepšení výsledků identifikace osob při zachování stejné sady 10 + 10 vzorových článků.

[2] Jedná se o články Hippomedón, Ferdinand z Ditrichštejna, Zarathuštra, Indiana Jones, Christoph Amberger, Kazimierz Bein, Oliver Palotai, Michail Vasiljevič Zimjanin, Ruslan Ponomarjov, Bohumil Zavadil (články o osobách) a Mezistátní utkání české hokejové reprezentace v sezóně 2002/2003, 1. září, 1307, Biologie, Aphrodite's Child, Geocaching, Evropská komise, Texaský masakr motorovou pilou (film, 2003), Měsíc (rozcestník), Strunatci (články, které nejsou o osobách).

Dále bude otestováno chování tohoto systému při použití **většího počtu vzorových článků**. Jako vzor bylo použito 46 pozitivních a 41 negativních článků, které byly vybrány tak, aby systém:

- nalezl i mystické postavy (např. Indiana Jones a Dagda (keltský bůh)),
- nalezl i články, které pojednávají o úzké skupině konkrétních lidí (např. Bratři Montgolfierové a Filémón a Baukis),
- nenalezl články, které pojednávají o široké skupině lidí (např. Vedlejší postavy v Percym Jacksonovi a 20 000 mučedníků z Nikomédie),
- nenalezl seznamy lidí (např. Seznam českých spisovatelů) a
- nenalezl články, jejichž část sice o osobě pojednává, ale celkově pojednávají o něčem jiném (např. Dášeňka čili život štěněte pojednává o knize, nikoliv o postavě).

Pro tuto sadu dat byly vygenerovány asociační pravidla obdobným způsobem jako v minulém bodě, ovšem byly zvýšeny požadované hodnoty *confidence* na 0,8, *lift* na 1,1 a *support* na 80 % nejvyšší dosažené hodnoty daného typu příznaku. V případě *support* bude rovněž tolerováno, pokud místo 80 % nejvyšší dosažené hodnoty bude hodnota rovna alespoň čísla *0,4*. Na obrázku 8.1 jsou znázorněny hodnoty *support* a *confidence* jednotlivých příznaků nalezených v těchto vzorových článcích. Tento graf ukazuje, proč není možné všem typům příznaků nastavit stejný minimální požadavek na hodnotu *support* – například metrika *support* příznaků typů *Nadpis* se pohybuje mezi 0,2 a 0,4, zatímco v případě příznaků typů *Název šablony* tato metrika běžně přesahuje hodnotu 0,5.



Obrázek 8.1: Závislost mezi hodnotami support, confidence a typem příznaku při identifikaci osob

Celkem bylo na základě této sady vzorových článků nalezeno 104 796 článků osob. Jak z vygenerovaného seznamu osob, tak z vygenerovaného seznam stránek, které nejsou osoby, bylo náhodně vybráno 160 stránek, u nichž byl výsledek identifikace ručně zkontrolován. Na

seznamu osob byla nalezena jedna stránka, která ve skutečnosti osobou nebyla, na seznamu ostatních článků žádná chyba zjištěna nebyla.

Dále byl obdobným způsobem vygenerován seznam osob **na anglické Wikipedii**. K tomuto účelu bylo použito 300 + 300 vzorových článků a stejné požadavky na metriky *support*, *lift* a *confidence* jako při poslední identifikaci českých osob (která vycházela z 46 + 41 vzorových článků). Tímto způsobem bylo nalezeno 1 384 324 článků osob. Z tohoto seznamu osob bylo náhodně vybráno 100 článků. Všechny z těchto článků byly osoby. Stejným způsobem bylo ze seznamu ostatních článků náhodně vybráno 100 článků, mezi kterými byla nalezena jedna osoba. DBpedie našla na anglické Wikipedii 1 818 074 osob, DBpedie nalezla 983 101 osob a YAGO nalezlo 1 010 473 osob. Ve srovnání s těmito systémy lze počet osob nalezených tímto skriptem považovat za přijatelný.

8.4 Identifikace států

Dále byla zvolena entita, která bude obsahovat relativně málo článků, a to na české Wikipedii, která obsahuje méně článků než anglická (v další kapitole bude identifikována jiná málo početná entita na anglické Wikipedii). Identifikování menších entit s použitím tohoto systému je o něco složitější, neboť je nutné přesně definovat, které články do hledané entity patří a které nikoliv.

Při náhodném výběru několika desítek vzorových článků nebude možné z této sady článků rozpoznat rozdíl mezi články o státech a články podobných entit (např. o krajích, městech, ostrovech apod.), které nemusí být při náhodném výběru článků zahrnuty do vzorového souboru. Nicméně po prvním vygenerování seznamů článků patřících a nepatřících do dané entity bude ihned patrné, které typy článků skript neidentifikuje správně. Tyto články je poté možné **doplnit do vzorových souborů** a spustit vygenerování příznaků a identifikaci článků znovu.

Tímto způsobem bylo nalezeno 1 538 článků o státech. Z vygenerovaných souborů bylo opět náhodně vylosováno několik článků, na kterých byla ověřena úspěšnost tohoto systému. 11 ze 50 (22 %) článků na seznamu států ve skutečnosti státy nejsou – jedná se většinou o podobné články, jako jsou Severoněmecký spolek a Valparaíso (region). Mezi 100 články na seznamu ostatních článků žádné státy nalezeny nebyly (což ovšem nebylo pravděpodobné vzhledem k tomu, že Wikipedie obsahuje přes 5 milionů článků).

8.5 Identifikace psích ras

Posledním typem entity, na které bude tento systém otestován, je identifikace ras psů na anglické Wikipedii. Jak již bylo uvedeno v kapitole 3.2.3, jedná se o entitu, která je na DBpedii definována v [5], ale v současnosti do ní není zařazen ani jeden článek. Většina psů na DBpedii je zařazena do nadřazené entity Mammal (např. Leonberger a Chinese Crested Dog), i když např. článek Siberian Husky je na DBpedii zcela mylně zařazen do skupiny Organisation resp. do žádné skupiny v případě DBpedie Live.

Jako vstupem tohoto systému bylo zvoleno cca 100 článků pojednávajících o psích rasách a 96 článků, které o psích rasách nepojednávají, přičemž při volbě těchto článků bylo třeba dbát na to, aby seznam článků nepojednávajících o psích rasách obsahoval články podobné identifikovaným článkům (zejména články o konkrétních psech, články o rasách koček a koní).

Vzhledem k tomu, že tato entita obsahuje ještě méně článků, než je entita obsahující státy, bylo třeba snížit minimální požadovanou hodnotu *support* na 50 % nejvyšší dosažené hodnoty *support* u stejného typu příznaku k tomu, aby vůbec nějaké relevantní příznaky byly vygenerovány.

Celkem bylo tímto způsobem nalezeno 538 článků o psích rasách. Cca 2 % těchto článků (testováno na 100 náhodně vybraných článcích) ve skutečnosti o psí rase nepojednávají, přičemž při náhodné kontrole 100 ostatních článků další psí rasy nalezeny nebyly.

Nicméně systém YAGO [11] obsahuje entitu DogBreeds, ve které se nachází 387 psích ras. Seznam těchto ras lze získat pomocí následujícího SPARQL dotazu, který může být spuštěn např. na serveru DBpedie Live[3]:

```
PREFIX yago: <http://dbpedia.org/class/yago/>
SELECT ?p where {
    ?p rdf:type yago:DogBreeds .
}
LIMIT 1000
```

Při porovnání výsledků obou systémů bylo 302 psích ras identifikováno oběma systémy, 236 psích ras pouze tímto systémem a 85 psích ras pouze systémem YAGO. U všech třech skupin psích ras byla zjištěna chybovost maximálně v jednotkách článků. U některých záznamů, které nebyly identifikovány oběma systémy, byla tato chyba způsobena přejmenováním článků (tento skript pracuje s posledním dumpem, data získaná z YAGO nikoliv; cca desítky záznamů). Ukázky konkrétních článků rozdělených do skupin podle toho, který systém je identifikoval, lze nalézt v příloze F. Úplný seznam všech nalezených článků o psích rasách lze nalézt na přiloženém paměťovém médiu ve složce `data/en/dogs/`.

8.6 Výběr prahových hodnot standardních metrik

V předcházejících podkapitolách 8.3 až 8.5 byly při generování seznamů entit voleny minimální požadované hodnoty metrik *support*, *confidence* a *lift*. Rád bych věnoval pár slov tomu, jakým způsobem byly tyto hodnoty zvoleny a jaké lze použít hodnoty pro identifikaci dalších entit.

Základním přístupem při volbě těchto hodnot je metoda pokus-omyl. Každá extrahovaná entita je jiná a nelze použít stejné hodnoty pro všechny typy entit. V případě, že vygenerované seznamy článků nebudou uspokojivé, lze tyto hodnoty změnit a vygenerovat nový seznam příznaků[4].

Zásadní vliv na tyto údaje má **počet článků**, které budou do hledané entity patřit. Čím početnější entita je, tím obecnější příznaky je třeba volit (při vyhledávání všech osob nemá smysl generovat příznak, který pokrývá francouzské sochaře ze 17. století) a naopak. Na obecnost generovaných příznaků má nejvyšší vliv hodnota *support* – čím nižší bude, tím méně obecné příznaky budou generovány. Její hodnota se může pohybovat mezi 0 a 1. Téměř vždy je lepší tuto hodnotu nastavit podílem nejvyšší hodnoty *support* daného typu pravidla, jako to bylo činěno v předcházejících kapitolách a popsáno v kapitole 8.3. Pro středně až velmi rozsáhlé skupiny článků (lidé, města, události apod.) doporučuji tuto

[3] http://live.dbpedia.org/sparql

[4] V tomto případě je časově výhodnější nejprve vygenerovat všechny příznaky pomocí `wi make --use-all-rules` a po změnách prahů metrik vygenerovat nový seznam použitelných příznaků ze souboru všech příznaků pomocí přepínače `--from-file` (viz kapitola 4).

hodnotu ponechat na 80 % nejvyšší dosažené hodnoty daného typu příznaku. Pro malé entity je vhodné tuto hodnotu o řádově o desítky procent snížit.

Hodnota *confidence* určuje, jak kvalitní pravidla mají být vybírány a může se pohybovat mezi 0,5 a 1. Pro většinu případů by výchozí hodnota 0,95 měla být vyhovující. V případě, že by systém nenalezl skoro žádné příznaky, lze tuto hodnotu snížit. Metrika *lift* zajišťuje, aby nebyly vybírány zcela nesmyslné příznaky a s její výchozí hodnotou 1,1 by nemělo být nutné manipulovat.

Kapitola 9

Závěr

V této práci byl popsán systém, který dokáže identifikovat typy entit článků na základě dvou vstupních souborů, kde jeden soubor obsahuje seznam několika vzorových článků, které do hledané entity článků patří a druhý soubor obsahuje seznam několika vzorových článků, které do hledané entity článků nepatří.

K nalezení všech článků, které do hledané entity článků patří, bylo nejprve nutné analyzovat články, které byly uvedeny ve vzorových souborech. Ve všech vzorových článcích byly nalezeny příznaky, ze kterých byla vytvořena asociační pravidla, která mohou daný vzorový článek identifikovat. Z multimnožiny asociačních pravidel nalezených ve všech vzorových článcích byly s využitím standardních metrik asociační analýzy vybrány ty pravidla, která reprezentují co největší část vzorových článků. Poté byly tyto pravidla použity k nalezení všech článků dané entity na Wikipedii.

Úspěšnost tohoto systému byla ověřena na několika vygenerovaných entitách, a to na osobách a státních útvarech na české Wikipedii a na osobách a rasách psů na anglické Wikipedii. Žádný systém není bezchybný, ale úspěšnost tohoto systému při jeho vhodné konfiguraci byla velmi uspokojivá. Celkem tento systém při použití dumpu Wikipedie z května 2018 nalezl 104 796 česky psaných osob, 1 384 324 anglicky psaných osob, 1 538 česky psaných států a 538 anglicky psaných psích ras, přičemž až na výjimky se chybovost této identifikace pohybuje kolem jednoho procenta.

Předností tohoto systému je jeho možnost použití pro identifikaci libovolné entity a případná snadná možnost podpory dalšího jazyka. Systém by bylo možné rozšířit například o podporu vyhledávání definičních slov v dalších jazycích či o další typy příznaků.

Literatura

[1] Agrawal, R.; Imieliński, T.; Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.*, ročník 22, č. 2, Červen 1993: s. 207–216, ISSN 0163-5808, doi:10.1145/170036.170072.
Dostupné z: http://doi.acm.org/10.1145/170036.170072

[2] Alexa Internet, Inc. 1996: *Alexa Top 500 Global Sites.* [Online; navštíveno 16.04.2018].
Dostupné z: https://www.alexa.com/topsites

[3] Berners-Lee, T.: *Uniform Resource Locators (URL).* RFC 1738, 1994.

[4] DBpedia Developers: *DBpedia Live Mirror.* [Online; navštíveno 13.05.2018].
Dostupné z: https://github.com/dbpedia/dbpedia-live-mirror/

[5] DBpedia Developers: *Ontology Classes.* [Online; navštíveno 13.05.2018].
Dostupné z: http://mappings.dbpedia.org/server/ontology/classes/

[6] Elsahar, H.; Demidova, E.; Gottschalk, S.; aj.: Unsupervised Open Relation Extraction. In *The Semantic Web: ESWC 2017 Satellite Events*, editace E. Blomqvist; K. Hose; H. Paulheim; A. Ławrynowicz; F. Ciravegna; O. Hartig, Cham: Springer International Publishing, 2017, ISBN 978-3-319-70407-4, s. 12–16.

[7] ExplosionAI UG (haftungsbeschränkt): *spaCy · Industrial-strength Natural Language Processing in Python.* [Online; navštíveno 04.05.2018].
Dostupné z: https://spacy.io/

[8] ExplosionAI UG (haftungsbeschränkt): *Training spaCy's Statistical Models · spaCy Usage Documentation.* [Online; navštíveno 07.05.2018].
Dostupné z: https://spacy.io/usage/training

[9] Gao, S. S.; Sperberg-McQueen, C. M.; Thompson, H. S.: *W3C XML Schema Definition Language (XSD) 1.1.* 2012, [Online; navštíveno 10.05.2018].
Dostupné z: https://www.w3.org/TR/xmlschema11-1/

[10] Harris, S.; Seaborne, A.: *SPARQL 1.1 Query Language.* 2013, [Online; navštíveno 12.05.2018].
Dostupné z: https://www.w3.org/TR/sparql11-query/

[11] Hoffart, J.; Suchanek, F. M.; Berberich, K.; aj.: *YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia.* Max Planck Institute for Informatics, 2012.

[12] 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)*, 2007.

[13] KNOT@FIT: *Skript na získavanie prvých viet/odstavců z textov z Wikipédie*. [Online; navštíveno 12.05.2018].
Dostupné z: https://knot.fit.vutbr.cz/wiki/index.php/Decipher_wikipedia

[14] KNOT@FIT: *Výzkumná skupina znalostních technologií*. [Online; navštíveno 08.05.2018].
Dostupné z: https://knot.fit.vutbr.cz/

[15] Lehmann, J.; Isele, R.; Jakob, M.; aj.: *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. Semantic Web 1 (2012) 1–5, 2012.

[16] Medelyan, O.; Milne, D.; Legg, C.; aj.: *Mining meaning from Wikipedia*. 2009, ISSN 1071-5819.

[17] MediaWiki: *API:Main page — MediaWiki, The Free Wiki Engine*. [Online; navštíveno 20.04.2018].
Dostupné z:
https://www.mediawiki.org/w/index.php?title=API:Main_page&oldid=2746422

[18] MediaWiki: *API:MediaWiki API help, action=query — MediaWiki, The Free Wiki Engine*. [Online; navštíveno 20.04.2018].
Dostupné z: https://en.wikipedia.org/w/api.php?action=help&modules=query

[19] OpenLink Software: *OpenLink Virtuoso*. [Online; navštíveno 13.05.2018].
Dostupné z: https://virtuoso.openlinksw.com/

[20] Python Software Foundation: *Python 3.6.5 documentation*. [Online; navštíveno 08.05.2018].
Dostupné z: https://docs.python.org/3/

[21] RDF Working Group: *Resource Description Framework (RDF)*. 2012, [Online; navštíveno 20.04.2018].
Dostupné z: https://www.w3.org/RDF/

[22] Tan, P.-N.; Steinbach, M.; Karpatne, A.; aj.: *Introduction to Data Mining (2nd Edition) (What's New in Computer Science)*. Pearson, 2018, ISBN 9780133128901.

[23] The Daring Fireball Company LLC: *Markdown Syntax Documentation*. [Online; navštíveno 20.04.2018].
Dostupné z: https://daringfireball.net/projects/markdown/syntax

[24] Wikipedia contributors: *Database download — Wikipedia, The Free Encyclopedia, kapitola Where do I get it?* [Online; navštíveno 20.04.2018].
Dostupné z: https://en.wikipedia.org/w/index.php?title=Wikipedia:Database_download&oldid=835600135#English-language_Wikipedia

[25] Wikipedia contributors: *Statistics — Wikipedia, The Free Encyclopedia*. [Online; navštíveno 16.04.2018].
Dostupné z: https://en.wikipedia.org/wiki/Special:Statistics

[26] Wikipedia contributors: *Template namespace — Wikipedia, The Free Encyclopedia*. [Online; navštíveno 20.04.2018].
Dostupné z: https://en.wikipedia.org/w/index.php?title=Wikipedia:Template_namespace&oldid=832376965

[27] Wikipedia contributors: *What is an article? — Wikipedia, The Free Encyclopedia*. [Online; navštíveno 20.04.2018].
Dostupné z: https://en.wikipedia.org/w/index.php?title=Wikipedia:What_is_an_article%3F&oldid=832542381

[28] Wikipedia contributors: *Wikislovník*. [Online; navštíveno 01.05.2018].
Dostupné z:
https://cs.wiktionary.org/wiki/Wikislovn%C3%ADk:Hlavn%C3%AD_strana

[29] Wikipedie: *Statistika — Wikipedie: Otevřená encyklopedie*. [Online; navštíveno 16.04.2018].
Dostupné z: https://cs.wikipedia.org/wiki/Speci%C3%A1ln%C3%AD:Statistika

[30] Wu, F.; Weld, D. S.: *Automatically Refining the Wikipedia Infobox Ontology*. ACM, 2008, ISBN 978-1-60558-085-2.

Příloha A

Ukázka dump souboru Wikipedie

Následuje zkrácená ukázka dump souboru anglické Wikipedie `pages-articles.xml`. Místa označená `[...]` představují vynechanou část textu.

```
<mediawiki [...] version="0.10" xml:lang="en">
  <siteinfo>
    <sitename>Wikipedia</sitename><dbname>enwiki</dbname>
    <base>https://en.wikipedia.org/wiki/Main_Page</base>
    [...]
    <namespaces>
      <namespace key="-1" case="first-letter">Special</namespace>
      <namespace key="0" case="first-letter" />
      <namespace key="1" case="first-letter">Talk</namespace>
      [...]
    </namespaces>
  </siteinfo>
  <page>
    <title>AccessibleComputing</title>
    <ns>0</ns>
    <id>10</id>
    <redirect title="Computer accessibility" />
    <revision>
      <id>767284433</id><parentid>631144794</parentid>
      <timestamp>2017-02-25T00:30:28Z</timestamp>
      <contributor>
        <username>Godsy</username><id>23257138</id>
      </contributor>
      <comment>[[Template:This is a redirect]] has been[...]</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text [...]>#REDIRECT [[Computer accessibility]] [...]</text>
      <sha1>ds1crfrjsn7xv73djcs4e4aq9niwanx</sha1>
    </revision>
  </page>
  <page>[...]</page> <page>[...]</page> [...]
</mediawiki>
```

Příloha B

Nejčastější slova v prvních větách

Tyto tabulky obsahují seznam slov, které se nejčastěji vyskytují v prvních větách článků české Wikipedie v závislosti na tom, zda dané slova byly nalezeny ve slovníku Wiktionary (viz kapitola 6.2.2).

Tabulka B.1: Slova nalezená v Wiktionary

Výskytů	**Slovo**
260973	je
235245	v
157592	a
72071	na
63904	byl
57050	z
50941	se
38772	ve
27743	od
27077	který
23444	okrese
22966	nebo
22662	roku
21851	český
21644	s
18900	byla
18818	pro
18530	jako
17689	do
14930	která
13803	za
13611	roce
13035	ledna
12948	de
12707	které
12412	též
12409	politik

Tabulka B.2: Slova nenalezená v Wiktionary

Výskytů	**Slovo**
6956	km
5943	m
4156	reprezentant
3723	studiové
2876	fr
2796	č
2270	Československa
2199	FC
2160	provincii
2138	rockové
2069	Říšské
2058	hrající
2015	doslova
1909	zkráceně
1896	vydané
1896	př
1855	označuje
1821	turnaj
1819	provincie
1771	katolíků
1699	tzv
1543	sněmu
1521	Hradec
1487	rodným
1483	okruhu
1480	III
1448	R

Příloha C

Nejčastější definiční slova

Tyto tabulky obsahují seznam definičních slov (viz kapitola 6), které byly mezi všemi českými a anglickými větami nalezeny nejčastěji.

Tabulka C.1: Def. slova na anglické Wikipedii

Výskytů	**Slovo**
218536	village
118114	politician
113508	player
110471	film
108907	album
99690	footballer
79950	moth
60155	town
58829	list
54647	station
48529	municipality
45552	song
45220	writer
44489	actor
42644	community
40783	studio album
40599	genu
39735	commune
39363	district
39280	surname
39011	beetle
36945	member
36646	school
34938	actress
34070	unincorporated community
32484	football player
31599	species of beetle

Tabulka C.2: Def. slova na české Wikipedii

Výskytů	**Slovo**
11348	politik
8528	město
8109	vesnice
7588	část
7130	obec
4812	fotbalista
4264	spisovatel
4220	kanton
4180	francouzský kanton
4103	klub
3984	reprezentant
3719	herec
3699	skladatel
3664	film
3602	malá vesnice
3293	druh
3283	název
3059	útočník
2997	obránce
2927	skupina
2843	herečka
2753	fotbalový klub
2678	záložník
2667	rod
2548	malíř
2528	seznam
2495	československý politik

Příloha D

Získané příznaky článku K. Bein

Tato příloha obsahuje seznam všech příznaků, které byly vygenerovány pro článek *Kazimierz Bein* na české Wikipedii (včetně těch, které nebyly vybrány jako vhodné pro identifikaci ostatních článků):

Tabulka D.1: Seznam vygenerovaných příznaků

Typ	Pt.	Hodnota
article		none
category	sh.	Muži
category	sh.	Narození <num>
category	sh.	Narození 1872
category	sh.	Oční lékaři
category	sh.	Polští esperantisté
category	sh.	Polští lékaři
category	sh.	Polští překladatelé
category	sh.	Přek. do esperanta (A)
category	sh.	Přek. z polštiny (A)
category	sh.	Úmrtí <num>
category	sh.	Úmrtí <num>. června
category	sh.	Úmrtí v Lodži
category	sh.	Úmrtí 15. června
category	sh.	Úmrtí 1959
category	1.	Muži
category	1.	Narození
category	1.	Oční
category	1.	Polští
category	1.	Překladatelé
category	1.	Úmrtí
category	p.s.	června
category	p.s.	esperanta
category	p.s.	esperantisté
category	p.s.	lékaři
category	p.s.	Lodži
category	p.s.	Muži
category	p.s.	<num>
category	p.s.	polštiny
category	p.s.	překladatelé
category	p.s.	1872
category	p.s.	1959
header		Překlady do esperanta
header		Související články
is-a		lékař
is-a		překladatel
template	sh.	Autoritní data
template	sh.	SORT... (B)
template	sh.	Infobox - spisovatel

Vysvětlivky: **article** = Typ členu v první větě, **category** = Kagegorie, **header** = Kapitoly, **is-a** = Definiční slovo, **template** = Šablona, **Pt.** = Podtyp příznaku, **sh.** = přesná shoda textu, **1.** = shoda 1. slova, **p.s.** = shoda posledního slova, **(A) Překl.** = Překladatelé, **(B) SORT...** = DEFAULTSORT:Bein, Kazimierz

Příloha E

Příznaky k identifikaci osob v ČJ

Tato tabulka obsahuje seznam všech příznaků, které byly použity při identifikaci osob na české mutaci Wikipedie spolu s jejich standardními metrikami:

Tabulka E.1: Seznam vygenerovaných příznaků

Příznak	**V.**	**Supp.**	**Conf.**	**Lift**
is-a::obor	F	0.047	1.0	1.87
category::first-word::<num>	F	0.047	1.0	1.87
header::Životopis	T	0.093	1.0	2.15
category::first-word::Narození	T	0.349	1.0	2.15
category::first-word::Přírodní	F	0.047	1.0	1.87
category::first-word::Gravitace	F	0.047	1.0	1.87
title::parentheses::1910	T	0.023	1.0	2.15
title::first-two-words::Aun Schan	T	0.023	1.0	2.15
category::exact::Narození <num>	T	0.349	1.0	2.15
category::first-word::Muži	T	0.349	1.0	2.15
title::first-two-words::Seznam českých	F	0.023	1.0	1.87
category::last-word::vědy	F	0.047	1.0	1.87
header::Dílo	T	0.093	1.0	2.15
category::exact::<num>	F	0.047	1.0	1.87
category::last-word::Gravitace	F	0.047	1.0	1.87
category::exact::Přírodní vědy	F	0.047	1.0	1.87
title::parentheses::rozcestník	F	0.023	1.0	1.87
category::exact::Gravitace	F	0.047	1.0	1.87
is-a::věda	F	0.047	1.0	1.87
category::first-word::<num>.	F	0.047	1.0	1.87
is-a::herec	T	0.070	1.0	2.15
title::first-two-words::Nobelova cena	F	0.023	1.0	1.87
category::last-word::Muži	T	0.349	1.0	2.15
is-a::spisovatel	T	0.070	1.0	2.15
is-a::pohyb	F	0.047	1.0	1.87
title::first-two-words::Polská fotbalová	F	0.023	1.0	1.87
category::exact::Muži	T	0.349	1.0	2.15
title::first-two-words::Štefan Luby	T	0.023	1.0	2.15

Příloha F

Porovnání nalezených psích ras ve srovnání se systémem YAGO

V této příloze jsou uvedeny ukázky článků, které při identifikaci psích ras byly nalezeny oběma systémy, které byly nalezeny pouze tímto systémem a které byly nalezeny jen systémem YAGO [11]:

Oba systémy (302)

- Akita (dog)
- Australian Shepherd
- Azawakh
- Bedlington Terrier
- Belgian Shepherd
- Curly Coat. Retriever
- Dogue de Bordeaux
- English Mastiff
- Georgian Shepherd
- Halden Hound
- Lakeland Terrier
- Norwegian Buhund
- Portuguese Podengo
- Redbone Coonhound
- Sabueso Español
- Segugio dell'Appennino
- Taigan
- Vanjari Hound

Jen tento sys. (236)

- Amer. Pit Bull Terrier
- Austral. Silky Terrier
- Blue Paul Terrier
- Cursinu
- Dutch Shepherd
- Ecuador. Hairless Dog
- Golden Dox
- Greyhound
- Groenendael dog
- Hierran Wolfdog
- Hokkaido dog
- Hygen Hound
- Lucas Terrier
- New Gui. singing dog
- Norwegian Elkhound
- Olde Boston Bulldogge
- Red Setter
- Schipperke

Pouze YAGO (85)

- Akbash Dog
- Armant (dog)
- Bakharwal Dog
- B. Sheph. Dog (Laek.)
- Bulldog Campeiro
- Bully Kutta
- Cane Di Oropa
- Dogo Sardesco
- Dutch Shepherd Dog
- Form. Mountain Dog
- Kaikadi (dog)
- Newfoundland (dog)
- Pardog
- P. de Presa Mallorquin
- Saluki
- Samoyed (dog)
- Sm. Gr. Domestic Dog
- Sulimov Dog

Příloha G

Plakát

URČOVÁNÍ TYPŮ ENTIT NA ZÁKLADĚ EXTRAKCE INFORMACÍ Z WIKIPEDIE

Autor: Petr Rusiňák

Cíl práce

- Identifikovat typy článků (entit) na Wikipedii
- Vytvoření systému, který není závislý na konkrétní entitě
- Systém by měl být, taktéž rozšiřitelný na různé jazyky.

Vstupní soubory

Seznamy vzorových článků, které:

- patří do hledané entity
- nepatří do hleadné entity

Výstupní soubory

Seznamy:

- příznaků pro identifikování článků
- všech článků patřících do hledané entity
- všech článků nepatřících do hledané entity

Typy příznaků

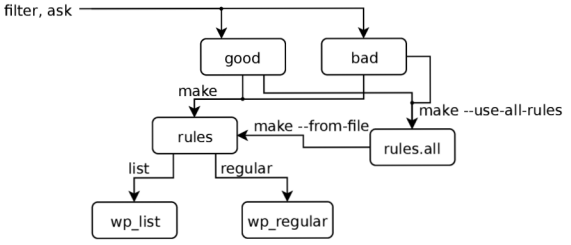
- Šablony (název, klíče pojmenovaných parametrů)
- Kategorie (shoda, první a poslední slovo)
- Kapitoly
- Název článku (první dvě slova, text v závorce)
- Definiční slovo
- Typ členu v první větě

Výsledky

Tento systém nalezl:

- 104 796 článků osob (ČJ)
- 1 384 324 článku osob (AJ)
- 1 538 článků o státech (ČJ)
- 538 článků psích ras (AJ)

Schéma systému



- good, bad = soubory se vzorovými články
- rules = seznam vygenerovaných příznaků
- wp_list = seznam článků patřících do dané entity
- wp_regular = seznam článků nepatřících do dané entity

Obrázek G.1: Plakát popisující tento systém

Příloha H

Obsah přiloženého CD

- `data/` – ukázka vstupních a výstupních souborů systému
- `latex/` – zdrojové soubory tohoto textu
- `src/` – zdrojové soubory systému
- `Readme.md` – uživatelský manuál
- `xrusin03-Urcovani-typu-entit-clanku-z-Wikipedie.pdf` – elektronická podoba tohoto textu