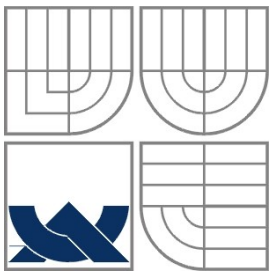
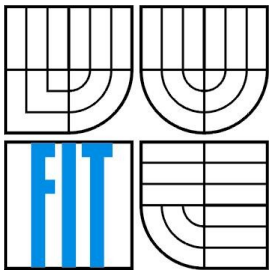


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA A AGREGACE DAT PARLAMENTU ČR

ANALYSIS AND AGGREGATION OF CZECH PARLIAMENT DATA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

IRENA TALAŠOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. PAVEL OČENÁŠEK, Ph.D.

BRNO 2016

Abstrakt

Cílem této práce je vytvořit různé statistiky a přehledy o fungování Parlamentu ČR v rámci jeho legislativní činnosti. Data pro tyto statistiky jsou získávána ze stránek Parlamentu a ukládána do databáze. Uživateli jsou výsledky prezentovány především ve formě tabulek, grafů a dalších metod na veřejně přístupné webové stránce. Výsledky by měly být co nejpřehlednější, nejsrozumitelnější a poskytovat zajímavé a netradiční informace.

Abstract

The aim of this work is to create a variety of statistics and reports on the functioning of Parliament of Czech Republic. Data for these statistics are gathered from the pages of the Czech Parliament and stored in to a database. Results are presented mainly in the form of tables, graphs and other methods on a publicly accessible Web site. Results should be as clear as possible, comprehensible and provide interesting and unusual information.

Klíčová slova

Parlament ČR, statistiky, přehledy, Poslanecká sněmovna, Senát, zákony, legislativa, stoplist, PHP, databáze, JSON, tabulky, grafy

Keywords

Czech Parliament, statistics, reports, Chamber of Deputies, Senate, laws, legislation, stoplist, PHP, database, JSON, tables, graphs

Citace

Irena Talašová: Analýza a agregace dat Parlamentu ČR, bakalářská práce, Brno, FIT VUT v Brně, 2016

Analýza a agregace dat Parlamentu ČR

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením Ing. Pavla Očenáška, Ph.D.
Další informace mi poskytli Ing. Vladimír Bartík, Ph.D. a
doc. RNDr. Pavel Smrž, Ph.D.
Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Irena Talašová
25.4.2016

Poděkování

Chtěla bych poděkovat svému vedoucímu práce, panu Ing. Pavlu Očenáškovi, Ph.D., za jeho rady, trpělivost, čas a podněcování k lepším výsledkům. Dále také patří velké díky panu Ing. Vladimíru Bartíkovi, Ph.D. a panu doc. RNDr. Pavlu Smržovi, Ph.D., za poskytnutí dalších informací. Také jsem vděčná své rodině a svým blízkým za podporu během celého studia.

Obsah

1 Úvod

V současné době je dění v Poslanecké sněmovně Parlamentu ČR a Senátu Parlamentu ČR sledováno širokou veřejností, avšak většinou pouze prostřednictvím informací zprostředkovanými televizí, rádiem nebo internetovými portály. Tvorba statistik a souhrnných informací rozhodně není tak častá, ačkoli je zajímavá a může člověku poskytnout zajímavé údaje, na které by jinak ani nepřišel nebo by o nich nepřemýšlel. Bohužel na tuto doménu se stále mnoho zdrojů nezaměřuje nebo jen ojediněle. K dispozici jsou většinou jen ty statistiky, které poskytne sama Poslanecká sněmovna nebo Senát na svých internetových stránkách.

Cílem této práce je lokalizovat relevantní informace na stránkách Parlamentu, tyto údaje zpracovat a vytvořit z nich přehledové statistiky, které budou tato data reprezentovat. Uživateli budou výsledky prezentovány pomocí statistických metod pro reprezentaci dat na veřejně přístupné webové stránce.

První část této práce poskytuje teoretický úvod k legislativní činnosti Parlamentu ČR; jedná se o kapitolu 2. Je zde naznačeno fungování, činnost a pravomoci Poslanecké sněmovny a poté také Senátu. Důležité pro tuto práci jsou také informace uvedené v podkapitolách 2.3 a 2.4, ve kterých je zmíněn legislativní proces návrhu zákona a jednotlivé typy parlamentních tiskopisů. V poslední části druhé kapitoly se nacházejí obecné informace o vyhledávání informací v textu (podkapitola 2.6) a také jedna konkrétní metoda – metoda IDF. Následující kapitola 2.7 popisuje možnosti extrakce informací z HTML a jednotlivé možnosti přibližuje čtenáři.

Druhá část této práce, obsažená v kapitole 3, pojednává o aktuálních existujících řešeních podobného problému (podkapitola 3.1). Je zde uvedeno, kde můžeme nalézt statistiky týkající se Parlamentu ČR a o jaký typ statistik se jedná. V podkapitole 3.2 je uvedena logická struktura stránek Parlamentu ČR. Pro extrakci informací je důležité vědět, jakou mají jednotlivé stránky strukturu a také, jak jsou zde data uložena a v jakém formátu; o tom se lze dočíst v podkapitole 3.3.

Třetí část práce je uvedena v kapitole 4. Zabývá se popisem statistických veličin a nástrojů pro reprezentaci dat, které jsou v této práci použity. Okrajově jsou zde popsány i nástroje, které použity při implementaci nebyly, ale o kterých během vývoje bylo uvažováno jako o dalších možnostech.

Čtvrtá část pojednává o návrhu aplikace. Najdeme ji v kapitole 5. Obsahuje přehled prováděných statistik, použité technologie a také zevrubně popisuje princip fungování celé aplikace.

Pátá část se nachází v kapitole 6 a zabývá se konkrétní implementací aplikace. Jsou zde popsány jednotlivé moduly, z nichž se program skládá, včetně konkrétních principů implementace. Najdeme zde i ukázky výsledných statistik a informace o průběhu testování aplikace.

Poslední, pátá část se nachází v kapitole 7. Obsahuje závěr, další možná rozšíření a přínos práce pro autora.

2 Teoretický úvod

Tato kapitola se zabývá principem fungování Poslanecké sněmovny, Senátu ČR a legislativním procesem návrhu zákona. Dále budou popsány typy a významy tiskopisů, které budou důležité pro tuto práci. Ke konci budou představeny základy vybraných technik a principů práce s textem, které jsou dále používány.

2.1 Poslanecká sněmovna

Následující text vychází z informací uvedených ve zdroji [1] a také z Ústavy České republiky[1]. Podle Ústavy České republiky je zákonodárná moc svěřena do rukou prezidenta a Parlamentu ČR, který má dvě komory - Poslaneckou sněmovnu a Senát. Polaneckou sněmovnu tvoří 200 poslanců volených na čtyři roky v rámci poměrného systému. V kompetenci má vyjadřovat se k mezinárodním smlouvám, vyslovovat důvěru nebo nedůvěru vládě, navrhovat zákony a další.

Během hlasování musí být přítomna alespoň třetina poslanců, tedy alespoň 67 z nich se musí účastnit hlasování. Pro schválení návrhu zákona je nutná nadpoloviční většina hlasů přítomných poslanců. V případě přehlasování prezidentského veta nebo vyslovování nedůvěry vládě je třeba alespoň 101 hlasů. Pro ratifikaci smluv a rozpuštění sněmovny je třeba 120 hlasů.

Zvolen může být každý občan ČR, který dosáhl 21 let a nebyl zbaven svéprávnosti. Hlasy z jednotlivých krajských obvodů se přepočítávají na mandáty. Zvolený poslanec navíc získává imunitu – nemůže být trestně stíhán. Poslance nelze odvolat, odchází až po ukončení volebního období nebo odstoupením. Nelze slučovat funkce poslance v Poslanecké sněmovně, Senátu, prezidenta ČR, soudce a další.

Sněmovna má jednoho stálého předsedu, pět místopředsedů, komise a stálé výbory. Sněmovna nesmí být rozpuštěna tři měsíce před koncem volebního období, jinak ji za speciálních okolností může rozpustit prezident, například pokud zasedání bylo přerušeno na dlouhou dobu nebo si sama Sněmovna přeje být rozpuštěna.

Poslanecké kluby slouží v Poslanecké sněmovně ke sdružování poslanců, kteří patří do stejné politické strany. Takto mohou vystupovat jednotně s jednotnými cíli. Každý poslanec může patřit pouze do jednoho poslaneckého klubu. Každý klub má svého předsedu a místopředsedu.

Vláda ČR je vrcholný orgán výkonné moci. Skládá se z předsedy vlády, nebo-li premiéra, místopředsedů a ministrů. Má zákonodárnou iniciativu. To znamená, že může podávat návrhy zákonů a také se může k zákonům vyjadřovat.

1 http://www.psp.cz/docs/laws/constitution.html

2.2 Senát

Tyto informace lze nalézt ve zdroji [2]. Senát má 81 členů. Charakteristikou je, že se jedná o instituci trvalou, nemůže být tedy rozpuštěna, ale každé dva roky je z jedné třetiny obnovována. Senátorem se může stát kterýkoliv občan ČR starší 40 let, který nebyl zbaven svéprávnosti. Senátoři jsou voleni v 81 mandátních obvodech na šest let. Jednou za dva roky tedy probíhají volby ve třetině obvodů. Volby se konají nejvýše ve dvou kolech, do druhého postupují dva nejúspěšnější kandidáti. Výsledky voleb však mohou být známé již v prvním kole, a to pokud některý kandidát získal nadpoloviční většinu hlasů. Pokud některému ze senátorů zanikne mandát, konají se volby doplňovací.

Mezi základní kompetence Senátu patří:

- Projednává návrhy zákonů postoupené Poslaneckou sněmovnou.
- Navrhuje zákony, které postupuje Poslanecké sněmovně.
- Přijímá zákonná opatření v případě, že je Poslanecká sněmovna rozpuštěna.
- Dává souhlas k ratifikaci mezinárodních smluv.
- Předseda senátu vyhlašuje volbu prezidenta republiky.
- A některé další.

Na první schůzi po volbách do Senátu se volí předseda a místopředsedové. Dále se na počátku každého funkčního období zřizuje Organizační, Mandátový a imunitní výbor, které předepisuje zákon, mimo tyto také další výbory. Poslanci jsou organizováni do výborů, které projednávají co je jim přikázáno nebo co si sami stanoví. Poslanci mohou být členem pouze jednoho výboru, to však neplatí v případě Organizačního, Mandátového a imunitního výboru. Předseda a místopředsedové Senátu jsou členy pouze Organizačního výboru. Senát může vytvořit podvýbor, pokud mu potřebuje uložit konkrétní problematiku k řešení.

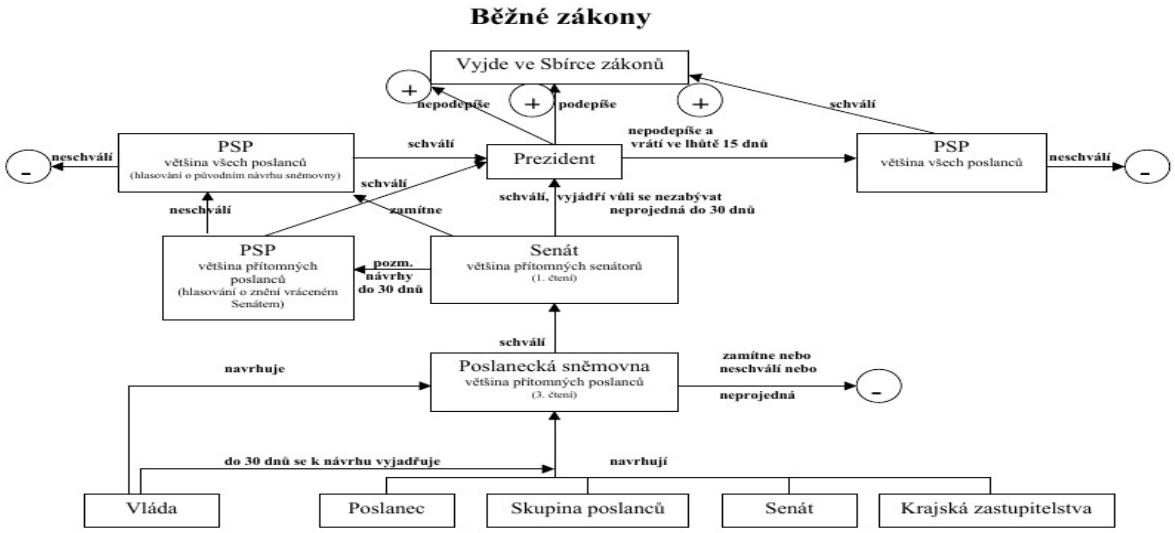
Dále Senát vytváří komise, pokud se provádí úkoly, které spadají pod více orgánů nebo naopak žádnému nepřísluší. Členy komise mohou být senátoři, ale i další osoby.

Senátoři se mohou sdružovat v senátorských klubech na základě příslušností k jednotlivým politickým stranám. Každý může být členem pouze jednoho senátorského klubu. Pro vznik nového klubu je třeba alespoň pěti senátorů.

2.3 Legislativní proces návrhu zákona

Tento text vychází ze zdroje [1] a [5]. Návrh zákona může podat poslanec, skupina poslanců, Senát, vláda a zastupitelstvo vyššího územního samosprávného celku. Samotný průběh návrhu a schvalování běžného návrhu zákona v Poslanecké sněmovně (dále PS) a Senátu probíhá v několika krocích:

- První čtení v PS – během něj navrhovatel přednese novelu nebo návrh zákona poslancům a vede se takzvaná obecná rozprava. Poté ho Sněmovna může úplně zamítnout nebo ještě vrátit navrhovateli na přepracování. Jinak se prostřednictvím některého z poslanců dá návrh na projednání některému výboru.
- Druhé čtení v PS – pověřený výbor přednese svůj názor na novelu a doporučení, jestli by měl být přijat nebo zamítnut. V této fázi dochází k pozměňování a doplňování návrhu. Uskutečňuje se další obecná rozprava. Sněmovna může po celou dobu návrh zamítnout nebo vrátit výboru.
- Třetí čtení v PS – koná se další rozprava, během které je možné opravit legislativně technické chyby, gramatické a další vyplývající z pozměňovacích návrhů. Na závěr se hlasuje o pozměňovacích návrzích a o tom, jestli bude návrh přijat nebo ne.
- Pokud návrh není přijat, sněmovna se jím dále nezabývá. Pokud je přijat, postoupí ho Senátu, který má časově omezenou lhůtu na to, aby se k němu vyjádřil. Buď se rozhodne návrhem nezabývat, v tom případě návrh pokračuje rovnou k prezidentu republiky. Nebo návrh zamítne či ho vrátí Poslanecké sněmovně spolu s pozměňovacími návrhy. V obou posledních případech je návrh vrácen a Poslanecká sněmovna o něm hlasuje znovu. Pokud návrh znovu schválí, návrh putuje k prezidentovi.
- Prezident republiky má právo veta, to znamená, že návrh může ve lhůtě patnácti dní vrátit Poslanecké sněmovně. Po podpisu prezidenta je návrh schválen a vyjde ve Sbírce zákonů.
- Pokud prezident návrh vrátí zpět do PS, ta o něm hlasuje a pokud znovu hlasuje většina poslanců pro, návrh je přijat i bez podepsání prezidentem.



Obrázek 2.1: Schéma legislativního procesu.

Princip schvalování návrhů zákonů popsaný výše platí pro běžné zákony. Schéma legislativního procesu pro běžné zákony je k vidění na obrázku 2.1. Obrázek pochází ze zdroje http://www.psp.cz/sqw/hp.sqw?k=173. Další typy návrhů jako schvalování státních rozpočtů, ústavních zákonů, mezinárodních smluv a zákonná opatření Senátu se mírně svým průchodem Poslaneckou Sněmovnou a Senátem liší. Bližší informace lze nalézt ve zdroji [5].

2.4 Typy tiskopisů

Následující text pojednává o sněmovních tiscích, se kterými se bude pracovat v rámci této práce. Budou zde uvedeny jednotlivé typy a jejich krátká charakteristika. Následující text vychází ze zdrojů [3] a [4].

Sněmovní tisk znamená podklad pro jednání Sněmovny. Mezi sněmovní publikace se řadí sněmovní tisky a těsnopisecké zprávy o schůzích Sněmovny s uvedením jednotlivých vystoupení, usnesení a výsledků hlasování. Tyto publikace jsou přístupné pro veřejnost a pravidelně se zveřejňují na internetu. Tisky se poslancům doručují v elektronické podobě. Jsou opatřeny datem a číslem. Od tohoto okamžiku se jedná tak, jako že tisk byl doručen v dané době všem poslancům a běží zákonné lhůty. V případě dodatečných oprav se tisk doručuje poslancům opakovaně.

Typy tiskopisů:

- Návrhy zákonů – běžné, obecné návrhy zákonů.
- Mezinárodní smlouvy - smlouvy týkající se i jiných zemí.
- Rozpočty – státní rozpočty předložené vládou.
- Písemné a ústní interpelace – dotazy poslanců na členy vlády, jsou pro ně vyhrazeny čtvrtky.
- Zprávy – předložení informací o stavu sněmovně nebo konkrétní záležitosti.
- Stanoviska vlády.
- Usnesení výborů – vypracovaný názor výborů na danou problematiku nebo návrh zákona.
- Pozměňovací návrhy – pozměňovací návrhy návrhů zákonů.

Každý tisk je označen jedinečným identifikátorem v číselné podobě. V současné době se přešlo od sekvenčního číslování tisků na takzvané sdružené, což znamená, že sdružuje tisky dohromady tak, aby bylo jasné, že patří k sobě. Například 123/0, 123/1, 123/2, jsou související tisky, s číselným označením ve tvaru **T/Z**, přičemž **T** i **Z** jsou přirozená čísla.

2.5 Koalice

Pojem koalice obecně označuje spojení či spojenectví dvou a více organizací s cílem dosáhnout společného záměru. Ne jinak je tomu i v případě politických stran. Ty se sdružují do koalice podpisem koaliční smlouvy, ve které si ujasňují společné záměry a kroky, kterých chtějí díky spolupráci rychleji a snadněji dosáhnout. Strany, které se spojí do koalice, snadněji prosadí nejen své návrhy zákonů, ale také mohou výrazně zasahovat při hlasování o důvěře vlády. Tyto koalice mají většinou nadpoloviční počet členů v Poslanecké sněmovně. Koaliční smlouvy bývají uzavírány na celé volební období, v průběhu volebních období se střídají a koalice uzavírají jiné strany. Poslanci (příp. poslanecké kluby), kteří nejsou v koalici, jsou v opozici a snaží se svými hlasy vyvážit dohodnuté záměry stran. Úplné znění koaliční smlouvy pro toto období (rok 2013 – 2017) mezi stranami ČSSD, hnutím ANO 2011 a KDU-ČSL lze shlédnout v online přístupném dokumentu [6]. Koalici pro minulé volební období tvořily strany ODS, TOP09 a Věci veřejné[2].

2.6 Vyhledávání informací v textu

Dokument můžeme popsat množinou reprezentativních klíčových slov, kterým říkáme *index termy*. Každý *index term* má jinou důležitost pro popis daného dokumentu. Toto můžeme vystihnout například tak, že každému *index termu* přiřadíme – lépe řečeno vypočítáme – numerickou váhu (například technikou IDF (Inverse document frequency), TF-IDF (Term frequency–Inverse document frequency) a další) [7] , [8] a [9].

Nalezení *index termů* obnáší odstranění nevýznamových slov. Často se jedná o předložky, spojky, číslice, zájmena a často se opakující slova. Některá tato slova jsou obsažena v seznamech (tzv. stoplistech), které jsou již předpřipraveny některými autory a lze je použít jako výchozí bod pro vlastní, třeba rozsáhlejší seznam slov k odstranění. Nevýhodou pro práci v češtině je, že pro anglický jazyk jsou stoplisty poměrně časté, zatímco pro češtinu lze nalézt jen několik málo stoplistů, které nejsou nijak obsáhlé. Další věc, se kterou se při práci s textem v češtině potýkáme, je ohýbání slov. Jedno a totéž slovo jen vyskloňované či vyčasované bude algoritmus vyhodnocovat jako rozdílná slova. Je proto užitečné převést všechna slova nejprve na lemmata (tvar, který většinou najdeme ve slovníku např. 1. pád jednotného čísla většiny substantiv, infinitiv slovesa)[3]. Nalezením tvarů slov a převedením do základního tvaru se zabývá stemming. Analyzátory provádějící tuto činnost se rozlišují na ty, které berou v úvahu kontext slova a ty, které slova analyzují izolovaně. Teprve poté je vhodné přistoupit k výpočtu vah jednotlivých termů. Zde bude přiblížena pouze jedna váhovací metoda, metoda IDF [9].

2 http://www.vlada.cz/cz/media-centrum/dulezite-dokumenty/koalicni-smlouva-74245/

3 https://cs.wikipedia.org/wiki/Lemma_%28lingvistika%29

2.6.1 Váhování IDF

Jedná se o metodu více diskriminativní oproti jiným. Smyslem je hledání těch termů, které jsou méně frekventované v rámci všech dokumentů [7] ,[8] a [9]. Váha termu se vypočítá podle vzorce 2.1 [9].

$$IDF(t) = 1 + \log\left(\frac{n}{k}\right)$$

kde:
$n \ldots$ *celkový počet dokumentů*
$k \ldots$ *počet dokumentů, ve kterém se vyskytuje term* t

Vzorec 2.1: Výpočet váhy IDF.

2.7 Extrakce informací z HTML dokumentu

Následující text vychází z [10]. Dříve se techniky extrakce informací používaly zejména pro text typu text plain, tedy prostý text, jako například pro obsah e-mailových zpráv. S příchodem HTML jazyka se tyto techniky začaly využívat i v této doméně. Jazyk HTML je vytvořen tak, aby výsledný dokument byl přehlednější a snáze srozumitelný. Data jsou v HTML víceméně strukturovaná. Čtenář se díky tomu snadno orientuje v textu a může si vybírat části, které ho zajímají.

Z pohledu automatické extrakce informací je situace poněkud složitější, neboť zde jsou použity často jen části vět a delší souvislé bloky textu chybí. Na druhé straně HTML značky poskytují dodatečné informace a lze si jimi pomáhat při orientaci v textu.

Podle [11] lze rozlišit několik typů získávání informací a to manuální extrakce, automatická extrakce, wrappery a dokumentový objektový model DOM.

2.7.1 Manuální extrakce

Podle struktury konkrétní sady dokumentů určených ke zpracování se napíše program, který získá požadovaná data. Jelikož je založen na pozorování neměnné struktury dat, nevýhodou je okamžik, kdy dojde ke změně internetových stránek a program přestane fungovat. Výhodou však je, že získáme konkrétní a přesná požadovaná data.

2.7.2 Automatická extrakce

Sytém hledá vzory pro extrakci dat. Nejprve ale musí mít tréninkové příklady na kterých zjistí, kde jsou data uložena a tyto principy aplikuje na řešení úlohy. Používá se především pro rozsáhlé stránky.

2.7.3 Wrappery

Jedná se o nástroje, které obsahují množinu extrakčních pravidel. Podle těchto předdefinovaných pravidel se při průchodu wrapperu stránkou vybírají ty části, které pravidlům odpovídají. Pro každý

dokument musíme vytvořit vlastní množinu pravidel. Můžeme je rozdělit na řetězcové wrappery a wrappery využívající dokumentových objektových stromů. Více informací se lze dočíst v [11].

2.7.4 Dokumentový objektový model DOM

Jedná se o aplikační programové rozhraní pomocí něhož lze dynamicky dokumentu měnit strukturu a získávat z něj informace [11]. HTML nebo XML soubor je reprezentován jako stromová struktura, která po změně může být převedena zase zpět do původní podoby. Lze tedy rozlišovat a orientovat se podle nadřazených, podřazených či rovnocenných elementů.

3 Specifikace požadavků

Tato kapitola popíše strukturu stránek Parlamentu ČR z pohledu, který je důležitý pro tuto práci. Budou zde uvedeny způsoby uložení a formát důležitých dat a také budou představeny již existující podobná řešení. Všechny tyto principy zároveň tvoří specifikaci požadavků a charakter výsledné aplikace.

3.1 Existující řešení

Nacházíme se v době, kdy veřejnost sleduje dění v Parlamentu ČR nejčastěji prostřednictvím televize, rádia či novinek a zpráv na internetu. Zajímavé a netradiční údaje můžeme získat například prostřednictvím statistiky. Bohužel statistik v této oblasti není mnoho. Některé jsou zveřejněné přímo na stránkách Poslanecké sněmovny. Zde lze nalézt například počty hlasování jednotlivých poslanců, kolikrát hlasovali pro, kolikrát proti návrhu zákona nebo například kolikrát do Poslanecké sněmovny nedorazili vůbec[4]. Totéž lze najít na stránkách Senátu[5]. Co se týče jiných zdrojů, existují statistiky hlasování ve volbách nebo vývoj procentuálního zastoupení žen v Parlamentu, kde uvádějí, například kolik procent žen kandidovalo do Parlamentu během různých let a jak se vyvíjí procentuální zastoupení žen v Parlamentu [12]. Chybí však nějaké méně tradiční a méně obyčejné statistiky, které by čtenáři přinesly zajímavější a neotřepané informace.

3.2 Logická struktura stránek Parlamentu ČR

Výchozím bodem čerpání informací o Parlamentu ČR jsou dvoje internetové stránky – stránky Poslanecké sněmovny [1] a stránky Senátu ČR [2]. Na každé z těchto stránek najdeme informace specifické pro danou instituci. Jelikož je princip fungování Poslanecké sněmovny a Senátu podobný, jsou i zveřejňované informace podobného typu a na stránkách Senátu nalézáme informace týkající se i Poslanecké sněmovny a obráceně. Stejné informace jsou tedy uloženy na obou místech, jen v jiné struktuře. Příkladem toho jsou informace o tiscích.

3.2.1 Stránky Senátu

Podíváme-li se blíže na stránky Senátu, zjistíme, že informace jsou zde rozděleny do pěti hlavních kategorií, kterým odpovídají položky hlavního menu, a jedné domovské stránky, která sdružuje nejdůležitější a pravděpodobně nejhledanější informace čtenářů.

4 http://www.psp.cz/sqw/pstat.sqw?o=7&id=6146&id_posl=1306

5 http://www.senat.cz/informace/pro_media/statistiky/index.php?ke_dni=20.4.2016&O=10

První kategorie s názvem *Senát PČR* nejprve vysvětluje základní informace o Senátu a dále se věnuje podrobnějším věcem. Najdeme zde důležité informace o senátorech. Zejména je to seznam všech senátorů s uvedením volebního obvodu, jména, jeho internetové stránky, politické příslušnosti a notifikačního systému (systém pro objednání). Také je zde ke každému poslanci napsána adresa jeho kanceláře a asistent nebo asistentka a případně kontaktní informace. Velmi zajímavý je odkaz zasedacího pořádku senátorů, kde jsou vypsaní jednotliví senátoři a lze se pomocí barev orientovat, do kterého senátorského klubu patří. Po kliknutí na jméno senátora se zobrazí jeho osobní stránka, kde lze najít užitečné informace jako například, kdy byl zvolen a za jaký klub a jeho mandát. Také je zde uvedeno jeho členství ve výborech a zastupovaných funkcí v rámci Senátu. Dále zde najdeme odkazy na informace, jak tento konkrétní senátor hlasoval, které návrhy navrhoval a jeho vystoupení na schůzích Senátu. Dále lze u jednotlivých senátorů získat informace o jejich hlasování na jednotlivých schůzích a také přepis jejich řeči v Senátu. Také lze zjistit jestli je dotyčný předkladatelem nějakého návrhu zákona či ústní interpelace a další informace o jeho osobě. V této kategorii také nalezneme, jaké jsou orgány Senátu, jak probíhají volby a něco o vztahu Senátu a EU.

Druhá kategorie je důležitá, pokud chceme sledovat především aktuální dění. Najdeme zde aktuální i již proběhlé schůze a jejich program, kalendář akcí, harmonogram pravidelných akcí, konference, semináře a další akce. Dokonce zde lze shlédnout online přenos z jednání. V neposlední řadě tady najdeme také strukturovanou informaci, co se právě v aktuální den v Senátu děje i s rozpisem hodin.

Třetí kategorie se zabývá dokumenty a legislativou. Předně je zde odkaz na seznam všech senátních tisků, které lze filtrovat podle několika kritérií jako například druh tisku, období nebo jeho aktuální stav. Dále zde najdeme dokumenty Senátu, výborů a komisí, které lze před zobrazením taktéž filtrovat, a legislativu EU, což znamená seznam všech projednávaných dokumentů EU. Důležité jsou také záznamy o hlasování, které se zobrazí po vyplnění parametrů vyhledávání jako výběr období a případně číslo schůzce. Zobrazí se přehled dokumentů, o nichž se hlasovalo, na které schůzi to bylo a také číslo hlasování. Nejdůležitější je výsledek hlasování u něhož si lze zobrazit detail. Další zajímavostí zde uváděnou jsou těsnopisecké zprávy, v nichž najdeme stenozáznamy (přepis pronesených řečí poslanců) ze schůzí Senátu a stenozáznamy z veřejných slyšení. Zbývají už jen informace jako komentáře k zákonům, zákony vrácené Senátem, důležité dokumenty a podobné.

Čtvrtá kategorie *informace a zajímavosti* slouží jako studnice informací pro média a veřejnost, která by se ráda dozvěděla něco z historie Senátu, něco o budovách a prostředí, ve kterých sídlí Senát, nebo pro ty, kteří by rádi Senát jeli navštívit. Nabízí se zde také časopis Senát, kulturní akce, fotogalerie, videogalerie a spousta dalších věcí.

Poslední část pojednává o kanceláři Senátu. Je zde uveden kontakt, organizace, nabídky zaměstnání, informace o veřejných zakázkách a dokonce nabídka nepotřebného majetku.

3.2.2 Stránky Poslanecké sněmovny

Jak již bylo řečeno, struktura stránek Poslanecké sněmovny a Senátu je víceméně podobná, liší se pouze v orientaci dat na daný orgán Parlamentu a také ve formátu uložení dat. Nyní zde budou popsány především tyto rozlišnosti s tím, že zbytek je podobný jako v textu výše. Na stránkách Poslanecké sněmovny se nejprve dostaneme na výchozí domovskou stránku, kde se dozvídáme základní informace o PS, kalendář akcí, aktuality a především nástroj pro filtrování a zobrazení tisků Poslanecké sněmovny podle typu tisku nebo doby jeho podání.

Kategorie *Poslanecká sněmovna* sdružuje základní informace o jejím fungování: popis voleb, jejich výsledky a organizace. Dále zde najdeme popis jednání PS, jednacího řádu, proces přijímání zákonů, přehled zákonodárné činnosti, historii parlamentarismu a ústavnosti a spoustu dalších informací přibližující a vysvětlující činnost PS.

Druhá kategorie s názvem *Poslanci a orgány* představuje předsedu a místopředsedu PS a dále jednotlivé poslance a to buď podle poslaneckých klubů nebo podle abecedního seznamu jejich jmen. Další možností je i vyhledávání poslanců podle výborů, komisí, stálých delegací, dle meziparlamentních skupin přátel a dle krajů. Po otevření stránky konkrétního poslance se zobrazí nejdůležitější údaje o něm, jako členství ve výborech, podvýborech a klubech včetně data počátku členství. Dále také, kterých návrhů je předkladatel, jeho podané písemné interpelace (přepis řeči), jeho řeč ve sněmovně a přehled hlasování.

Třetí kategorie se nazývá *Jednání a dokumenty* a, jak z názvu vyplývá, je zde přehled schůzí PS, jejich program, kalendář akcí (harmonogram jednání i týdenní program) a zvukové i audiovizuální záznamy ze schůzí; znovu odkaz na vyhledávání sněmovních tisků, interpelací, dokumentů a zákonů, stejně jako zasedací pořádek a hlasování. Důležitý je odkaz na stenoprotokoly (přepis jednání), které jsou řazeny dle jednotlivých schůzí nebo podle jmenného rejstříku poslanců.

Čtvrtá kategorie pojednává o veřejnosti a o médiích. Jak již z názvu vyplývá, jsou zde například fotogalerie, informace pro návštěvníky, infocentrum, Parlamentní knihovna, výroční zprávy politických stran a další.

Poslední kategorií je *Kontakt*. Jsou zde uvedeny důležité kontaktní informace a adresa Poslanecké sněmovny.

3.3 Formát a způsob uložení dat na stránkách Parlamentu ČR

Většina informací na stránkách Poslanecké sněmovny a Senátu je dostupná a prezentovaná pouze ve formátu HTML. Jsou zde však některé výjimky, například hlasování senátorů za různá období lze stáhnout ze stránek Senátu v podobě zip archivu obsahující data ve formátu XML[6].

6 http://www.senat.cz/informace/pro_media/statistiky/hlasovani_xml.php?ke_dni=9.6.2015&O=10

Na stránkách poslanecké sněmovny je k nalezení stránka [13], kde se dají stáhnout data Poslanecké sněmovny a Senátu. Na zveřejňování dat se podle informace zde uvedené stále pracuje a další další data budou postupně přidávána. Je zde ke stažení poměrně velké množství dat, a to agenda poslanců a osob, hlasování, sněmovní tisky, interpelace a senátní tisky. Data jsou ve formátu UNL – každý řádek v souboru tedy odpovídá jednomu záznamu v databázi, oddělovačem je znak roury (|) [13]. Nejsou zde však uvedena všechna potřebná data pro tuto práci, proto nebudou použita. Dále se budeme zabývat pouze způsobem uložení dat, která budou pro tuto práci potřebná.

S výhodou lze využít to, že na stránkách Poslanecké sněmovny lze tisky vyhledávat jednak podle filtrů, ale především podle čísla tisku. Všechny tisky aktuálního volebního období lze nalézt přes jednotný URL odkaz, kde cislo_tisku odpovídá číslu tisku. Tento odkaz je uveden v ukázce 3.1.

```
http://www.psp.cz/sqw/historie.sqw?o=7&t=cislo_tisku
```

Ukázka 3.1: Odkaz na informace o tisku.

Takto nalezneme vždy stránku se základními informacemi o tisku a jeho průchodu Parlamentem[7]. Tohoto lze využít například při automatizovaném zpracování a vyhledávání informací. Konkrétní znění daného tisku je obvykle dostupné ve formátu pdf a doc. Takto jsou uloženy tisky jak na stránkách PS, tak Senátu, avšak na stránkách PS jsou informace lépe strukturované a přehlednější, a proto budeme v tomto ohledu přednostně pracovat se stránkami PS. V základním přehledu o tisku ve formátu HTML se dozvídáme, kdo je předkladatelem tisku, o jaký typ dokumentu se jedná, jeho název, kdy byl předložen, kdy se dostal k podpisu prezidenta; případně kdy byl ukončen nebo jestli stále probíhá. Na konci stránky jsou většinou uvedena klíčová slova tisku. Všechny tyto informace jsou rozčleněny do několika sekcí, lišící se podle typu dokumentu a jeho stavu; například u schváleného návrhu zákona, který nebyl vrácen, jsou to sekce: *Předkladatel*, *Poslanecká sněmovna* (údaje o průchodu PS), *Senát* a *Prezident*. Jedná-li se o dokument, je přítomna sekce *Dokument* místo *Předkladatel*. Pokud například Senát vrátil návrh do Poslanecké sněmovny, přibyde zde další sekce *Poslanecká sněmovna*. Pokud je návrh zamítnut nebo stále v jednání, je to uvedeno v textu a také barevně rozlišeno, červeným podbarvením pro probíhající návrh a černým podbarvením pro ukončený návrh, v diagramu průchodu Parlamentem. Nové tisky dostávají po řadě vyšší číslo než tisky před nimi. Nevýhodou je ovšem to, že tisky jsou postupem času aktualizovány a tuto skutečnost lze zjistit buď neustálým prohledáváním všech těchto odkazů, což by bylo velmi časově náročné, nebo na zcela jiném odkazu[8], kde jsou jednotlivé aktualizace uvedeny ke každému dni spolu s odkazem na detail tisku.

7 Například tisk 467: https://www.psp.cz/sqw/historie.sqw?o=7&t=464

8 https://www.psp.cz/sqw/tisky.sqw?tx=1

Jména poslanců a jejich příslušnost do poslaneckých klubů pro aktuální volební období lze získat z řazení poslanců podle klubů. Údaje jsou zobrazeny ve formě tabulky. Pokud potřebujeme seznam poslanců za jiné volební období (informace na stránkách se vztahují vždy pouze k aktuálnímu), je třeba jít do digitálního repozitáře, poté do digitální parlamentní knihovny[9] a tady najdeme seznam jednotlivých volebních období a k nim všechny důležité informace jako tisky daného období, seznam poslanců a jejich hlasování, stenoprotokoly a další. Údaje zde sahají až do roku 1848, liší se však svojí komplexností.

Jména senátorů jsou na stránkách Senátu zobrazena taktéž ve formě tabulky všech jmen senátorů, příslušnosti do klubů a dalších dodatečných informací. Seznam senátorů zde lze zobrazit aktuální k vybranému dni a to pomocí pole do něhož zapíšeme požadované datum[10].

Stenozáznamy se na stránkách PS zobrazují ve formátu HTML. Je možné je vybrat podle daného jména poslance[11]. Poté se zobrazí seznam schůzí, na nichž daný poslanec vystoupil, a pod každou schůzí název tisků, ke kterým promlouval, a odkazy na části textu stenozáznamu, kde se jeho řeč objevuje. Někdy promlouvá pouze jednou, jindy na daném odkazu nalezneme více vystoupení a někdy projev pokračuje až na další stranu stenozáznamu dané schůze.

9 https://www.psp.cz/eknih/index.htm

10 http://www.senat.cz/senatori/index.php?ke_dni=20.4.2016&O=10&lng=cz&par_2=2

11 http://www.psp.cz/eknih/2013ps/rejstrik/jmenny/index.htm

4 Vhodné statistické metody a nástroje pro reprezentaci dat

V této kapitole budou popsány jednotlivé statistické metody pro reprezentaci dat uživateli, které budou použity v této práci a dále nástroje pro reprezentaci statistických dat.

4.1 Statistické metody

Dále budou popsány jednotlivé použité statistické veličiny jako aritmetický průměr, vážený průměr a četnosti.

4.1.1 Aritmetický průměr

Jedná se o jednu z nejčastějších charakteristik polohy (charakteristika polohy jsou čísla, která charakterizují úroveň hodnot znaku v statistickém souboru) [14]. Je definován jako součet všech hodnot vybraného znaku vydělený celkovým počtem všech znaků nebo-li rozsahem souboru. Má-li statistický soubor rozsah n a statistický znak X nabývá hodnot $x_1, x_2, \ldots x_n$, potom aritmetický průměr je dán vzorcem 4.1. Zdroje [14] a [15].

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Vzorec 4.1: Aritmetický průměr.

4.1.2 Vážený aritmetický průměr

Aritmetické průměry souborů s rozdílnými rozsahy je třeba „vážit“ rozsahem příslušného souboru. Potom vážený aritmetický průměr z průměrů více souborů vypočítáme podle vzorce 4.2. Zdroj [14].

$$\overline{x}_v = \frac{\overline{x}_1 \cdot n_1 + \overline{x}_2 \cdot n_2 + \ldots + \overline{x}_k \cdot n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} \overline{x}_i \cdot n_i}{n}, \text{kde}$$

$n_1, n_2, \ldots, n_k$ *jsou rozsahy souborů*
$\overline{x}_1, \overline{x}_{2,\ldots}, \overline{x}_k$ *jsou aritmetické průměry souborů*
$n = n_1 + n_2 + \ldots + n_k$

Vzorec 4.2: Vážený aritmetický průměr.

4.1.3 Četnosti

Existují dva typy četností [15]:

1. Absolutní četnost – celkový součet všech případů dané hodnoty.
2. Relativní četnost – procentuální vyjádření absolutní četnosti.

4.2 Nástroje pro reprezentaci dat

Pokud chceme vyjádřit četnost, použijeme reprezentaci pomocí čísla nebo procent. Dalším častým nástrojem pro vyjádření výsledků některých statistických metod jsou tabulky. Poskytují velmi přehledný a elegantní způsob reprezentace dat.

Highcharts[12]

Jedná se o jednoduché a flexibilní, přesto profesionální a propracované API pro tvorbu různých typů grafů, které se dají použít například na webových stránkách. Grafy jsou velmi přehledná varianta pro reprezentaci statistických výsledků. Lze použít například koláčové, bublinové i méně tradiční prostorové a další zajímavé grafy.

Chart.js[13]

Chart.js je javascriptová knihovna pro tvorbu grafů. Lze ji jednoduše vložit do webové stránky a načítat a zobrazovat data. Neobsahuje tolik možností, typů grafů a přizpůsobení jako Highcharts.

Google charts tools[14]

Další možností je využít nástrojů od firmy Google. Jedná se o interaktivní grafy zaměřené na použití ve webovém prohlížeči a v mobilních telefonech. Grafy jsou kompatibilní s HTML5, použitelné na iOS i Android zařízeních. Je zde nabízena řada nástrojů pro propojení dat a grafů v reálném čase. Vhodný typ grafu si lze vybrat z galerie dostupných grafů a pak ho pomocí javascriptu vložit do webové stránky.

12 http://www.highcharts.com/

13 http://www.chartjs.org/

14 https://developers.google.com/chart/

5 Návrh aplikace

Tato kapitola se zabývá návrhem aplikace. Bude zde uvedena vnitřní struktura, charakteristika a principy činnosti aplikace.

5.1 Použité technologie

PHP[15]

Jazyk PHP je velmi rozšířený a oblíbený. V dnešní době s ním pracuje většina hostingových služeb a tak většinou nejsou problémy s kompatibilitou a nemožnosti použití. V tomto jazyce bude programována většina aplikace: všechny moduly starající se o extrakci dat ze stránky, jejich zpracování i uložení zpět do databáze. PHP bylo vybráno pro tuto práci nejen kvůli své velké rozšířenosti, kompatibilitě s frameworky, dalšími nástroji a knihovnami, ale i proto, že je velmi dobře dokumentováno v elektronické podobě i ve formě knižních publikací.

Nette[16]

Tato práce bude psána v prostředí Nette frameworku za účelem zpřehlednění a zjednodušení zdrojového kódu. Bude taktéž použit šablonovací systém Latte, jenž mimo jiné eliminuje bezpečnostní rizika. Celá práce se takto stane jednak přehlednější pro údržbu a úpravu kódu a jednak se zlepší rozšiřitelnost a znovupoužitelnost kódu.

MariaDB[17]

Data budou ukládána do databáze MariaDB. Jedná se o odnož databázového systému MySQL. Její výhodou je, že je zdarma a pod GNU GPL licencí. Zachovává kompatibilitu s MySQL a tudíž většinou není problém s nasazením na běžném hostingu. Z těchto důvodů byla vybrána právě tato databáze pro uchovávání údajů.

Composer[18]

Jedná se o nástroj na správu závislostí v PHP. Prostřednictvím něhož můžeme definovat libovolně složité závislosti jednotlivých knihoven a on je pak za nás nainstaluje.

15 http://php.net/manual/en/

16 https://nette.org/cs/

17 https://mariadb.org/

18 https://doc.nette.org/cs/2.3/composer a https://getcomposer.org/

JSON

Bude použit jako transportní formát během výběru dat z databáze a jejich vykreslení do grafů a tabulek. Byl vybrán kvůli své jednoduchosti, přehlednosti a podpoře javascriptových knihoven.

Highcharts

Tato již dříve zmiňovaná technologie pro jednoduché zobrazování grafů a přehledného prezentování dat uživateli bude použita v této práci kvůli svému profesionálnímu vzhledu a velkému výběru typu grafů. Byla vybrána také proto, že výsledný graf jde ve svém zdrojovém kódu upravovat, a tak můžeme docílit přesně takového grafu, jaký je potřeba.

Bootstrap[19]

Jedná se o jeden z nejpopulárnějších CSS, HTML a JS frameworků pro vývoj responzivních aplikací. Díky němu se grafické prvky na stránce přizpůsobí zařízení, ať už jsou prohlíženy na telefonu, tabletu či na notebooku. Tyto prvky jsou navíc velmi vzhledově přitažlivé pro uživatele. V této práci bude použit pro tabulky, obrázky a další prvky. Bootstrap navíc definuje velkou škálu CSS tříd pro vytvoření základního designu.

Select2[20]

Toto rozšíření JQuery knihovny je zaměřené na select boxy. Umožňuje select boxy více přizpůsobit, například tak, že je možné vytvořit select boxy umožňující výběr více možností s nabídnutou nápovědou uživateli.

JQuery[21]

Jedná se o javascriptovou knihovnu, která velmi zjednodušuje práci s javascriptem. Díky ní se práce stane daleko jednodušší a přesnější. Další velkou výhodou je, že minimalizuje rozdíly mezi jednotlivými prohlížeči. V této práci využijeme především její schopnost zachycovat a reagovat na události a poté měnit strukturu stránky spolu se schopností pracovat s ajaxem.

Majka[22]

Jedná se o morfologický analyzátor slov vyvíjený na FI Masarykovy univerzity v Brně. Bude použit pro analýzu slov a převedení na základní tvar slova nebo-li lemma.

19 http://getbootstrap.com/

20 https://select2.github.io/

21 http://jquery-navod.cz/kategorie-ostatni-clanky/1-uvodni-clanek

22 https://nlp.fi.muni.cz/ma/

5.2 Přehled provedených statistik

Cílem práce je provést několik statistik z dat získaných ze stránek Parlamentu ČR. Byly vybrány tyto statistiky:

1. Délka projednávání návrhů zákonů za aktuální volební období (2013 – 2016).

   Zde se zaměříme na již schválené návrhy zákonů v tomto volebním období. Co se týká typů tisků, budou brány v potaz poslanecké, senátní i vládní návrhy zákonů, mezinárodní smlouvy a návrhy státního rozpočtu. Naopak nebude zahrnuto zákonné opatření Senátu, zprávy, interpelace, stanoviska vlády, usnesení výborů a návrhy, které byly podány zastupitelstvem některého kraje. Posledně zmíněné typy tisků budou odfiltrovány ještě před zpracováním informací, neboť nejsou relevantní vzhledem k účelu této statistiky. Jejím cílem je ukázat závislost toho, kdo návrh podával a jeho příslušnosti do koalice, opozice či skupiny obou, na potřebné době ke schválení návrhu zákona. Bude tedy třeba zjistit podavatele tisku a vyhodnotit jestli patří do koalice, opozice nebo jde o smíšený návrh a také zjistit dobu, za kterou návrh prošel od podání až k podpisu prezidenta. Výsledkem bude graf zobrazující na ose X datum podání návrhu a na ose Y počet dní průchodu Parlamentem. Barevně budou rozlišeny koaliční, opoziční a smíšené návrhy. Pro ještě lepší přehlednost budou počty přijatých zákonů a jejich průměrná doba schvalování zobrazena pro koalici, opozici a mix v tabulce.
2. Délka projednávání návrhů zákonů za minulé volební období (2010 – 2013).

   Jelikož aktuální volební období nezahrnuje ještě tolik tisků, stejná statistika jako v předchozím bodě bude vypracována také pro minulé volební období. Výhodou tedy je, že lze tyto období srovnat a také se podívat na statistiku, která je vyhotovena z většího objemu dat.
3. Další statistiky aktuálního volebního období.

   Tato statistika navazuje na předchozí a ukazuje další zajímavosti a detaily aktuálního volebního období. Patří sem graf všech podaných návrhů zákonů, tedy i těch neúspěšných nebo stále v jednání. Zpracovávané typy tiskopisů jsou stejné jako v bodě 1. Tento graf zobrazuje procentuální podíl podaných návrhů poslanci (jednotlivcem, skupinou), senátory, vládou a krajskými zastupitelstvy z celkového počtu podaných návrhů. Tyto informace jsou také shrnuty v tabulce, kde ke každé skupině jsou uvedeny počty a jejich procentuální podíl. Další statistika se zaměřuje na ukázání rozdílů mezi podanými, zamítnutými a aktuálními návrhy v závislosti na podavateli (poslanci, senátoři, vláda, krajská zastupitelstva). Poslední statistika této série si klade za cíl ukázat stejné závislosti jako v předchozím případě, avšak znovu rozdělené na koalici, opozici a smíšené skupiny. Takto lze porovnat, kdo podává

návrhy nejčastěji, komu se nejrychleji schvalují, komu se schvalují s největší pravděpodobností a další zajímavosti.

4. Řečnictví v Poslanecké sněmovně.

   Zde se v stenozáznamech Poslanecké sněmovny zaměříme na rozpravy, které se vedly k nějakému sněmovnímu tisku. Jednotlivá slova projevů poslanců budou spočítána. Výstupem bude počet slov u každého poslance, které ve sněmovně pronesl ve vztahu k návrhům zákonů a taky jeho počet vystoupení, nebo-li počet návrhů, které komentoval. Tyto výsledky si uživatel bude moci zobrazit vlastním výběrem a také budou pro zajímavost vyhotoveny tabulky zachycující několik nejlepších řečníků s nejvíce slovy v celé Poslanecké sněmovně a také několik řečníků za každou politickou stranu.

5. Diskutovanost návrhů zákonů.

   Zde budeme uvažovat stejné typy tisků jako v předchozích případech a všechny typy tisků, co se týká jejich stavu. Každý tisk bude charakterizován počtem slov, které k němu pronesli všichni poslanci během jeho projednávání napříč průchodem Poslaneckou sněmovnou. Tisky budou zobrazeny v podobě datových řad, kde na ose Y bude diskutovanost tisku (tedy počet slov) a na ose X bude daný tisk ohraničen dvěma body – datem podání a datem podpisu prezidenta u úspěšných návrhů, datem zamítnutí u ukončených a aktuálním datem u probíhajících zákonů. U každého tisku lze zobrazit detail ukazující přesná data, přesný počet slov a také klíčová slova návrhu. V závěru tedy budeme moci z grafu odvodit počty pronesených slov a dobu průchodu Parlamentem v závislosti na klíčových slovech tedy tématu návrhu zákona. Jelikož je takovýchto dat velké množství, je třeba graf rozdělit na různá časová období podle podání návrhů a také podle diskutovanosti na více a méně diskutované návrhy v daném období. Tato kritéria si bude moci uživatel sám zvolit. Případně si bude moci vybrat z nabízených klíčových slov a zobrazit návrhy s danými klíčovými slovy za celé aktuální volební období. Nakonec je v této statistice představeno deset nejdiskutovanějších návrhů zákonů v tomto volebním období a jejich klíčová slova.

6. Váhování IDF.

   Texty týkající se projevů k tiskům v Poslanecké sněmovně budou upravovány do podoby jednotlivých lemmat, na která bude aplikována váhovací technika IDF. Bude představeno několik slov s nejvyšší hodnotou IDF i, pro ukázku, s nejnižší hodnotou.

7. Profesní slovní zásoba.

   Každému poslanci, který komentoval nějaký tisk v Poslanecké sněmovně, bude spočítána profesní slovní zásoba, tedy počet různých slov (nepočítá se ohýbání slov), které pronesl. Jelikož však ti, kteří mluvili jen málo, mají pochopitelně nižší tuto slovní zásobu. Nakonec

bude spočítán vážený průměr slovních zásob poslanců. Také budou představena nejčastější slova, která v Poslanecké sněmovně vzhledem k tiskům zazněla.

5.3 Získávání nových dat

Jelikož jsou grafy a statistiky zmíněné v bodě 1 a 3 v předchozí podkapitole zaměřené na aktuální volební období a je zde žádoucí mít co nejvíce dat, bude prováděna automatická aktualizace podkladových dat pro tyto statistiky a tím i zobrazované výsledky budou aktuální. Zbylé statistiky jsou natolik výpočetně náročné a přínos neustálé aktualizace by byl tak malý, že u nich nebude na webhostigu, kde budou výsledky zobrazeny, aktualizace prováděna. Aktualizace se bude dát provést spuštěním odpovídajících modulů, které uloží nová data do db nebo aktualizují stará. Odtud budou již data v aktualizované podobě získávána pro prezentaci statistik.

Získávání nových dat a aktualizace starých je prováděna pomocí dvou modulů. První modul bude v pravidelných časových intervalech (například několik dní nebo po přístupu návštěvníka na webovou prezentaci projektu) zjišťovat nově přidané tisky a tisky, které se nějakým způsobem změnily. Nejprve tento modul zjistí z databáze datum, které naposledy analyzoval. Toto datum vyhledá na stránce, kde jsou uloženy záznamy o aktualizacích tisků (byla popsána v podkapitole 3.3) a od tohoto data nalezne identifikační čísla záznamů, které se změnily a uloží je do databáze. Poslední zpracované datum uloží pro použití při dalším vyhledávání aktualizací. Výsledkem je tedy tabulka identifikačních čísel tisků, které se změnily a je tedy třeba je znovu zpracovat.

Moduly Poslanci a Senátoři se taktéž budou v časovém intervalu zhruba jedenkrát za týden aktualizovat, neboť je třeba kontrolovat, jestli některý poslanec či senátor nezměnil své jméno (například po svatbě) nebo příslušnost k politickému klubu.

5.4 Extrakce dat a jejich zpracování

Modul navazující na ten z předešlé podkapitoly očekává jako vstup tabulku databáze obsahující čísla tisků buď nových nebo aktualizovaných tak, jak bylo její vytvoření popsáno v předešlé podkapitole. Z jednotlivých identifikačních čísel návrhů modul vytvoří odkaz a načte data z odkazované stránky. Vzhledem k velkému množství dat, které se ze stránek stahuje bude třeba již stažená data uložit, aby v případě jejich potřeby byly vzaty z cache a ne znovu načítány vzdáleně ze stránek. Pokud daná stránka ještě nebyla použita, data jsou načtena ze stránky a uložena, příště již budou přístupná a automaticky se načtou z cache (pro tento účel určené složky). V této chvíli je nutné extrahovat potřebné informace pro statistiky popsané v bodě 1, 2 a 3 v předešlé podkapitole z HTML stránky tisku a uložit je do databáze. K extrakci bude použit DOM parser a pro některé specifické informace také vlastní analyzátor.

Co se týká zbylých statistik, potřebná data budou získána odlišným způsobem. Pro každého poslance v aktuálním volebním období budou vyhledány ze stenozáznamů schůzí Poslanecké sněmovny jeho projevy k jednotlivým návrhům zákonů. Bude spočítán počet jeho projevů a počty slov, kterým přispěl k různým návrhům zákonů. Takto získáme jednak počet slov u všech poslanců a také celkové počty slov, které poslanci pronesli k daným návrhům a uložíme je do databáze. S jednotlivými slovy budeme navíc kromě počítání pracovat, abychom mohli využít váhovací techniky IDF a také spočítat slovní zásobu poslanců.

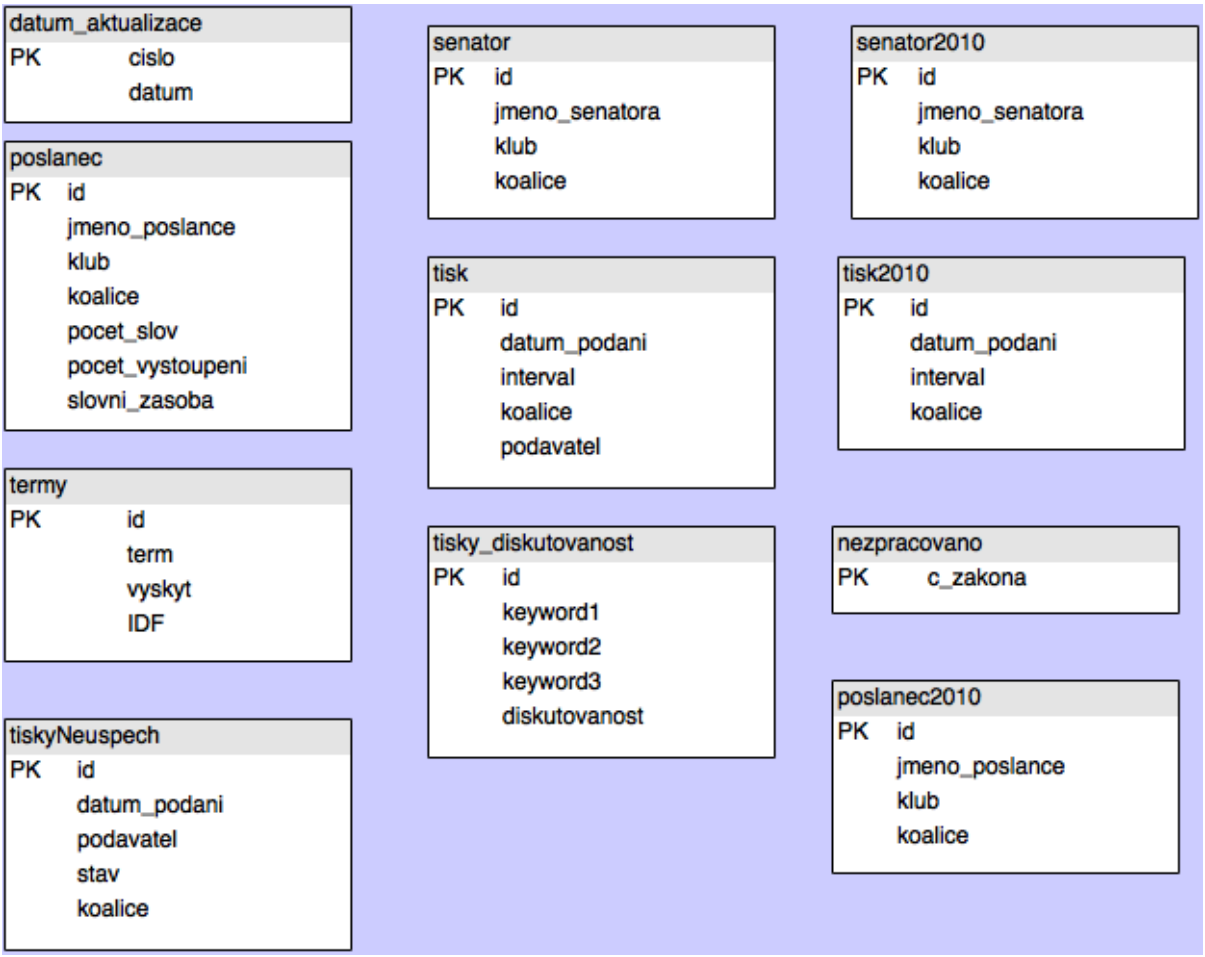
Dále bude potřeba modul, který ze stránek Poslanecké sněmovny zjistí jména a příslušnost do poslaneckých klubů poslanců a uloží vše do databáze. Toto bude třeba provést i pro senátory, zde budou data vyhledávána na stránkách Senátu, a to v obou případech pro aktuální i minulé volební období (tedy léta 2010 – 2013).

Poslední popisovaný modul stáhne ze stránek Poslanecké sněmovny všechny aktuálně dostupné tisky, nalezne a získá jejich klíčová slova a tyto informace uloží taktéž do databáze.

V databázi bude několik tabulek sloužících pro uchování extrahovaných informací, a to:

- **datum_aktualizace** – poslední datum, kdy byla provedena aktualizace,
- **poslanec** – id, jména a příjmení všech poslanců, jejich klub, jestli patří do koalice nebo do opozice, počet pronesených slov, počet komentovaných návrhů zákonů a slovní zásoba,
- **poslanec2010** – id, jména a příjmení poslanců za minulé volební období a jejich klub spolu s informací o příslušnosti do koalice,
- **senator** – id, jméno a příjmení senátorů, příslušnost do klubu a do koalice,
- **senator2010** – stejné informace jako v tabulce senator, ale pro období od roku 2010 až 2013,
- **termy** – jednotlivé termy, počet dokumentů, ve kterých se term vyskytl a vypočítaná hodnota IDF,
- **tisk** – záznamy o přijatých tiscích; jejich id, datum podání návrhu, počet dní do konce, podavatel a příslušnost podavatele do koalice,
- **tisk2010** – záznamy o tiscích jako výše ale pro minulé volební období,
- **tiskyNeuspech** – tisky zamítnuté nebo stále v projednávání, jejich id, datum podání, podavatel a jeho příslušnost do koalice a také stav tisku,
- **tisky_diskutovanost** – id všech návrhů zákonů, jejich klíčová slova a počty slov, které k nim poslanci pronesli,
- **tmp_termy** – pomocná tabulka pro vyhodnocení slovní zásoby poslanců, po dokončení výpočtu je prázdná.

Schéma databáze je také patrné z ilustrace 5.1. Jedná se o soubor tabulek, které nejsou vzájemně propojeny cizími klíči, neboť charakter aplikace to nevyžaduje.



Ilustrace 5.1: Schéma databáze.

5.5 Tvorba statistik a prezentace výsledků uživateli

Z informací umístěných a aktualizovaných v této databázi se pomocí statistických metod vytvářejí statistiky, jejichž výsledky budou předávány skriptu, který zajišťuje vykreslování grafů. Výsledky budou umístěné na osobní webové stránce, kde k nim bude mít přístup široká veřejnost. Výsledky v podobě grafů, tabulek a čísel by měla být přehledná, intuitivní a uživatelům poskytovat zajímavý a neotřelý pohled na dění v Poslanecké sněmovně a Senátu.

6 Implementace

V této kapitole bude podrobněji rozebrán postup implementace jednotlivých částí aplikace. Výsledné statistiky budou prezentovány na osobní webové stránce *www.anerii.eu*.

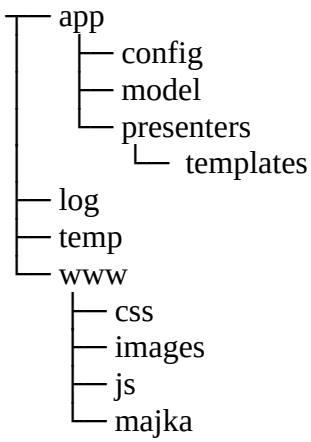
6.1 Struktura aplikace

Struktura výsledné aplikace je silně závislá na struktuře zdrojových stránek a v případě změny jejich struktury bude třeba upravit zdrojový kód aplikace. Pro eliminaci problému změny některé stránky těsně před dokončením práce a také dalších důvodů popsaných dále, byly důležité stránky ukládány do paměti cache.

Aplikace dodržuje návrhový vzor *model–view-presenter* což je podobné jako *MVC*, *model - view-controller*[23].

- *Model* – vrstva, která pracuje s daty. Je oddělena od zbytku aplikace, vrací data příslušným presenterům.
- *View* – vykresluje a zobrazuje požadovaná data pomocí šablon.
- *Presenter* – propojuje předchozí dvě vrstvy, dotazuje se na data modelu a předává je do *view* na vykreslení.

Adresářová struktura zdrojového kódu aplikace se skládá z následujících důležitých složek, vycházejících ze struktury Nette frameworku, jak je naznačeno v ukázce 6.1:



Ukázka 6.1: Adresářová struktura aplikace

Složka app sdružuje hlavní zdrojové kódy programu. Ve složce config nalezneme nastavení připojení k databázi a některá nastavení samotného Nette. Složky model a presenters obsahují

23 https://doc.nette.org/cs/2.3/quickstart/home-page

jádro aplikace a ve složce templates najdeme šablony prezenterů, což odpovídá již zmíněnému konceptu *model-view-presenter*. Složka log slouží pro logování aplikace a uložení chybových výpisků. Zaznamenávají se zde některé zpracované tisky a především případné chyby, které by se mohly vyskytnout. Ve složce temp najdeme záznamy již navštívených stránek pro pozdější přístup, aby se nemusely znovu stahovat z internetu. Složka www obsahuje další potřebné nástroje, javascriptové knihovny, použité obrázky a především aplikaci Majka. Soubor readme obsahuje dodatečné informace o aplikaci. Další soubory a složky obsažené ve zdrojovém adresáři, které zde nebyly popsány se váží více k Nette frameworku než ke zde popisované aplikaci, a proto nejsou zmíněny.

Aplikaci je možné rozdělit do několika modulů (tyto moduly nejsou chápány jako moduly v terminologii Nette), které pracují víceméně samostatně, a právě takto budou jednotlivé moduly, dále rozdělené na vrstvy *model-view-presenter*, popisovány. Pro správný běh aplikace, musí webový server splňovat některá kritéria, jako daná minimální verze PHP a některé důležité funkce serveru. Jejich detailní popis je k nalezení v [16].

Ve všech modulech se používají stejné funkce pro stažení HTML stránky a uložení do proměnné v programu, jak je naznačeno v ukázce 6.2:

```
$ch = curl_init($adresa);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
$html = curl_exec($ch);
curl_close($ch);
```

Ukázka 6.2: Získání obsahu HTML stránky.

Proměnná $adresa obsahuje URL adresu požadované stránky a $html výsledný řetězec vytvořený z obsahu webové stránky. Problémem však na všech stránkách Poslanecké sněmovny je, že zde není uvedena znaková sada, a tudíž se některé znaky nezobrazují správně a ani DOM parser nepracuje správně, a to přestože funkci na parsování přímo z HTML stránky má. Musí se tedy vždy použít funkce, která nejprve vše převede na kódování UTF-8, a až poté je možné s výsledným řetězcem pracovat.

6.2 Třída DatabaseModel

Tato třída poskytuje jednotné rozhraní pro ukládání informací do databáze ze všech modulů. Takto jsou všechny funkce měnící obsah databáze na jednom místě a je daleko snadnější tyto činnosti kontrolovat a v případě potřeby upravovat. Vytvořením pouze jednoho rozhraní pro provádění změn v databázi se kód stává přehlednější. Jednotlivé metody této třídy budou však vysvětleny spolu s moduly, které jejich funkce využívají.

6.3 Modul poslanci

Tento modul se stará o získání seznamu všech poslanců a jejich příslušnosti do poslaneckých klubů za aktuální volební období i za to minulé. Takto získaná data jsou dále využívána dalšími moduly. Tento modul sám o sobě žádná data uživateli neprezentuje, a proto k němu žádná vrstva *view* ani *presenter* nenáleží.

Model

Srdcem *modelu* je cyklus for, který po řadě načítá stránky jednotlivých poslaneckých klubů přítomných v Poslanecké sněmově. Každá stránka klubu má vlastní, neměnný odkaz, který si musí program pamatovat. Celá HTML stránka se uloží do proměnné, se kterou musíme pracovat a upravovat ji, abychom získali požadované informace. Nyní potřebujeme odstranit vše, kromě jednotlivých záznamů o poslancích. K tomu použijeme funkci preg_match_all a regulární výraz ve tvaru, který je předveden v ukázce 6.3.

```
/<tr>\s*<th>.*?<\/a>/
```

Ukázka 6.3: Regulární výraz pro získání jednotlivých záznamů o poslancích.

Takto získáme pouze odkazy na stránky jednotlivých poslanců, jejichž část zachycující identifikační číslo poslance je přibližně ve tvaru, který zobrazuje ukázka 6.4. Třemi tečkami jsou označeny místa, kde pokračují znaky, které jsou v tomto významu nepodstatné.

```
<a href="detail.sqw?id=6151&...">Jméno Poslance...<\a>
```

Ukázka 6.4: Část získaných odkazů zachycující ID poslance.

Tyto získané výsledky projdeme v cyklu. Postupně aplikujeme regulární výrazy na odstranění nepotřebných částí a z každého odkazu získáme identifikační číslo poslance, pod kterým je uložen na stránkách Poslanecké sněmovny, a jeho jméno. Identifikační číslo, potřebné především pro ukládání seznamu poslanců do databáze a vytváření odkazů na stránky poslance, lze získat pouze z tohoto odkazu, neboť jinde uvedené není. Jelikož víme, který poslanecký klub právě zpracováváme, ke každému poslanci uložíme i údaje o klubu a tedy jeho příslušnosti do koalice či opozice. Nakonec všechny tyto informace za použití metody vloz_do_db_poslanec ze třídy DatabaseModel uložíme do tabulky poslanec a vždy inicializujeme sloupce pocet_slov, pocet_vystoupeni a slovni_zasoba na hodnotu 0 pro další použití v modulu zaměřující se na diskutovanost. Tento samý postup opakujeme i pro kluby, které byly přítomny v Poslanecké sněmovně za minulé volební

období. I zde platí, že každý má svůj vlastní odkaz. Inicializaci však neprovádíme, neboť diskutovanost a řečnictví těchto poslanců za toto volební období není počítáno.

6.4 Modul senátoři

Tento modul naplňuje tabulku `senator` v databázi identifikačními čísly senátorů, jejich jmény, senátorskými kluby a příslušností do koalice či opozice.

Model

Seznam všech senátorů spolu s podrobnými údaji si lze na stránkách Senátu zobrazit aktuální k libovolnému dni. Výhodou je, že zde se formát odkazů nemění tak, jako tomu bylo u Poslanecké sněmovny. Mění se pouze datum zahrnuté v odkazu – to, ke kterému se má seznam senátorů zobrazit aktuální. Například odkaz z ukázky 6.5 zobrazí senátory, kteří byli členy Senátu dne 1.1.2012.

```
http://www.senat.cz/senatori/index.php?ke_dni=1.1.2012&O=10&lng=cz&par_2=2
```

Ukázka 6.5: Odkaz pro zobrazení členů Senátu platných ke dni 1.1.2012.

Díky tomuto nemusíme měnit odkazy a pro získání dat aktuálního a minulého volebního období stačí například pro každý rok v těchto obdobích upravit odkaz a stránku načíst. Každou takto staženou stránku uložíme do řetězce a použijeme DOM HTML parser pro získání sloupců jmen a klubů z tabulky všech poslanců. Použijeme metodu `find`, jako je znázorněno v ukázce 6.6.

```
$dom->find('tr td[2] a, tr td[4]')
```

Ukázka 6.6: Funkce find pro získání sloupců tabulky.

Pomocí této metody vyhledáme data z druhého a čtvrtého sloupce tabulky a následně vše uložíme do jednoho řetězce. Zjistíme id senátora z odkazu regulárním výrazem, který je naznačen v ukázce 6.7.

```
/<a href=.*?=(\d+)">/
```

Ukázka 6.7: Regulární výraz pro získání id senátora.

Dále postupně odstraňujeme přebytečné HTML značky, až získáme jen senátorovo jméno a jeho senátorský klub. Podle klubu a podle data, ke kterému se poslanci zobrazují, se určí příslušnost do koalice či opozice senátora (v různých letech jsou koaliční smlouvy podepsány mezi jinými

stranami). Získané informace nakonec uložíme do databáze; do tabulek senator pro období 2013 - 2016 a senator2010 pro období 2010 - 2013 za použití metody vloz_do_db_senator třídy DatabaseModel.

Moduly poslanci a senátoři jsou konstruovány tak, aby jednou za čas samostatně prováděly aktualizaci seznamů poslanců a senátorů. Je to důležité, neboť někteří poslanci mohou změnit politický klub nebo se během volebního období přejmenovat. Je tedy žádoucí, aby moduly tyto skutečnosti hlídaly a případně provedly změnu v databázi. Moduly budou jednou týdně spouštěny hostingovou službou cron, která je automaticky v předem daný čas spustí. Po spuštění extrahují požadované informace, a pokud se budou lišit od těch uložených v databázi, provedou patřičné úpravy. Na stránkách Poslanecké sněmovny zůstávají URL odkazy stále stejné, zatímco URL na požadovanou stránku Senátu sestavíme z funkce pro zjištění aktuálního data a neměnné části stále stejného odkazu, jak je v ukázce 6.5.

6.5 Modul zjištění nových tisků

Jedná se o třídu ZjistiNoveDokumenty. Tento modul by měl být spuštěn pouze jednou, a to ještě před ostatními moduly, jelikož provádí přípravu databáze tak, aby další části mohly na jeho činnosti stavět. Vychází ze stavu prázdné databáze, pouze tabulky poslanec a senator již mohou být naplněny daty. Jeho činnost je prostá. Do tabulky nezpracovano uloží všechna identifikační čísla tisků aktuálního volebního období, která existují.

Model

Jak již bylo řečeno, na stránky jednotlivých tisků se lze dostat přes jeden odkaz, u něhož se mění pouze poslední část značící číslo tisku. Princip činnosti modulu je tedy takový, že pomocnou proměnou $cislo, značící hledané číslo tisku, na začátku cyklu nastavíme na hodnotu 1, tedy první možné číslo tisku. Vytvoříme odkaz pomocí této proměnné, stáhneme odkazovanou stránku a na konci hodnotu proměnné $cislo zvýšíme o 1 a celý cyklus necháme proběhnout znovu. Jelikož se již zde stahují celé stránky tisků, se kterými budeme dále taktéž pracovat, již zde si budeme všechna data ze stažené stránky ukládat do cache, kde klíč tvoří URL odkazu. Takto zajistíme to, že pokud se odkaz nenalezne v cache a tím i požadovaná data, je třeba z něj stránku stáhnout a uložit.

Jakmile máme dostupná data ze stránky, musíme z obsahu zjistit, jestli tento tisk existuje. Pokud neexistuje, pak je v textu tato informace uvedena (v těle HTML najdeme informaci: Tisk $cislo neexistuje.). V případě, že existuje, uložíme jeho číslo do databáze. Nabízela by se možnost ukončit provádění cyklu a hledání existujících tisků, jakmile se nalezne první tisk, který neexistuje, avšak občas se objeví situace, kdy je tisk z nějakého důvodu smazán a tak existují tisky i po něm. Proto nastavíme limit, že pokud pět po sobě jdoucích tisků nebude existovat, dostali

jsme se k nejvyššímu číslu existujícího tisku a činnost cyklu a tím i modulu skončí. Číslo pět bylo vybráno z toho důvodu, že v celé sadě tisků existuje pouze jeden takovýto smazaný dokument; takže pokud by ještě nějaký přibyl, byla by dostatečně velké rezerva, a pokud by se náhodou stalo, že by neexistovalo více tisků po sobě, dozvěděli bychom se tuto informaci z logovacích souborů.

6.6 Modul zjištění aktualizací

Tento modul dokáže zjistit, jaké tisky byly aktualizovány nebo nově přidány od posledního data, které má uložené jako datum poslední aktualizace. Předpokládá tedy, že předchozí modul nalezl všechny existující tisky a tento modul bude již vybírat ke zpracování ty, které se změnily. Datum, kdy proběhlo vyhledávání všech tisků se na počátku nastaví jako datum poslední aktualizace a od něj už si datum tento modul hlídá sám. Osvědčilo se spuštění modulu po každém přistoupení uživatele na stránku projektu, neboť takto má uživatel vždy aktuální informace a běh skriptu není nijak výpočetně náročný. Pokud by se však zvýšil počet přístupů na stránku projektu, výhodnější by bylo jej spouštět automaticky pomocí služby cron například jednou za několik dní.

Model

Nejprve z databáze zjistíme datum poslední aktualizace a až k tomuto datu budeme tisky brát v úvahu. Vytvoříme cyklus `while`, který bude končit, jakmile nastavíme příznak ukončení cyklu. Nově přidané tisky se zobrazují na pěti stranách, z nichž každá má stejný odkaz lišící se pouze číslem strany. Sestavíme tedy odkaz pro první stránku, kde `$strana` značí číslo hledané strany. Toto je viditelné v ukázce 6.8.

```
'http://www.psp.cz/sqw/tisky.sqw?str='.$strana.'&O=7&PT=U&N=1&F=H&RA=20&tx=1'
```

Ukázka 6.8: Sestavení odkazu na stránky aktualizovaných tisků.

Jakmile máme data ze stránky stažené, použijeme DOM parser, abychom získali jenom některé sloupce tabulky, ve které jsou informace uloženy. Použijeme cyklus `foreach` a metodu `find` pro získání jen některých buněk. Jejich obsah spojíme do výsledného řetězce, jak je viditelné v ukázce 6.9.

```
foreach($dom->find('td[align="right"],td[colspan="4"]') as $e) {
        $retezec .= $e->plaintext . '<br>';
}
```

Ukázka 6.9: Získání vhodného obsahu tabulky.

Pomocí regulárního výrazu zjistíme nové datum poslední aktualizace, které uložíme do databáze. Tento výraz zachycuje ukázka 6.10.

```
/^\d+\.\s\D+\s\d{4}\b/
```

Ukázka 6.10: Regulární výraz pro zjištění data poslední aktualizace.

Poslední datum aktualizace hledáme na této stránce a pokud jej nalezneme, odstraníme všechny údaje pozdějšího data a nastavíme příznak ukončení cyklu. Získáme všechna čísla tisků, která byla aktualizována na aktuální stránce a uložíme je do databáze. Pro odstranění všech záznamů od hledaného data použijeme následující regulární výraz spolu s proměnnou PHP `$datum` a získaný výsledek nahradíme prázdným řetězcem, jak je patrné z ukázky 6.11.

```
/(?:^|<br>){$datum}.*/
```

Ukázka 6.11: Odstranění všech záznamů od data poslední aktualizace.

Pokud jsme na aktuální stránce nenašli hledané datum poslední aktualizace, probíhá cyklus znovu, zvýší se číslo stránky, až dokud neprojdeme i poslední stránku. Zde by činnost modulu skončila, i kdyby dané datum z nějakého důvodu nenašel.

6.7 Modul zpracování tisků

Jedná se o klíčový modul pro první část prováděných statistik, tedy statistik týkajících se především doby průchodu návrhů zákonů Poslaneckou sněmovnou. Tento modul je spuštěn vždy, jakmile se v tabulce `nezpracovano` nacházejí nějaká identifikační čísla tisků. Data získaná tímto modulem jsou základem pro některé statistiky, a proto se zde budeme věnovat i popisu jejich zobrazení ve vrstvách *presenter* a *view*. Nyní zde bude přiblížen princip činnosti tohoto modulu.

Model

Jak již bylo řečeno, výchozím předpokladem pro činnost tohoto modulu je tabulka `nezpracovano` naplněná identifikačními čísly tisků určených ke zpracování dříve popsanými moduly. Tisky se zpracovávají postupně od nejnižšího čísla zákona až do té doby, než jsou všechny tisky zpracovány a tabulka vyprázdněna. Nejprve potřebujeme načíst obsah patřičné stránky tisku, abychom z něj mohli získat důležité informace. Pokud bychom chtěli pouze testovat funkčnost programu, mohli bychom data vzít z cache, kam je uložil modul pro získávání nových tisků. Avšak pro normální

funkčnost a běh skriptu je třeba data stahovat přímo ze stránek Poslanecké sněmovny, protože jen tak zjistíme změny, které u tisků nastaly.

Dostali jsme se tedy do bodu, kdy máme data a potřebujeme z obsahu zjistit, o jaký typ dokumentu se jedná, abychom věděli, jak s ním dále naložit a jaká data z něj extrahovat. Nebudou zahrnuty zprávy, interpelace, stanoviska vlády, usnesení výborů a návrhy, které byly podány zastupitelstvem některého kraje. Všechny tyto tisky obsahují sekci *Dokument*. Použijeme tedy DOM parser, a pokud tuto sekci nalezneme, aktuálně zpracovávaný tisk se zahodí. Dále nás nezajímají zákonná opatření Senátu, jejichž obsah je odlišný, proto použijeme funkci, která v příslušné sekci vyhledá zmínku o tom, že se jedná právě o tento typ tisku. Pro pozdější možnou kontrolu zahozených tisků použijeme logovací funkci, která do souboru zaznamená informaci o právě zahozeném tisku, jeho identifikačním číslu a aktuálním datu. Sekci *Dokument* nalezneme pomocí funkce z ukázky 6.12.

```
foreach($dom->find('h2 class="section-title"') as $e) {
        if ($e->plaintext == "Dokument") {
            // ulozime do souboru informaci o zahozenem tisku
             $this->log_zahod($tisk, "Dokument");
         }
}
```

Ukázka 6.12: Nalezení sekce Dokument.

Nyní tedy víme, že se jedná o typ dokumentu, který chceme zpracovávat, potřebujeme ale dále zjistit, jaký je jeho stav projednávání v Parlamentu, a podle toho zjistit příslušné informace a uložit do příslušné tabulky. Pokud je tisk stále v jednání Parlamentu, v grafu znázorňujícím jeho průchod Parlamentem až k vyhlášení ve sbírce zákonů je červeně označen aktuální stav, v němž se zákon právě nachází. V HTML kódu jsou u některého ze stavů uvedeny třídy „mark current“ nebo „mark person current“. Pokud byl tisk zamítnut, v grafu je to naznačeno černým stavem a v kódu nalezneme třídy „mark terminated“ nebo „mark person terminated“. Pro ilustraci vyhledání některých těchto tříd pomocí DOM parseru je zde uvedena ukázka 6.13.

```
$dom->find('//span[@class="mark terminated"]')
$dom->find('//span[@class="mark current"]')
```

Ukázka 6.13: Nalezení zamítnutých nebo nedokončených návrhů pomocí DOM parseru.

Jakmile zjistíme, že se jedná o takto dosud úspěšně neukončený zákon, musíme zjistit datum, kdy byl návrh zákona předložen Poslanecké sněmovně, kdo jej podal a jeho aktuální stav (zamítnuto nebo v projednávání). Bude nás tedy zajímat sekce *Předkladatel*, a proto získáme pouze tu z obsahu stránky, a to pomocí regulárního výrazu z ukázky 6.14.

```
/<h2 class="section-title">Předkladatel<\/h2><div class="section-content simple">.*?
<\/div>/
```

Ukázka 6.14: Výraz pro získání sekce Předkladatel.

Nyní tuto získanou část dále analyzujeme a zjistíme podavatele a to, jestli patří do koalice. Jelikož se ale obsah dokumentu liší podle typu tisku, musíme zjistit jeho typ. Pokud se v sekci *Předkladatel* nachází na příslušném místě informace o tom, že zastupitelstvo některého kraje podalo tento návrh zákona, tuto informaci si zapamatujeme a hned také víme, že o koalici nejde. Podobné je to v případě, že se jedná o mezinárodní smlouvu, návrh podaný vládou nebo návrh státního rozpočtu. V ten okamžik víme, že podavatelem je vláda a počítáme, že patří do koalice. Složitější je, pokud zjistíme, že podavatelem jsou senátoři nebo poslanci. V tom případě musíme zjistit, jestli se jedná jen o jednotlivce nebo o skupiny poslanců nebo senátorů. Při určování koalice se v případě podavatele jednotlivce podíváme do databáze a zjistíme, jestli daný člověk náleží do koalice nebo opozice, a tuto informaci později, až budeme mít všechna data získána, uložíme do databáze k příslušnému návrhu zákona. Pokud je předkladatelem skupina zákonodárců, musíme zjistit příslušnost do koalice a opozice všech. Pokud jsou všichni, kdo se podíleli na návrhu, v koalici či opozici, bude i výsledek koalice či opozice. Pokud se jedná o různé příslušníky, návrh zákona byl podán smíšenou skupinou. Nakonec z této sekce zjistíme ještě datum podání návrhu a všechny zjištěné informace můžeme pomocí příslušné metody třídy `DatabaseModel` uložit do databáze do tabulky `tiskyNeuspech`. Také použijeme logovací funkci, která všechny tyto informace uloží tak, že zaznamená každý tisk na nový řádek souboru.

Nyní víme, že se jedná o návrh zákona, který již byl úspěšně ukončen a byl vydán ve sbírce tisků. Musíme ale zase rozlišovat typy tisků, které zpracováváme, neboť struktura uložených informací je rozlišná. Obecně ale budeme u všech tisků zjišťovat informaci, kdo návrh zákona podal, jeho příslušnost do koalice, kdy byl návrh podán a kdy se dostal k podpisu prezidenta – tedy počet dní, který návrh zákona strávil v Parlamentu.

Tak jako v případě neúspěšných tisků, nejprve získáme sekci *Předkladatel*, se kterou budeme pracovat. Pokud zjistíme, že jde o mezinárodní smlouvu, víme, že podavatel je vláda a patří do koalice. Zjistíme dále datum podání pomocí regulárního výrazu zachyceného v ukázce 6.15.

```
/sněmovně dne (\d{1,2}\.\s*?\d{1,2}\.\s*?\d{4})\./
```

Ukázka 6.15: Výraz pro zjištění data podání mezinárodní smlouvy.

Dále nás zajímá sekce *Prezident*, kterou najdeme pomocí výrazu z ukázky 6.16.

```
/<div class="section"><h2 class="section-title">Prezident republiky<\/h2>.*?<\/div>/
```

Ukázka 6.16: Výraz pro nalezení sekce Prezident.

Nyní zbývá podobným způsobem jako v případě nalezení data podání zjistit datum doručení k podpisu prezidenta a získáme kompletní požadované informace.

Pokud se jedná o návrh státního rozpočtu, je situace podobná, jen s několika malými úpravami ve vyhledávání ve zdrojovém kódu.

Pokud se jedná o senátní návrh zákona, je jméno navrhovatele uloženo jiným způsobem, a proto ho extrahujeme speciální funkcí, která vrátí jméno zástupce navrhovatele. Toto jméno musíme vyhledat v tabulce senátorů a zjistit jeho příslušnost do koalice. Pokud je podavatelem vláda, data získáme podobným způsobem, až na několik odlišností. Pokud zákon předložila skupina poslanců, musíme zjistit všechna jejich jména, příslušnosti do koalice a poté vyhodnotit, jestli se jedná o skupinu smíšenou nebo celou patřící do koalice či opozice. V případě podání návrhu pouze jedním poslancem je situace obdobná, jen jednodušší, protože nemusíme vyhledávat více lidí.

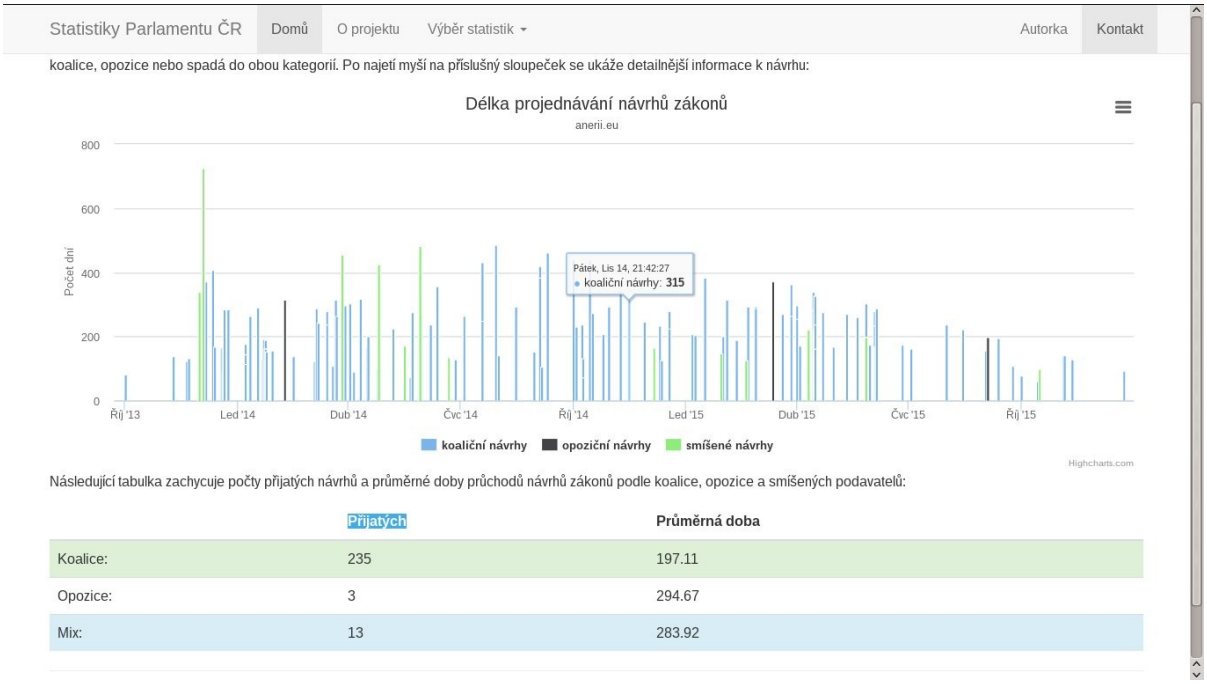
Jestliže byly všechny potřebné informace správně získány a jsou neprázdné, vypočítáme časový interval mezi datem podání návrhu a datem, kdy se návrh dostal k podpisu prezidenta. Všechny tyto zjištěné údaje uložíme do databáze do tabulky `tisk`. Zároveň používáme logovací funkci, abychom správnou činnost a analýzu textů mohli kdykoliv zkontrolovat. Nakonec aktuálně zpracovávané číslo tisku smažeme z tabulky `nezpracovano` a takto se zpracováváním tisků pokračujeme, dokud tabulka není prázdná.

Presenter

Statistiky vytvořené z dat získaných převážně z tohoto modulu budou prezentovány na třech stránkách webové prezentace práce, budeme proto potřebovat tři presentery, ve kterých zavoláme funkce, které naplní připravené proměnné daty z databáze tak, aby byly použitelné ve vrstvě *view*. Metody, které plní proměnné daty z databáze, se nacházejí ve vrstvě *model*, pro lepší přehlednost jsou však zmíněny ve vrstvě *presenter*. Tyto metody často vracejí výsledky ve formě vícedimenzionálních polí. Jejich úkolem je vybrat požadovaná data z databáze, případně je nějak upravit nebo z nich například vypočítat průměr a vrátit výsledek vhodný pro zobrazení. Předpřipravují například pole všech úspěšných návrhů zákonů s jejich dobou průběhu, datem podání a dalšími informacemi a také ostatní data potřebná pro tabulky a grafy.

View

V této vrstvě zobrazujeme data. První statistika týkající se doby průchodu návrhů zákonů Parlamentem v závislosti na příslušnosti podavatele do koalice či opozice používá zobrazení dat pomocí grafu a posléze také tabulky. Graf se vykresluje pomocí javascriptového kódu, ve kterém je nutné uvést a případně si vytvořit vlastní jazykový set nastavující například popisky os a další věci týkající se grafu. Dále také uvedeme, že se bude jednat o sloupcový typ grafu a nastavíme formát os. Poté data ve formátu JSON předáme na vstup tohoto grafu. Další grafy pro statistiky týkající se počtu návrhů zákonů pracují na stejném principu, jen je nutné modifikovat typ a nastavení zobrazování grafů. Na obrázku 6.1 a 6.2 jsou demonstrovány některé výsledné statistiky.



Obrázek 6.1: Ukázka výsledného zobrazení délky projednávání návrhů zákonů.

Obrázek 6.2: Ukázka výsledné statistiky přijatých, zamítnutých a probíhajících návrhů zákonů.

6.8 Modul zjištění klíčů

Dále budeme potřebovat tabulku všech existujících tisků a jejich klíčových slov, které je třeba extrahovat ze stránky tisků. Pro tyto účely slouží tabulka tisky_diskutovanost. Bude obsahovat identifikační číslo tisku a až tři klíčová slova návrhu a počet slov pronesených poslanci k danému tisku. Tento poslední sloupec však bude naplněn daty až při běhu následujícího modulu.

Model

Princip činnosti tohoto modulu je podobný jako princip činnosti modulu, který zjišťuje nové tisky. V tomto případě se však data ukládají do tabulky tisky_diskutovanost a po ověření z obsahu stránky, že tisk existuje, se z něj extrahují uvedená klíčová slova, která se uloží do databáze pod příslušným identifikačním číslem tisku. Pro vyhledání klíčových slov použijeme regulární výraz z ukázky 6.17.

```
/EUROVOCu:.*?<p>/
```

Ukázka 6.17: Výraz pro zjištění klíčových slov tisku.

Na výsledek tohoto regulárního výrazu použijeme DOM parser, kterým vyhledáme všechny odkazy. Textem uvnitř uzlu těchto odkazů jsou požadovaná klíčová slova. Část této funkce je vidět v ukázce 6.18.

```
$dom = str_get_html($matches[0]);
foreach($dom->find('a') as $e) {
            $text = $e->plaintext ;
 // ulozime klicova slova do db
 // pokracovani kodu ...
}
```

Ukázka 6.18: Část funkce pro vyhledání klíčových slov z odkazů.

6.9 Modul zjištění diskutovanosti a práce s textem

Tento modul zpracovává projevy poslanců týkající se sněmovních tisků. Zjišťuje celkový počet pronesených slov k jednotlivým tiskům a také počty slov, které poslanci vyřkli při komentování návrhů zákonů. Dále je s těmito projevy pracováno za účelem aplikování techniky vyhledávání informací v textu, s použitím váhovací metody IDF, a také je počítána profesní slovní zásoba poslanců. Data získaná tímto modulem jsou základem pro několik statistik, jejichž tvorba a zobrazení zde budou taktéž popsány.

Model

Jelikož množství dat, se kterým budeme pracovat, je velké a také již zveřejněné stenografické záznamy se s časem nemění, pouze přibývají další, všechna stažená data, tedy celé HTML stránky, se budou při prvním použití ukládat do cache na straně serveru. Pokud by znovu vyvstala potřeba s nimi pracovat, budou primárně získávány z cache a až v případě nedostupnosti budou znovu načtena z internetu.

Budeme využívat stenografické záznamy ze schůzí řazených podle jmen poslanců. Je třeba implementovat cyklus, který proběhne pro každého poslance v databázi. Takto postupně projdeme všechny záznamy. Nejprve zjistíme z databáze identifikační číslo a jméno právě zpracovávaného poslance. Pomocí identifikačního čísla sestavíme odkaz vedoucí na stránku, kde je seznam všech schůzí, během kterých tento poslanec ve sněmovně vystoupil. Každá schůze je navíc rozčleněna na jednotlivá vystoupení týkající se některého tisku nebo vystoupení za jiným účelem a pod nimi nalezneme odkazy do příslušných míst, přesněji stránku stenozáznamu, kde se toto vystoupení nachází. Jakmile tedy sestavíme odkaz a získáme stránku, odstraníme vše kromě názvů jednotlivých

vystoupení na schůzích a odkazů do stenozáznamů. Funkce realizující tyto operace jsou ke shlédnutí v ukázce 6.19.

```
preg_match_all('/<dd>.*?<\/d(d|iv)>/s', $html,$m);
//vytahneme jednotlive nadpisy tematu a odkazy na prepisy, ulozime do pole
foreach ($m[0] as $retezec) {
        $retezec = preg_replace('/<\/dd>/',"<br>", $retezec,-1);
        if ((preg_match_all('/<b>.+?(<br>\s*<br>|<\/div>)/s', $retezec,
            $matches3)) >0){
            // vysledky spojime
             $pole = array_merge($pole,$matches3[0]);
        }
}
```

Ukázka 6.19: Získání odkazů na stenozáznamy.

Dále budeme pracovat pouze se záznamy týkajícími se sněmovních tisků. Každý záznam se buď netýká žádného tisku nebo se týká pouze jednoho nebo dvou tisků. Pokud se žádného tisku netýká, nebudeme s ním pracovat; pokud jsou zde tisky uvedeny, uložíme si jejich konkrétní čísla pro pozdější použití. V daném záznamu tedy hledáme, je-li v textu uvedeno: „/sněmovní tisk <číslo_tisku>/“ nebo „/sněmovní tisky <číslo_tisku> a <číslo_tisku>/“, kde <číslo_tisku> zastupuje odkaz s číslem tisku. Takto získané tisky budeme postupně zpracovávat. Extrahujeme si jednotlivé odkazy do stenozáznamů týkající se daného tisku. Jednotlivé odkazy budeme postupně procházet a získávat z nich data.

Přistoupíme tedy na první odkaz a dostaneme se na odkazované místo ve stenozáznamech. Nejprve musíme určit, co se nachází na stránce, na kterou jsme se dostali. Vyhledáme, jestli se zde nachází nějaký nový nadpis, tedy začátek jednání o jiném tisku nebo jiném tématu. Pokud takovýto nadpis nalezneme, analyzujeme ho a zjistíme, jestli je to začátek právě zpracovávaného tisku. Pokud ano, vše od něj do začátku stránky odstraníme, aby se projevy poslanců k předchozímu tisku nemísily s aktuálně zpracovávaným. Pokud nalezneme začátek jiného tématu, znamená to, že jsme se dostali až na konec současně zpracovávaného, a tak smažeme vše od tohoto nadpisu do konce stránky a nastavíme příznak říkající, že tento záznam tisku, nebo-li toto téma, je již zpracované.

Dále se zaměříme na konkrétní data na stránce. Potřebujeme vyhledat úseky řečí poslance, kterého aktuálně zpracováváme. Struktura stránky je taková, že jsou zde úseky řečí uvozeny jménem poslance, který je pronesl. Tyto úseky mohou být vzhledem ke své pozici na stránce trojího typu. Buď mohou být mezi řečmi jiných poslanců nebo mohou končit na aktuální stránce a pak také, po zalistování, na další stránce pokračovat, již bez uvedení jména poslance. Všechny tyto situace je třeba detekovat zvlášť, jelikož mají jinou strukturu. Tyto tři situace detekujeme pomocí regulárních výrazů a funkcí, které jsou představeny v ukázce 6.20.

```
1. preg_match_all("/\(pokračuje $prijmeni\)(.*?)<a href=\"\/sqw\/detail\.sqw\?id/s",
    $stranka_odkaz, $rec)
2. preg_match_all("/$row->jmeno_poslance:(.*?)<a href=\"\/sqw\/detail\.sqw\?id/",
    $stranka_odkaz, $reci)
3. preg_match_all("/$row->jmeno_poslance:(.*?)\*\*\*/",$stranka_odkaz, $rec)
```

Ukázka 6.20: Funkce detekující tři typy úseků záznamů ve stenozáznamech.

Proměnná $row->jmeno_poslance zastupuje celé jméno poslance a $prijmeni pouze jeho příjmení. Jelikož takovýchto úryvků řečí může být na stránce více, nalezený text po zpracování smažeme a vyhledáváme další, dokud je něco nalezeno.

Jakmile získáme text řeči poslance, potřebujeme spočítat počet slov. Nejdříve byla použita funkce str_word_count, ale při testování bylo zjištěno, že nepodporuje české znaky, a tak slova nesprávně rozděluje. Počet vypočítaných slov tedy neodpovídá realitě. Proto, po odstranění nadbytečných mezer a znaku nového řádku, bylo nutno upravený text nejdříve rozdělit pomocí funkce explode na základě mezer a poté slova spočítat. Takto jsme vypočítali počet slov úryvku a v průběhu nalézání dalších úryvků k danému tisku od zpracovávaného poslance k tomuto číslu připočítáváme další pronesená slova, až nakonec zjistíme, kolik slov celkem tento poslanec pronesl k tomuto tisku. Toto číslo připočítáme k hodnotě pocet_slov uložené v databázi k danému poslanci a také k hodnotě diskutovanosti zpracovávaného čísla tisku nebo obou tisků. Takto po zpracování všech tisků a všech poslanců zjistíme celkové počty pronesených slov jednotlivými poslanci k sněmovním tiskům a také diskutovanost jednotlivých tisků. V průběhu zpracovávání jednotlivých tisků ukládáme jejich odkazy do pole jako klíč, abychom nezpracovávali některé odkazy vícekrát. Pokud je uvedený odkaz, který jsme díky průchodu až na konec projednávání tématu už zpracovali, přeskočíme ho.

Jakmile jsou zpracovány všechny úryvky ze stránky, zkontrolujeme, jestli je nastaven příznak, že byl nalezen začátek nového tématu. Pokud ano, zpracovávání tisku končí; pokud ne, nalezneme odkaz na další stránku, kde pokračuje zápis jednání aktuálního tisku, a pokračujeme stejným způsobem zpracovávání, dokud nedojdeme na konec stenozáznamu nebo k novému tématu.

Jednotlivé úryvky budeme během zpracování předávat metodě, která bude pracovat se slovy, převádět je na lemmata a nakonec vypočte jejich výsledné váhy. Z textu tedy nejprve odstraníme číslice a interpunkční znaky jako otazník či vykřičník. Všechna slova porovnáme s vytvořeným stoplistem - tedy seznamem slov, která budou ignorována – a slova, která jsou v něm obsažena, z textu vymažeme. Dále potřebujeme jednotlivá slova převést na základní tvar tzv. lemma. K tomu použijeme externí aplikaci Majka[24], která bude spouštěna pomocí příkazu exec. To, že tento příkaz na hostingu obvykle není dostupný, je další důvod, proč tento modul není na hostingu spouštěn a neprobíhá zde automatická aktualizace. Majka může mít na vstupu pouze jedno slovo, které je jí

24 https://nlp.fi.muni.cz/ma/

předáno spolu s dalšími parametry. Výstupem získáme řetězec několika tvarů slova spolu s dalším upřesněním, o jaký slovní druh se jedná, a dalšími informacemi [17]. Z tohoto výstupu extrahujeme pouze první slovo, kterým je právě hledané lemma. Jakmile toto slovo získáme, předáme jej databázové funkci, která v tabulce termy zjistí, jestli je zde tento term již uložen. Pokud ano, zvýšíme počet výskytů v dokumentu o hodnotu 1. Pokud není, term uložíme a počet výskytů nastavíme na hodnotu 1. Jako jeden dokument uvažujeme jedno vystoupení, tedy jednu řeč poslance k danému tisku. Zároveň počítáme celkový počet dokumentů pro další použití. Také kontrolujeme a vytváříme seznam slov, která se v tomto dokumentu vyskytla, a pokud se objeví znovu, už je nezpracováváme. Termy, které se v dokumentu vyskytují, se zároveň ukládají do pomocné tabulky tmp_termy, která je po zpracování každého poslance vyprázdněna. Předtím však zjistíme počet různých termů, tedy slov, které poslanec použil, a tedy jeho profesní slovní zásobu, kterou uložíme do databáze. Po zpracování všech poslanců jsou v tabulce termy všechna slova použitá všemi poslanci a počet jejich výskytů ve všech dokumentech. Vypočítáme každému termu váhu IDF podle vzorce uvedeného v podkapitole 2.6.1 a tuto váhu uložíme ke každému termu do databáze.

Z důvodu urychlení práce s databází je více databázových dotazů sdruženo do jedné transakce.

Presenter

Presentery získávají data dle různých kritérií z databáze a připravují je pro konečné zobrazení. Jejich prostřednictvím jsou vybíráni například poslanci s nejvyšším počtem slov, tisky k zobrazení podle časových i obsahových kritérií, které jsou kompletovány s daty z jiných tabulek. Takto připravená data jsou zobrazována pomocí šablon.

View

První zde zmíněná statistika se zaměřuje na řečnictví v Poslanecké sněmovně. Výběr je tvořen JQuery knihovnou Select2, tedy boxem, ze kterého můžeme podle nápovědy vybírat jména poslanců, která chceme dynamicky zobrazit v tabulce, spolu s počtem použitých slov a počtem vystoupení poslance. Aby se tabulka mohla měnit podle požadavků uživatele, je na stránce použita knihovna JQuery, která změny v select boxu zachytí, zašle ajaxový požadavek *presenteru* pro získání dat a následně zajistí jejich překreslení zpět do tabulky. Hlavní část kódu, který se nachází v šabloně, lze vidět v ukázce 6.21.

```
<script n:syntax="double">
$('select[name="myselect"]').change(function() { volej($(this).val()); });
function volej(param){
 var res = param;
 $.get( {{link zmen!}} , { "res[]": res } ,function(data) {
            $('#myTable').find("tr:gt(0)").remove();
                        $.each( data, function( key, value ) {
                           $('#myTable > tbody:last-child').append("<tr class=
                                  success ><td> " + key +
                           " </td><td> " + value[0] + "</td><td>" + value[1]

                             +"</td><td>"+value[2]+"</td></tr>");
                         });
             });
           if (res === null){
                  $ ('#myTable').find("tr:gt(0)").remove();
            };
  };
//...
</script >
```

Ukázka 6.21: Část kódu pro zachycení požadavků uživatele a překreslení výsledků do šablony.

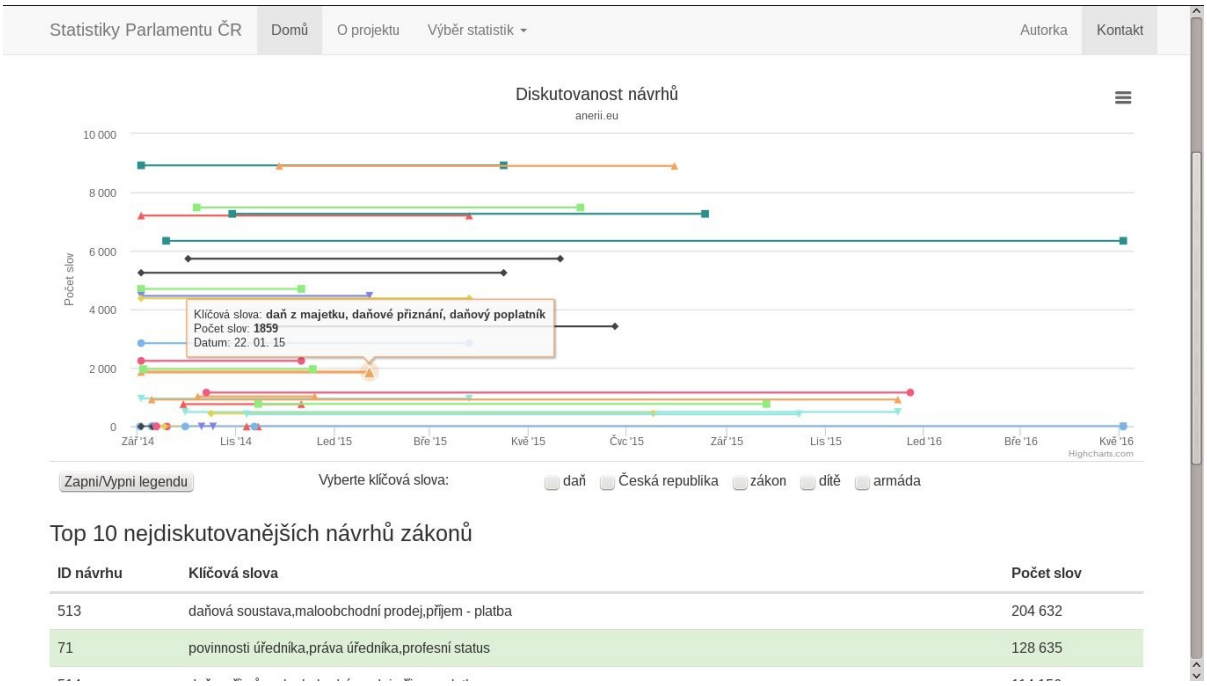
Dále jsou zde statistiky zobrazující tabulky poslanců s největším počtem slov, podle jednotlivých poslaneckých klubů i v celé Poslanecké sněmovně.

Další statistika se zaměřuje na diskutovanost návrhů zákonů. Hlavním cílem je zobrazit závislost klíčových slov na počtu pronesených slov k danému návrhu. Jelikož je však návrhů velké množství, bylo třeba jejich zobrazování rozdělit na několik časových období a také na více a méně diskutované návrhy. Tyto kombinace si uživatel sám vybere a pomocí JQuery jsou tyto požadavky předány ajaxovým požadavkem *presenteru* a následně vyhledány v databázi. Po získání dat JQuery zajistí překreslení grafu. Jádro této funkce je zachyceno v ukázce 6.22.

```
<script n:syntax="double">
var $button = $('div[id="button"]');
var $clone = $button.clone();
$button.button();
if(jQuery.isEmptyObject(values)){
        $('input[name=radioOptions]').removeAttr("disabled");
        $('select[id="selecttime"]').removeAttr("disabled");
        $("#mene").prop("checked", true)
        var value = $( 'input[name=radioOptions]:checked' ).val();
        var value2 = $('select[id="selecttime"]').val();
        $.get( {{link obdobi!}} , { hodnota : value2, checked: value } ,function(data) {
                $button.replaceWith($clone);
                 f(data);
        });
    }else{
        $('input[name=radioOptions]').attr("disabled", true);
        $('select[id="selecttime"]').attr("disabled", true);
        $.get( {{link klicovaSlova!}} , { klslova : values} ,function(data) {
                    $button.replaceWith($clone);
                     f(data);
        });
    }
//...
</script>
```

Ukázka 6.22: Část kódu zajišťující překreslení grafu podle požadavků uživatele.

Poslední statistiky zachycují výsledky metody váhování IDF. V tabulce je zobrazeno 20 termů s nejvyšší váhou IDF a také pro zajímavost 5 termů s nejnižší hodnotou IDF. Je zde také prezentován výsledek váženého průměru slovních zásob poslanců a také nejčastější slova, která zazněla v Poslanecké sněmovně v rámci diskusí ohledně sněmovních tisků. Také zde můžeme vybrat konkrétní poslance pro zjištění jejich profesní slovní zásoby a tyto výsledky zobrazit. Na obrázku 6.3 můžeme vidět ukázku výsledné statistiky diskutovanosti návrhů zákonů.



Obrázek 6.3: Ukázka výsledné statistiky diskutovanosti návrhů zákonů.

6.10 Testování

Testování bylo prováděno jednak během celého vývoje jednotlivých modulů a také po celkovém dokončení všech částí. Při testování modulů zpracovávajícího tisky a diskutovanost návrhů bylo potřeba logovacích funkcí, které zaznamenávaly činnost a získané výsledky modulu. Po shlédnutí tisků, které nebyly správně zpracovány, bylo jasné, že ne všechny tisky se podaří správně zpracovat, jelikož se na stránkách vyskytují různé chyby, díky kterým analýza neproběhne korektně. S těmito tisky se musí počítat, jejich počet je však malý. Další velké množství chyb se nachází ve zveřejněných záznamech z jednání. I tyto chyby ztěžují analýzu textu a i zde je možné chybné zpracování takovýchto částí. Navíc vzhledem k opravdu velkému množství zpracovávaných dat není možné „ručně“ ověřit správnost výsledků. Po rozsáhlém testování mi však nejsou známy další problémy a tyto vzniklé chyby jsou nevýznamné v porovnání s množstvím zbylých dat. Časová náročnost běhu všech modulů, kromě modulu diskutovanosti se pohybuje v řádu desítek sekund, maximálně několik minut. Situace modulu diskutovanosti je však jiná. Zde byla doba běhu asi 25 hodin na mém osobním počítači s následujícími parametry: procesor Intel Core i7-2570QM 2,2 GHz, OS linux Mageia 5 (32-bit), paměť 8 GB DDR3, disk WD Black2 (systém včetně webového serveru a databáze na SSD části disku).

7 Závěr

Cílem této bakalářské práce bylo vytvořit soubor statistik, které budou jednoduchou a přehlednou formou prezentovat některé zvláštní a netradiční poznatky získané z oblasti legislativní činnosti Parlamentu ČR. K tomuto účelu slouží několik modulů, které získávají a analyzují data, která jsou zobrazena především pomocí tabulek a grafů umístěných na veřejných webových stránkách tak, aby byly dostupné pro širokou veřejnost.

Nejprve bylo třeba nastudovat fungování Parlamentu ČR, abych mohla navrhnout co nejzajímavější statistiky, které budou implementovány. Dalším krokem bylo zjistit, jak jsou data uložená na stránkách Poslanecké sněmovny a Senátu a jestli je reálné požadované informace z nich získat. Nutná byla také hlubší znalost regulárních výrazů, jejichž konstrukce byla důležitou součástí práce. Aplikaci bylo nutné pro přehlednost implementovat ve formě několika modulů, z nichž každý se orientuje na získávání jiných dat a tedy tvorbu rozličných statistik. Mezi implementované statistiky patří délka projednávání návrhů zákonů v závislosti na příslušnosti podavatele do koalice za toto volební období i za volební období minulé. Nechybí ani grafy ukazující v číslech i procentech, kolik návrhů zákonů podali poslanci, senátoři, vláda a zastupitelstva krajů, kolik jejich návrhů bylo přijato, zamítnuto nebo je stále aktuálních. Druhá část statistik se zabývá diskutovaností návrhů zákonů v závislosti na jejich klíčových slovech během různých časových období. Také jsou prezentovány údaje o počtech slov, které poslanci pronesli během projednávání návrhů zákonů, a také celkové počty jejich vystoupení ve sněmovně. Tyto vytyčené cíle byly úspěšně dosaženy.

Po dohodě s vedoucím práce a díky informacím poskytnutým od pana Ing. Vladimíra Bartíka, Ph.D. a doc. RNDr. Pavla Smrže, Ph.D. byla nad rámec zadání implementována technika vyhledávání informací v textu pomocí váhovací metody IDF. Součástí úprav zpracovávaných textů stenozáznamů, které bylo třeba provést, bylo použití externí aplikace z Masarykovy univerzity – Majka, s jejíž pomocí byly získány základní tvary jednotlivých slov a následně jim byly přiřazeny váhy. Dále byla pro zajímavost zjištěna profesní zásoba poslanců a několik nejčastějších slov, která poslanci ve vztahu k návrhům zákonů pronesli.

Všechny vytvořené statistiky, tedy výsledky této bakalářské práce, jsou umístěny na osobní webové stránce dostupné na adrese *www.anerii.eu* tak, aby byly přístupné všem, kteří se jen trochu zajímají o dění v Parlamentu. Statistiky v této oblasti nejsou příliš časté, lze najít pouze základní a velmi obecné, spíše nic neříkající, přehledy. Proto bude snaha tento projekt dále udržovat i po odevzdání práce, neboť v něm spatřuji příležitost přiblížit okolnímu světu legislativní činnost a dění v Parlamentu, tedy instituci, která silně ovlivňuje životy nás všech.

Mým přínosem v práci bylo zjištění chyby v DOM parseru, který nekorektně zpracovává HTML dokument s diakritikou, který v hlavičce neobsahuje nastavení použité znakové sady. Při konzultacích práce, vedoucí vzhledem k dosaženým výsledkům navrhl možnost publikace práce.

7.1 Další vývoj projektu a možná rozšíření

Další vývoj projektu spatřuji především v dalším rozšiřování repertoáru statistik. Další statistiky by se například mohly týkat hlasování poslanců a senátorů. Případně by bylo zajímavé analyzovat legislativní činnost EU nebo porovnat legislativní činnosti České republiky s nějakým jiným státem například SR. Další zajímavou možností by bylo na získaná data aplikovat techniku „data mining“ pro zjištění zajímavých souvislostí v získaných datech.

Pokud se mi v budoucnu naskytne příležitost pokračovat v tomto tématu například v diplomové práci, ráda bych tuto možnost zvážila.

7.2 Přínos práce pro autora

Práce pro mě byla velkým přínosem, protože jsem musela proniknout do pro mě nových koutů informatiky i politického dění u nás. Vyzkoušela jsem si práci v Nette frameworku, což by se mohlo hodit pro příští projekty, jelikož velmi usnadňuje a zpřehledňuje vytvořený kód. Dále bylo zajímavou zkušeností studium fungování regulárních výrazů, vzdálené načítání stránek a extrahování informací z HTML a také práce s javascriptovými grafy, jejichž vzhled bylo třeba upravovat. Také jsem si vyzkoušela práci s databází, ajaxem a knihovnou JQuery. V neposlední řadě jsem získala několik cenných zkušeností s váhovací techniku IDF pro získávání informací z textu, dále také s prací s textem a jeho převedení na základní tvar.

Literatura

[1] *Poslanecká sněmovna Parlamentu ČR* [online]. [cit. 2016-01-25]. Dostupné z: http://www.psp.cz/sqw/hp.sqw

[2] *Senát Parlamentu ČR* [online]. [cit. 2016-01-25]. Dostupné z: http://www.senat.cz

[3] *Legislativní helpdesk* [online]. Úřad vlády České republiky [cit. 2016-01-26]. Dostupné z: https://help.odok.cz/vykladovy-slovnik/-/wiki/Výkladový slovník/Sněmovní tisk

[4] Sněmovní tisky. *Poslanecká sněmovna Parlamentu ČR* [online]. Úřad vlády České republiky [cit. 2016-01-26]. Dostupné z: http://www.psp.cz/sqw/hp.sqw?k=1303

[5] Přijímání zákonů. *Poslanecká sněmovna Parlamentu ČR* [online]. Úřad vlády České republiky [cit. 2016-04-18]. Dostupné z: http://www.psp.cz/sqw/hp.sqw?k=173

[6] Koaliční smlouva mezi ČSSD, hnutím ANO 2011 a KDU-ČSL na volební období 2013-2017. *Vlada.cz* [online]. [cit. 2016-04-18]. Dostupné z: http://www.vlada.cz/assets/media-centrum/dulezite-dokumenty/koalicni_smlouva.pdf

[7] LIU, Bing. *Web data mining: exploring hyperlinks, contents, and usage data*. New York: Springer, 2007. ISBN 35-403-7881-2.

[8] HAN, Jiawei, KAMBER, Micheline a PEI, Jian. *Data mining: concepts and techniques*. 3Rd ed. Boston: Elsevier, 2012. Morgan Kaufmann series in data management systems. ISBN 978-0-12-381479-1.

[9] BARTÍK, Vladimír. *Dolování z textu a na webu. Brno. Vysoké učení technické v Brně, Fakulta informačních technologií, Ústav informačních systému.*

[10] BURGET, Radek. *Extrakce informace z HTML dokumentů na základě logické struktury dokumentu*. Brno, 2001. Vysoké učení technické v Brně, Fakulta informačních technologií, Ústav informačních systému.

[11] MAZAL, Zdeňek. *Extrakce textových dat z internetových stránek*. Brno, 2011. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací. Vedoucí práce byla Ing. Lucie Fojtová.

[12] Ženy ve volbách. *STATISTIKA&MY* [online]. [cit. 2016-04-19]. Dostupné z: http://www.statistikaamy.cz/2013/12/zeny-ve-volbach/

[13] Data Poslanecké sněmovny a Senátu. *Poslanecká sněmovna Parlamentu ČR* [online]. [cit. 2016-04-19]. Dostupné z: http://www.psp.cz/sqw/hp.sqw?k=1300

[14] DRÁPELA, Karel a ZACH Jan. *Statistické metody I (pro obory lesního, dřevařského a krajinného inženýrství)*. 1. vyd. Brno: Mendelova zemědělská a lesnická univerzita, 1999. ISBN 80 7157-416-3.

[15] Základy statistiky. *Matematika.cz* [online]. [cit. 2016-04-19]. Dostupné z: http://www.matematika.cz/zaklady-statistiky

[16] Požadavky Nette Framework. *Nette* [online]. [cit. 2016-04-19]. Dostupné z: https://doc.nette.org/cs/2.3/requirements

[17] *Free natural language morphology* [online]. [cit. 2016-04-19]. Dostupné z: https://nlp.fi.muni.cz/ma/

Přílohy

A) Přiložené DVD

K této práci je přiloženo nepřepisovatelné DVD. Jeho obsahem je zejména složka nette-bc, která obsahuje zdrojový kód programu. Vnitřní struktura této složky byla popsána v podkapitole 6.1. Obsahuje také soubor readme, který sdružuje některé důležité informace o programu. Dále na disku najdeme písemnou zprávu této bakalářské práce ve formátu PDF a také ve zdrojovém tvaru a v neposlední řadě také návod k instalaci spolu se souborem databaze.sql pro vytvoření struktury databáze.

B) Instalační příručka k aplikaci

Pokud chceme spustit aplikaci, nejprve je vhodné si vytvořit databázi a v ní tabulky a jejich sloupce, podle návrhu databáze, který byl uveden v podkapitole 5.4. K tomuto účelu lze využít přiložený skript *databaze.sql*. Jakmile máme tuto databázi připravenou a zdrojový kód umístěný v adresáři webového serveru (dle konfigurace document_root webového serveru[25]), musíme se ujistit, že náš webový server disponuje všemi potřebnými funkcemi pro běh aplikace v Nette. Konkrétní informace nalezneme ve zdroji [16]. Nyní můžeme přistoupit ke spouštění jednotlivých modulů, čímž se budou plnit tabulky v databázi, a výsledky můžeme kontrolovat na webové stránce projektu. Spuštění modulu Poslanci a Senátoři provedeme zavoláním CronPresenteru například pomocí internetového prohlížeče. Pokud máme projekt například na stránkách *anerii.eu*, CronPresenter zavoláme zadáním *anerii.eu/nette-bc/www/cron* do internetového prohlížeče. Pozor, na hostingu však nelze spouštět moduly s dlouhou dobou běhu kvůli časově omezenému provádění skriptu. Další moduly kromě diskutovanosti se spouští při vstupu návštěvníka na stránky projektu, takže pro spuštění těchto modulů zavoláme HomepagePresenter (zadáním *anerii.eu/nette-bc/www/)*. V něm je aktuálně zakomentovaná funkce pro spuštění modulu zjišťující všechny tisky, neboť je potřeba pouze při první instalaci programu. Nyní ji však nesmíme zapomenout odkomentovat a poté zakomentovat. Spuštění modulu zjištění klíčových slov a diskutovanosti se provádí stejným způsobem taktéž v HomepagePresenteru. Získali jsme všechna potřebná data a výsledky jsou již ke shlédnutí na stránkách projektu.

25 http://www.karelia.com/support/sandvox/help/z/Document_Root.html