Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

BAKALÁŘSKÁ PRÁCE

PLZEŇ, 2012 **Jaromír Novotný**

P R O H L Á Š E N Í

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

Plzeň, 2012 _____

Jaromír Novotný

Poděkování

Rád bych poděkoval Ing. Lucii Skorkovské, která mi pomohla ujasnit směřování mé bakalářské práce a také za vedení a rady během jejího zpracování.

Anotace

V této bakalářské práci je popsán základ automatické sumarizace, typy používaných metod a vlastnosti těchto metod. Dále se práce zaměřuje na úspěšnost vyhodnocování sumarizačních metod a popisuje také metody, jež se používají při zpracování novinových článku. Věnuje se zde větší pozornost metodám statistickým a grafovým, které jsou zde také implementovány (jedna statická a jedna grafová metoda). Následně jsou porovnávány výsledky obou metod a jejich použitelnost pro reálné úlohy.

Klíčová slova: automatická sumarizace, používané metody, ohodnocování metod, statistické metody, grafové metody,

In this Bachelor Thesis the fundamentals of automatic summarization are described, types of used methods and their properties. Next, the thesis focused on evaluation of summarization methods and describes methods, which are used for newspaper articles. There is bigger attention to the statistic and graph methods, which are also implemented (one of the static and one of the graph methods). The results of both implemented methods and their usability for real tasks is compared.

Key words: automatics summarization, used methods, evaluation of summarization methods, statistic methods, graph methods

Obsah

Kapitola 1

Úvod

Sumarizaci (česky souhrn) můžeme chápat jako zkrácenou verzi nějakého textu (např. novinového článku), jenž je vytvořen extrakcí nebo abstrakcí originálního textu a ponechává si při tom nejdůležitější informace.

Začátek práce je věnován sumarizaci obecně. Se sumarizací se setkáváme v běžném životě, aniž bychom si to třeba uvědomovali, proto je zde uveden příklad, kde se sumarizace například vyskytuje. Důkazem toho, že sumarizace je hodně rozšířená je to, že existuje mnoho definic, zde jsou uvedeny dvě definice.

Teoretická část práce obsahuje také možnosti rozdělení sumarizačních metod z různých pohledů. Každý rozdělení obsahuje krátký popis a v nějakých případech i sumarizační metody (patřící do této části rozdělení), které se používají pro automatickou sumarizaci, neboli pro sumarizaci, která je prováděna podle těchto metod automaticky nějakým přístrojem, například počítačem (naprogramovaný algoritmus).

Tato práce se také zabývá sumarizací novinových článků. V této části jsou také uvedeny sumarizační metody, které se používají právě při tvorbě sumarizace novinových článků. U každé metody je stručný popis.

Ohodnocování výsledků je další teoretická část práce, v které se zabýváme ohodnocováním sumarizačních metod. Pro ohodnocování výsledků existuje více metod, jejich dělení a stručný popis je v této části uvedeno.

V další část je věnována sumarizačním metodám statistickým a grafickým. Tato část obsahuje popis těchto metod, příklady metod a hlavně implementaci jedné statistické a jedné grafové metody. Implementace těchto metod obsahuje detailní popis každé z vybraných metod a algoritmus pro praktickou realizaci.

Nakonec se práce zabývá výsledky. Nejdříve jsou zde vloženy originální články, které byly použity jako vstup do implementovaných metod. Dále tyto výsledky ohodnocuji, popisuji vhodnost obou metod pro články a také uvádím celkový závěr učiněný z těchto výsledků.

Cílem práce bylo seznámit se se sumarizací z teoretického hlediska a hlavním cílem práce bylo tedy zaměřit se na statistické a grafové metody, vybrat zastupitele statistické a grafové metody a nakonec provést implementaci vybraných metod. Poté aplikovat implementované metody na originální články, získat výsledky a tyto výsledky ohodnotit a popsat vhodnost metod pro reálné použití.

Kapitola 2

Teoretické základy sumarizace

2.1 Úvod do sumarizace

Příklad: Nejčastější případ, kde se můžeme setkat se sumarizací, nalezneme v novinách v podobě úvodní stránky. V tomto případě je sumarizace velmi krátká a vytvořena pro zběžný přehled článků vyskytujících se v novinách. Přestože je velmi krátká musí obsahovat nejdůležitější informace o článku, jako např. čeho se článek týká, kdy se tato událost stala a na jakém místě, atd.

Sumarizaci můžeme podle článku [19] popsat v následujících bodech:

1) Sumarizace může být vytvořená z jednoho nebo více dokumentů (článků).

2) Sumarizace musí zachovávat důležité informace originálu.

3) Sumarizace by měla být krátká.

V této práci se zabývám automatickou sumarizací, rozdíl je v tom, že sumarizaci vytváří program (tvořený určitým algoritmem), nikoliv člověk. Tudíž nám stačí vstup programu (originální text) a následně dostáváme sumarizaci.

2.2 Definice sumarizace

Sumarizaci lze podle článku [1] definovat následovně:

„ Text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer then half of the original text(s). ”

Definice uvedená v češtině:

„ Vytvořený text je zpracován z jednoho nebo více dokumentů, obsahuje nejdůležitější informace z originálního dokumentu(ů), a jeho délka nepřesahuje polovinu délky originálního dokumentu(ů). “

Sumarizaci můžeme definovat více způsoby, příkladem je další definice sumarizace textu uvedená v článku [2]:

> *„ Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks). “*

Definice uvedená v češtině:

> *„ Sumarizace je proces získávání nejdůležitější informace ze zdroje (nebo zdrojů), pro vytvoření zkrácené verze pro konkrétního uživatele (nebo více uživatelů) a pro konkrétní úlohu (nebo úlohy). “*

2.3 Rozdělení sumarizačních metod a jejich vlastnosti

Sumarizace lze dělit různými způsoby, záleží například na tom, jak chceme danou úlohu řešit nebo na vstupních datech. V různých článcích se používá rozdělení sumarizace pro daný úkol a nezahrnují se další možnosti. Lze dělit základně na dvě hlavní skupiny, ale pro lepší přehled a získání většího pohledu na sumarizaci zde uvedeme podrobné rozdělení podle článku [3]

2.3.1 Na základě rozsahu

- Jedno dokumentová sumarizace

- Více-dokumentová sumarizace

  Snižuje čas vyhledávání, zvláště když cíl uživatele je najít co nejvíce možných informací o daném tématu. Shluk podobných článků umožňuje vytvořit více-dokumentovou sumarizaci na základě podobností.

2.3.2 Podle jazyka

- Mono-jazykové

  Mono-jazykový typ textové sumarizace spočívá ve vytvoření sumarizace ve stejném jazyce, jako je zdrojový text.

- Multi-jazykové

Multi-jazykový typ textové sumarizace spočívá ve vytvoření sumarizace ze zdrojového textu, který je v jednom jazyce, do požadovaného jiného jazyka.

2.3.3Úrovně zpracování sumarizace

- Povrchní přístup

Informace reprezentovány prostřednictvím povrchních vlastností a jejich kombinacemi. Mezi povrchní vlastnosti můžeme uvést například pozičně významné termy, frekvenčně významné termy, termy specifické pro zpracovanou doménu nebo termy obsažené v uživatelově dotazu. Výsledný souhrn je extrakt.

- Hlubší přístup

Využívají sémantické zpracování k určení významných částí textu, zjišťují textové jednotky a jejich vzájemné vztahy jsou jako telurové relace, syntaktické relace apod. Také mohou využívat informaci o stavbě textu a rétorické struktuře, případně i hypertextových značek. Výsledný souhrn může být extrakt nebo abstrakt.

2.3.4 Účel, pro který je sumarizace vytvářena

- Hodnotící sumarizace

Do těchto souhrnů můžeme začlenit například kritiky, recenze, posudky. O daném dokumentu vyjadřují mínění autora, což je jejich charakteristickým rysem. Tato jejich vlastnost vylučuje hodnotící souhrny zařadit do skupiny automaticky generovatelných.

- Individuální sumarizace

Poskytují zkrácenou formu informace o hlavních tématech dokumentu (zachovávající jeho nejpodstatnější části). Uživatel by podle nich měl být schopen rozhodnout o tom, zda čtení celého textu má pro něj význam. Tímto si vysloužily časté využívání ve výstupech vyhledávacích systémů, kde nahrazují originální texty dokumentů. Obvykle obsahují do 10% z původního textu.

- Informativní sumarizace

Tento souhrn obsahuje většinu základních informací z originálního dokumentu. Tím získává uživatel dostatečný přehled o daném tématu a není nutné čtení původního dokumentu.

2.3.5 Sumarizace z pohledu uživatelů

- Sumarizace zaměřená na uživatele („user-focused“)

Tento souhrn je zaměřen na určité požadavky konkrétního uživatele nebo skupiny uživatelů.

- Sumarizace zaměřená na téma („topic-focused“)

- Sumarizace zaměřená na získávání odpovědí („query-focused“)

2.3.6 Podle použitého principu

- Heuristické metody

Tyto metody můžeme považovat za první pokusy o automatickou sumarizaci, začaly se vyskytovat již v polovině minulého století. Metody jsou extraktivního typu. Mezi ně můžeme zařadit například metody:

- Luhnova metoda (Luhn method), tato metoda pracuje na předpokladu, že výskyt stejných slov v dokumentu (nezahrnují se zde stop slova) představuje hlavní téma.
- Edmunson, tato metoda rozšiřuje dřívější práce na výskyt slov. Předpokládá se, že důležité informace se vyskytují například v názvu, úvodu, závěru, na začátku nebo na konci vět, jsou indikovány zdůrazňujícími slovy a slovy klíčovými. V metodě se používá trénovacích dat pro nastavení parametrů.

- Statistické metody

V začátcích sumarizace byly metody hlavně statistické ve své podstatě a zaměřovali se na frekvence výskytů nejdůležitějších konceptů v textu. Hlavní problém čistě statistických metod je, že nezohledňují kontext. Zohlednění kontextu se z velké části opírá na identifikaci a zachycení nejen duplikovaných termů, ale také na souvisejících termech. Tento koncept (známý jakou soudržnost), spojuje sémanticky související termy, jež jsou důležitou součástí uceleného textu. Nejjednodušší formou této soudržnosti je slovní soudržnost.

Dále se budeme statistickým metodám věnovat v další části práce.

- Grafové metody

Graf je podle článku [13] nejefektivnější reprezentace vztahů mezi dvěma proměnnými. Nejčastěji se umisťuje na osu x nezávislá proměnná (může být změněna) a na osu y závislá proměnná (reaguje na změny v nezávislé proměnné).

Grafové metody jsou relativně rozšířené a více se jim budeme věnovat v další části práce.

- Algebraické metody

Zde můžeme uvést jako příklad metodu založenou na Latentní sémantické analýze (LSA - Latent Semantic Analysis). Tato analýza zachytí hlavní téma dokumentu, následně věty obsahující toto hlavní téma jsou vybírány do výsledné sumarizace. Tato metoda byla původně vytvořena pro použití na jedno-dokumentové sumarizace a později upravena a přizpůsobena na více-dokumentovou sumarizaci.

Kapitola 3

Sumarizace novinových článků

3.1 Úvod do sumarizace novinových článků

Sumarizace novinových (zpravodajských) článků vzniká jako důsledek velkého přírůstku zpravodajských článků na internetu, jelikož v takovém množství je pro čtenáře velmi obtížné efektivně shromažďovat pro něj důležité informace. Zde nastupuje sumarizace, která nahrazuje celé články kratšími verzemi se stejným přínosem informací, tudíž čtenáři nemusejí číst celé články, ale pouze nejdůležitější části z nich (sumarizace). Dále také velký počet článků umožnil rozvoj automatické sumarizaci, jelikož pro lidi je velmi těžké (podle mého názoru je v tomto ohledu nejhorší časová náročnost) vytvářet sumarizace manuálně.

3.2 Příklad metod pro sumarizaci novinových článků

V literatuře můžeme nalézt dva hlavní přístupy k automatické sumarizaci a to: lingvistický přístup, statistický přístup a popřípadě jejich kombinace. Tyto přístupy se dají v podstatě definovat také jako abstrakce a extrakce.

3.2.1 Jednoduchá metoda neautomatické sumarizace

Pro zajímavost je v této práci uvedeno, jak můžeme vytvořit sumarizaci novinového článku (pro úvodní stránku novin) bez použití automatické metody. Uvidíme, že tento postup může být pro člověka náročný v případě delších, a nebo složitějších (více důležitých informací) článků. Tato metoda je popsána podle článku [4]:

Články v novinách se zaměřují na určitou událost, jež má v nějakých případech spojitost s minulostí a určitým místem. Délka článku je dána množstvím informací o daném tématu, priority tohoto článku a umístění v novinách. Při odkazování na novinový článek musíme tedy zachytit hlavní zprávu, jež je v tomto článku obsažena. Postup na vytvoření takovéto sumarizace můžeme popsat ve čtyřech následujících krocích:

1) Najděte v článku slova „kdo“, „co“, „kdy“, „kde“ a „proč“. Toto jsou nejzákladnější fakta, která nalezneme v novinovém článku, a proto by měla být zahrnuta v sumarizaci článku. „Kdo“ odkazuje na předmět článku, „co“ odkazuje na to, co bylo řečeno o předmětu článku, „kdy“ může odkazovat stejně tak na to, kdy byl článek napsán, jako na datum popisované události, „kde“ odkazuje na všechny oblasti, které mají spojení s předmětem článku a na to co se na tomto místě stalo, „proč“ odkazuje na důvod vzniku tohoto článku. Důležité je tato fakta popsat vlastními slovy.

2) Přidáme hlavní myšlenku. Autor novinového článku napsal článek za účelem předání zprávy a právě tato zpráva je hlavní myšlenkou. Hlavní myšlenka má přímý vztah k otázce „proč“ v článku, jelikož je to její rozšíření. Pro sumarizaci hlavní myšlenky by se neměly vytvářet víc jak tři věty. V nějakých případech může mít článek více hlavních myšlenek, v tomto případě se musí zanechat popis každé myšlenky.

3) Zahrneme další detaily podporující téma. Když článek přečteme alespoň dvakrát, měli bychom pochopit hlavní informaci, což je podstatné k rozpoznání detailů hlavních od detailů přidaných pro efekt. Detaily, které musíme nejdříve přidávat, jsou takové, jež obsahují nezbytné informace pro pochopení článku, jako například pracovní pozice předmětu článku nebo kolikaletý výzkum byl prováděn při objevení nového objevu. Dále můžeme přidat detaily, které podporují představivost.

4) Na konec sumarizace vložíme zakončující větu. Sumarizace by neměla končit v místě, kde končí samostatný článek, ale v místě konce příběhu tohoto článku.

3.2.2 Definice ontologie

Jelikož jsou v této práci uvedené metody, které jsou a nejsou založené na ontologii, je zde tedy uvedena definice ontologie.

V minulých letech se velmi rozšířil zájem o aplikování technik související s ontologií. Nicméně v literatuře stále chybí unikátní definice. Uvádíme zde tedy jednu možnost a to Gruberovu definici ontologie [6]:

> *„An ontology is an explicit specification of some topics. It is a formal and declarative representation, which in-cludes the vocabulary (or names) for referring to the terms in a specific subject area and the logical state-ments that describe what terms are, how they are related to each other*.”
>
> Český překlad: „*Ontologie je explicitní vyjádření nějakých témat. Je to formální a deklarativní reprezentace, která zahrnuje slovník (nebo jména) odkazující na termy týkající se určité oblasti a logické prohlášení, jež popisuje, jaké to jsou termy a jaké mají k sobě vzájemné vztahy. “*

V podstatě se dá říci podle článku [5], že ontologie rozčleňuje svět do několika objektů, aby je mohla lépe popsat. Definice popsání těchto objektů nebo jejich reprezentace závisí na jednotlivých aplikacích. Zde je ontologie navržena jako analyzování a shromažďování sémantických informací tříd z článku. Předpokládá se, že každý článek obsahuje několik podtémat, použijeme ontologii na identifikování těchto podtémat článku a následné zakódování každého možného podtématu za použití nepřekrývající části ontologie.

3.2.3 Metoda založená na ontologii

Za použití sémantické informace, zakódované v ontologii, tento systém určí, která témata jsou užitečná pro extrakci paragrafů. Navržení a vytvoření ontologie jsou první dva kroky potřebné pro vytvoření sumarizačního systému.

V článku [5] je navržen postup pro metodu založenou na ontologii.

3.2.4 Metoda, jež není založená na ontologii

Metoda, jež není založená na ontologii, může být například metoda extrakční a cvičící, jež vypočítá frekvenci termů, délky vět a správná podstatná jména. Po získání těchto hodnot můžeme použít následujícího vzorce aplikujícího na každý bod pro získání ohodnocení [5]:

$$G_j = L(w_1 f_{j1} + w_2 f_{j2} + \cdots + w_n f_{jn})$$

Kde G_j je stupeň j-tého bodu; f_{ji} je hodnota i-té vlastnosti z j-tého bodu; w_i je váha i-té vlastnosti. L je 1, pokud bod má dostatečný počet slov, jinak je jeho hodnota 0.

3.2.5 Více-dokumentová sumarizace

- Extrakce událostí odkazující informace

V tomto případě můžeme říci, že soubor článků zahrnuje ty, jež uvádějí výskyt událostí a ty, jež uvádějí následující události [7]. V tomto typu souboru článků udávají následující články odkazy na události, jež byly již popsány v předchozích článcích a přidávají další informace související s událostmi. Proto předpokládáme, že vyhledání událostí ve více článcích a odkazující informace mezi těmito událostmi, je užitečné pro sumarizaci.

Metoda lexikální soudržnosti (Lexical cohesion) metoda je jedna možnost jak se vypořádat s odkazovými informacemi pro sumarizaci jednoho dokumentu. Nicméně, s cílem vypořádat se s odkazovými informacemi v jiných dokumentech, nemůžeme použít informaci jako je vzdálenost mezi dvěma větami. Takže je navrhnuta metoda pro extrakci událostní informace z novinového článku a určení události za použití podobnostního měření mezi dvěma událostmi. V tomto článku je „událost“ definovaná následovně:

Událost je informace popisující fakta a podobné informace konkrétního data.

- Extrakce událostí

Událost můžeme podle článku [7] popsat jako jednotku reprezentující vztah mezi různými články a měla by obsahovat užitečnou informaci, která identifikuje stejné události. Za účelem získaní kvalitní událostní informace z vět dokumentu, je dobré zkoumat hlubší strukturu vět v dokumentu (diskursní a anaforická analýza). V přídavku, je aktuální informace užitečná pro rozlišování podobných událostí (vyjádření tisku v květnu je jiné než vyjádření v dubnu). Na základě této diskuze se vybraly následující sloty pro definování události.

- Kořen je slovo, jež dominuje události.
- Modifikátor slov, modifikuje kořeny slov. Slova jsou tříděna do několika skupin, jako předmět a objekt slov pro slovesa, adjektiva a adnominální slova pro podstatná jména.
- Zápor představuje vyjadřovací metodu.
- Hloubka je délka cesty mezi kořeny událostí a kořen věty v analytickém událostním stromu.
- Datum je datum reprezentující událost. Tento slot není vyžadován k identifikování události.
- Datum článku je datum, v němž byl článek vydán.
- Kousky reprezentují seznam pozic slov ve větě.

V této metodě pro získání událostní informace se použijí následující kroky:

1) Aplikujeme Cabocha [8] pro získání analytického závislostního stromu.
2) Vybereme slovesa a podstatná jména, jež jsou modifikátory slov jakožto kandidáty „kořenů“ pro události.
3) Zkontrolujeme, jestli negativní výraz je nebo není zahrnut v kořenech a nastaví „zápor“ podle této analýzy.
4) Extrahujeme „modifikátorovou“ informaci z analyticky závislostního stromu. V této části se nastaví typy modifikátorů, použitím POS štítku a po postavení části. Modifikovaná informace nezahrnuje pouze slova, jež přímo závisí na kořenovém slovu, ale také modifikátory pro modifikovaná slova. Modifikátory pro modifikované slovo jsou řazeny do stejné kategorie modifikovaných slov.
5) Když se může získat informace o datu z věty, nastaví se toto datum jako „datum“ pro události, jež mají závislost s datovými slovy.
6) „Datum článku“ je získán z informace o článku.
7) „Hloubka“ a „kousky“ jsou vypočítány pomocí porovnáváním událostních informací s analytickým závislostním stromem.

- Použití odkazující událostní informace

Již existuje algoritmus, který by počítal váhy důležitosti vět v jednom dokumentu založeném na PageRank algoritmu. Algoritmus PageRank je ten, jenž dokáže vypočítat důležitost WWW stránek, za použití odkazové analýzy. Základní koncept algoritmu je distribuční důležitost stránky skrz linkovou strukturu. Jinými slovy, stránka, která má nashromážděno hodně důležitých odkazů z jiných stránek a odkazy ze stránky s vyšší důležitostí má vyšší důležitost oproti odkazu z jedné stránky s nižší důležitostí.

Další popis a postup pro výpočet důležitosti odkazů nalezneme v [7].

- Zvážení pozice věty a prvního dotazu

Jelikož v novinových článcích se mohou důležité věty vyskytovat na začátku článku, může být použita pozice věty pro výpočet důležitosti každé věty. Algoritmus v němž se nastaví důležitost každé stránky, je navrhován jako citlivý na téma. Tento algoritmus je navržen tak, aby počítal důležitost každé stránky podle kategorie, do níž stránka patří. Tento algoritmus můžeme označit jako upravený PageRank algoritmus a jeho podrobnější popis lze nalézt v článku [7].

- Přeuspořádání a zhutnění na základě podobnosti událostí

Je potřeba mechanismus pro detekci redundantních vět, jelikož při získávání konečných vět pro sumarizaci je šance, že vybereme tyto redundantní věty. Tento redundantní popis musí být odstraněn. Více podrobností v článku [7].

- Antologická sumarizace

Použitím metody příklad na bázi přístupu pro sumarizaci článku má následující výhody: vysoká modularita, nemusí se počítat váha pro každé slovo a velké využití kontextu. Výsledky experimentů (provedených podle této metody a představených v článku [9]) uvádějí, že sumarizovaný text obsahuje přibližně 60-ti procentní přesnost podle úsudku člověka.

Příklad na bázi přístupu vytváří jazyk podle napodobování instancí, které vznikly v metodě strojového překladu na základě antologie. Tato myšlenka je odvozena z pozorování, že člověk při překladu používá znalosti i z překladů, jež provedl někdy v minulosti. V úkolu strojového překladu byl tento přístup implementován a doposud dosáhl efektivních výsledků.

Jelikož člověk při sumarizaci používá své znalosti a dřívější poznatky, začíná se více zaměřovat na sumarizační metody založené na antologii, v tomto případě příklad báze sumarizace. Metoda příklad báze sumarizuje vstupní text ve třech krocích:

1) Získání podobných instancí na vstupu novinového článku ze sbírky instancí

2) Srovnání odpovídajících frází mezi vstupem novinového článku a podobnou instancí.

3) Zkombinování odpovídajících frází do formy souhrnu.

Můžeme říci, že metoda příkladu báze má následující výhody:

1) Velká modularita

   Je požadována snadné zlepšování a údržba k vytvoření užitečného obecného systému. Framework příkladu báze usnadňuje zlepšování systému už pouze tím, že stačí přidat instance.

2) Využití podobnosti před důležitostí

   Většina minulých metod na automatickou sumarizaci byla založena na extrakci. Tyto metody vypočítávají důležitost každého slova a následně podle toho vybírají určité věty. Samozřejmě je velice obtížné počítat tyto důležitosti tak, aby odpovídali lidskému smyslu pro sumarizaci. Metoda příklad báze tedy nepočítá s důležitostí slov, ale pracuje místo toho s podobností. Je jednodušší získat podobnost mezi dvěma výrazy než důležitost jednoho. (Tento poslední výrok je pouze předpoklad v článku)

3) Vysoká využitelnost lokálního kontextu

   Obecná statická metoda se snaží počítat pravděpodobnost každého slova, jež se objevuje v sumarizovaném korpusu. To může ztížit zachování lokálního kontextu, jelikož se statistický přístup zaměřuje na globální pravděpodobnost. Nicméně přístup příkladu báze se pokouší nalézt lokální podobnostní instanci z kolekce instancí, jež může zvýšit vhodnost vstupního kontextu.

Mnoho metod sumarizací vytvořilo sumarizaci z jedné věty, která byla vybrána například z novinového článku. Na rozdíl od těchto metod, příklad báze vytváří výslednou větu (sumarizaci) kombinováním frází. Tudíž může vytvořit sumarizaci s vysokou kompresí, jež zahrnuje informaci z více částí zdrojového textu.

Kapitola 4
Ohodnocování výsledků sumarizačních metod

4.1 Popis ohodnocování výsledků

Ohodnocování výsledků sumarizačních metod slouží k posouzení rozdílů výsledků mezi jednotlivými sumarizačními metodami. Samozřejmě můžeme mluvit o ohodnocování, které provádí člověk na základě svých vlastních zkušeností. Tyto výsledky jsou proto u každého jedince jiné, a kvůli tomu můžeme říci, že nejsou dostatečně přesné ke stanovení závěrů, jaká metoda je lepší. Budeme se tedy raději zabývat automatickým ohodnocováním výsledků sumarizačních metod, které je prováděno danou metodou, která se řídí daným algoritmem. Při správném zvolení metody již můžeme stanovit závěry o přesnosti sumarizačních metod.

4.2 Rozdělení metod ohodnocování výsledků

Vyhodnocování kvality automatické sumarizace se může podle článku [17] dělit následovně:

- Přímé (intrinsic)
    - Zaměřené na lingvistickou kvalitu textu

      Toto hodnocení je nejčastěji ohodnocováno člověkem. Každému souhrnu přiřadí hodnotu z předem definovaného měřítka (tabulky).
        - Gramatika
        - Neredundantnost
        - Srozumitelnost
        - Struktura a souvislost
    - Hodnocení obsahu

      Tyto metody jsou považovány jako hlavní přístup k určování kvality souhrnů (v těchto metodách je často používáno srovnávání výsledného souhrnu s „ideálním“ souhrnem).
        - Ko-selekční přístup

          Tyto metody jsou používány při použití metod pro extrakci vět.

          Vyhledává počet vět, „ideální“ (správný), jež obsahuje automatický souhrn.
            - Přesnost
            - Úplnost
            - F-skóre
            - Relativní užitečnost

        - Podobnostní míry

            Tyto metody raději porovnávají aktuální slova ve větě, než celé věty. Mají výhodu, že dokážou porovnávat jak extrakty vytvořené člověkem, tak extrakty vytvořené automatickou sumarizací, s abstrakcí napsanou člověkem (jež obsahuje nově napsané věty).

            - Kosinová podobnost
            - Překrytí obsahu
            - Nejdelší společný podřetězec
            - Společné n-gramy (ROUGE)
            - Ohodnocování vět (pyramidy)
            - Hodnocení na základě LSA

- Nepřímé (extrinsic)

    Tyto metody posuzují kvalitu na základě uplatnění souhrnu vzhledem k určité úloze.

    - Metody pro kategorizaci dokumentů
    - Metody pro vyhledávání informací
    - Metody pro zodpovídání dotazů

4.3 Popis metod ohodnocování výsledků

4.3.1 Přímé

- Zaměřené na lingvistickou kvalitu textu

    - Gramatika

        Výsledný text by neměl obsahovat slova nebo hodnoty, které nejsou skutečná slova (tj. značky) nebo také interrupční chyby a gramaticky špatně napsaná slova.

    - Neredundantnost

        Výsledný text by neměl obsahovat redundantní slova.

    - Srozumitelnost

        Podstatná jména a zájmena by měla ve výsledném textu být naprosto jasná. Například můžeme uvést, že zájmeno „on“ by mělo být uvedeno pouze v případě, že bude odkazovat na někoho v rámci celého souhrnu.

- Struktura a souvislost

Souhrn by měl být dobře strukturovaný a měl by být také koherentní.

- Hodnocení obsahu

o Ko-selekční přístup

- Přesnost, úplnost a F-skóre:

Za hlavní metody vyhodnocování kvality souhrnů ko-selekčního přístupu jsou přesnost, úplnost a F-skóre.

Přesnost (P) je číslo reprezentující počet vět vyskytujících se v obou systémech a ideální souhrn vydělen počtem vět v systémovém souhrnu.

Úplnost (R) je číslo reprezentující počet vět vyskytujících se v obou systémech a ideální souhrn vydělen počtem vět v ideálním souhrnu.

F-skóre je hodnota, jež kombinuje přesnost a úplnost. Nejjednodušší možnost jak spočítat F-skóre je spočítat harmonický průměr přesností a úplnosti:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

Lze uvést i složitější formulaci na spočítání F-skóre:

$F = \frac{(\beta^2+1) \cdot P \cdot R}{\beta^2 \cdot P + R}$, kde β je váhový faktor, jenž upřednostňuje přesnost při hodnotě $\beta > 1$ a při hodnotě $\beta < 1$ podporuje úplnost.

Hlavním problémem této metody je, že P a R hodnoty většinou neodpovídají hodnotám, které jsou vytvořeny úsudkem člověka. Může nastat případ, že dva hodnotící souhrny mají podle lidského úsudku přibližně stejnou hodnost, ale podle této metody je jedna ohodnocena podstatně vyšším skóre než druhá.

- Relativní užitečnost:

Tato metoda řeší hlavní problém přesnosti a úplnosti, zaveden pojem Relativní užitečnost (RU). Sumarizační model této metody reprezentuje všechny věty vstupu a jejich spolehlivostní hodnoty, pro zařazení těchto vět do souhrnu. Jako příklad lze uvést dokument s pěti větami (1 2 3 4 5), jež je

reprezentován jako [1/5 2/4 3/4 4/1 5/2]. Druhé číslo v páru reprezentuje úhel, který označuje větu, která by podle rozhodnutí člověka měla být zařazena do souhrnu. Tato hodnota je nazvána *užitečnost* věty. Je závislá na vstupním dokumentu, délce souhrnu a na rozhodnutí člověka. Pro výpočet relativní užitečnosti, čísla rozhodnutí, pro (N ≥ 1) je požadováno, aby se všem *n* větám v dokumentu přiřadilo pomocné skóre. Nejvýznamnější věty *e* podle vzorce níže, se označují jako extrahované věty o velikosti *e*. Pak tedy můžeme definovat metrický výkonnostní systém:

$$RU = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} u_{ij}}{\sum_{j=1}^{n} \epsilon_j \sum_{i=1}^{N} u_{ij}}$$

Kde u_{ij} je užitečnostní skóre věty *j* od ohodnovatele *i*, dále ϵ_j nabývá hodnoty 1 pro nejvýznamnější věty *e* v závislosti na součtu všech užitečnostních skóre od všech ohodnovatelů (soudců), jinak nabývá hodnoty 0, δ_j nabývá hodnoty 1 pro nejvýznamnější věty *e* extrahovaných systémem, jinak nabývá hodnoty 0.

o Podobnostní míry

- Kosinová podobnost

  Základní metoda hodnocení kvalit v části, podobnostní míra je kosinová podobnost.

$$\cos(X, Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (x_i)^2} \cdot \sqrt{\sum_i (y_i)^2}}$$

  Kde X je hodnocený souhrn a Y originální text. Dokumenty X a Y jsou reprezentovány vektory v prostoru slov.

- Překrytí obsahu lze počítat pomocí vzorce:

$$overlap(X, Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|}$$

  X a Y jsou reprezentace založené na souborech slov nebo lemmat. $\|X\|$ je velikost souboru X.

- Nejdelší společný podřetězec („Longest Common Subsequence - LCS“)

$$lcs(X, Y) = \frac{length(X) + length(Y) - edit_{di}(X, Y)}{2}$$

  Kde X a Y jsou reprezentace založené na sekvenci slov nebo lemmat, *lcs(X,Y)* je hodnota reprezentující délku nejdelšího společného podřetězce mezi X a Y, *length(X)* je délka řetězce X a $edit_{di}(X, Y)$ je editovaná vzdálenost X a Y.

- Společné n-gramy (ROUGE)

  ROUGE (Recall-Oriented Understudy for Gisting Evaluation), tato metoda je používána jako automatická ohodnovací metoda. Metoda má řadu opatření, jež jsou založena na podobnosti n-gramů. Předpokládejme, že nějaký počet zhodnocovatelů vytvoří referenční souhrny – referenční nastavení souhrnu (RSS). ROUGE-n skóre souhrnu, jenž je kandidát na výsledný souhrn, počítáme následovně:

$$ROUGE - n = \frac{\sum_{C \epsilon RSS} \sum_{gram_n \epsilon C} Count_{match}(gram_n)}{\sum_{C \epsilon RSS} \sum_{gram_n \epsilon C} Count(gram_n)}$$

  Kde $Count_{match}(gram_n)$ je maximální hodnota n-gramů, které se současně vyskytují v souhrnech, jenž jsou kandidáty na výsledné souhrny, referenční přehled a $Count(gram_n)$ je hodnota n-gramu v referenčním souhrnu.

  Měli bychom upozornit, že průměrné n-gram ROUGE skóre, ROUGE-n je metrické stažení.

- Ohodnocování vět (pyramidy)

Metoda ohodnocování vět je nová poloautomatická metoda. Základní myšlenkou této metody je nalézt jednotky sumarizačního obsahu (SCU_S), které se používají pro srovnávání informací v souhrnech. SCU_S pochází z anotací korpusu a nejsou větší než klauzule. Anotace začíná identifikovat podobné věty, po té začíná přesnější prověřování, které může nalézt přesněji související podčásti. SCU_S, jenž se více vyskytuje v manuálně vytvořených souhrnech, dostává vyšší váhu, takže pyramida se vytvoří až po *SCU* anotaci manuálně vytvořených souhrnů. Na vrcholku pyramidy se nachází SCU_S, které se objevují v největším počtu souhrnů a tudíž mají nejvyšší váhu. Čím níže v pyramidě je *SCU*, tím nižší je jeho váha a také tím menší je jeho výskyt ve více souhrnech. Pro ohodnocení SCU_S v odborných souhrnech se pak porovnávají již existující pyramidy a porovnává se kolik informací je shodných mezi odborným souhrnem a souhrnem manuálně vytvořeným. Tato metoda je velmi slibná, ale nicméně vyžaduje ještě nějakou anotaci.

4.3.2 Nepřímé

- Kategorizace dokumentů

Kvalita automatických souhrnů může být měřena podle vhodnosti roztřídění do určitých kategorií. Určuje se, zda hodnocení směřuje k obecnému shrnutí a zjišťuje se, jestli účinně zachycuje informace v daném dokumentu, jenž je třeba následně správně zařadit. Je tedy potřeba corpus dokumentů s jejich tématy dle řazení. Výsledky získané tříděním souhrnů jsou obvykle porovnávány se získanými výsledky z třídění celých dokumentů (horní mez) nebo s náhodným výběrem vět (dolní mez). Třídění může být prováděno jak manuálně, tak i automaticky (programem). Při použití automatického třídění musíme mít na paměti, že toto třídění obsahuje některé podstatné chyby. Je proto důležité rozlišovat chyby vytvořené automatickým klasifikátorem a chyby souhrnu. To se často provádí jen na základě porovnání výkonu systému s horní a dolní mezí.

- Vyhledávání informací

Vyhledávání informací (Information Retrieval – IR) je další možnost hodnocení kvality souhrnů. Význam korelace (Relevance correlation) je metoda založená na IR a posuzuje relativní pokles v získaném procesu, podle třídění souhrnů z celého dokumentu. Pokud souhrn zachytí hlavní body dokumentu, pak IR proces označí soubor takovýchto souhrnů (místo označování celých dokumentů), jenž by měl mít přijatelně dobré výsledky. Rozdíl mezi tím jak je dobře souhrn a celý dokument zpracován, by měl sloužit jako možnost hodnocení kvality souhrnů.

Předpokládejme daný dotaz *Q* a korpus dokumentů *D*. Vyhledávač ohodnotí všechny dokumenty v *D* podle důležitosti jejich dotazu *Q*. Pokud místo korpusu *D*, použijeme odpovídající souhrny všech dokumentů (jsou tedy náhradou všech celých dokumentů) a výsledný korpus souhrnů *S* bude ohodnocován stejným vyhledávačem podle důležitosti jejich dotazu, tak dostáváme jiné výsledky. V případě, že nahrazující souhrny původních dokumentů jsou dobré, lze předpokládat, že hodnocení budou podobná. Existuje několik metod, které měří podobnost ohodnocování. Jelikož vyhledávače vytváří relevantní skóre v dodatku hodnocení, můžeme použít i silnější podobnostní testy, například lineární korelaci.

Relevantní korelace (RC) je definována jako lineární korelace relevantních výsledků přidělených stejným IR algoritmem v různých datových souborech.

- Zodpovídání dotazů

Vnější hodnocení dopadu sumarizace v úloze zodpovídání dotazů bylo provedeno v [18]. Autoři vybrali čtyři připouštěcí testy pro absolventy řízení (GMAT – Graduate Management Admission Test), cvičení čtení s porozuměním. Cvičení měla více možností odpovědí, takže každá odpověď měla být vybrána ze seznamu odpovědí vedle každé otázky. Autoři zjišťovali, kolik otázek zodpověděli testované osoby správně, za daných situací. Nejdříve jim byly ukázány originální pasáže, pak automaticky generovaný souhrn, dále abstrakt vytvořený člověkem (profesionál, jenž vytváří abstrakty, měl pokyn vytvořit informativní abstrakt), a konečně, testované osoby měli vybrat správné odpovědi, před očima měli pouze otázky, nic jiného. Výsledky zodpovídání dotazů za různých podmínek byly porovnány.

Kapitola 5

Statistické metody

5.1 Popis statistických metod

V začátcích sumarizace byly metody hlavně statistické ve své podstatě a zaměřovaly se na frekvence výskytů nejdůležitějších konceptů v textu. Hlavní problém čistě statistických metod je podle článku [10], že nezohledňují kontext. Konkrétně, hledání zohlednění dokumentu se z velké části opírá na identifikaci a zachycení nejen duplikovaných termů, ale také na souvisejících termech. Tento koncept (známý jakou soudržnost), spojuje sémanticky související termy, jež jsou důležitou součástí uceleného textu. Nejjednodušší formou této soudržnosti je slovní soudržnost. Poprvé představen koncept „Slovních řetězců“ (Lexical Chains).

Tyto slovní řetězce reprezentují slovní soudržnost mezi libovolným počtem příbuzných slov. Slovní řetězce lze rozpoznat podle souboru slov, jež jsou sémanticky spřízněné. Použití slovních řetězců v sumarizaci textů je účinné, jelikož tyto vztahy jsou snadno identifikovány ve vstupním textu a pro výpočet nejsou nutné rozsáhlé znalostní báze. Použitím slovních řetězců lze staticky nalézt důležité koncepty pouze sledováním struktury dokumentu (lepší než hlubší porozumění sémantiky). Vše co je pro jejich výpočet nutné je obecná znalostní báze, která obsahuje podstatná jména a jejich sdružení. Tato sdružení zachycují pojetí vztahu, jako synonyma, antonyma a hyperonyma.

Naproti extrémnímu přístupu takových statistických metod je důležité se pokusit o „sémantické porozumění“ vstupního dokumentu. Ale v tomto přístupu se musí vytvořit sémantická reprezentace a musíme mít k dispozici hlavní specifickou bázi znalostí, což může být problém.

5.2 Příklady statistických metod

5.2.1 Naive-Bayes

Tato metoda je odvozená z Edmunsonovy metody [20], která umožňuje učení z dat. Klasifikující funkce rozděluje každou větu na to, jestli je vhodná pro extrakci nebo ne, za použití Naive-Bayes klasifikátoru. Nechť s je konkrétní věta, S soubor vět, které vytváří souhrn, a $F_1, F_2, \dots, F_k$ funkce. Předpokládaná závislost funkcí:

$$P(s \in S \,|F_1, F_2, \dots, F_k) = \frac{\prod_{i=1}^{k} P(F_i|s \in S) \cdot P(s \in S)}{\prod_{i=1}^{k} P(F_i)}$$

Funkce jsou v souladu s Edmunsonovými, ale jsou dodatečně přiřazeny „délka věty“ a „přítomnost velkých písmen slov“. Každé větě je přiřazena hodnota, podle vzorce výše, a pouze n nejlépe hodnocených vět je extrahováno. K posouzení systému byl použit korpus technických dokumentů s manuálními abstrakty následujícím způsobem: pro každou větu v manuálním abstraktu, autor manuálně analyzuje shodu se skutečnou větou dokumentu a vytvoří mapování (přesná shoda s větou, vyhovující spojení dvou vět, neshodné, atd.). Automatická extrakce je pak hodnocena podle tohoto mapování. Analýza vlastností odhaluje, že systém používající pouze pozici a funkce, dohromady s délkou vět a vlastnostmi vět, má nejlepší výsledky.

5.2.2 LSA

LSA – Latentní sémantická analýza („Latent Semantic Analysis“) je jedna z metod, jež dokazují, že metody nemusí být např. ryze statistické. Tato metoda je statisticko - algebraitská metoda. Výsledná sumarizace z této metody je extrakce smyslu slov a podobnosti vět, při čemž používá informace o slovech použitých v kontextu.

V každém odborném článku se může popis metod lišit, byl tedy vybrán popis LSA metody z článku [11]:

Jedno-dokumentová metoda založená na LSA:

Následná sumarizace je založena na takzvaném termovém přístupu, v tomto přístupu je nejdůležitější informace nalezena pomocí nejdůležitějšího termu (nebo tématu) a následným výběrem z dokumentu nejdůležitějších informací obsahujících toto téma nebo term. LSA je plně automatická matematicko-statistická technika pro extrahování a reprezentaci. Základní myšlenka je taková, že souhrn všech slov v kontextu, v kterém se dané slovo může a nemusí objevit, zajistí vzájemné vazby, jež určí podobnosti významů slov a podle toho je k sobě přiřadí. LSA metoda byla použita již v několika odvětvích. V sumarizaci je reprezentace dokumentu rozdělena do dvou kroků. V prvním kroku vytvoříme matici tvořenou větami, kde každý sloupec

reprezentuje váhový vektor (výskytu termu) věty v úvaze v souboru dokumentů. Podmínky od uživatele je získat vyšší váhu. Druhý krok je aplikovat na vytvořenou matici Singulární rozklad (SVD - „Singular Value Decomposition“). SVD odvodí latentní sémantické struktury dokumentu reprezentovaného maticí, rozčlení původní dokument do r lineárně nezávislých vektorů, které reprezentují hlavní témata dokumentu.

SVD může zachytit vztahy mezi termy, takže termy a věty můžeme dělit na sémantické základy spíše než na pouhá slova. Jestliže je ve slovním spojení nějaký vzor (slovo nebo část textu opakující se v dokumentu),tento vzor bude zachycen a reprezentován jedním ze singulárních vektorů. Velikost odpovídající singulární hodnoty označuje důležitý úhel tohoto vzoru v dokumentu. Každá věta obsahující tento slovní kombinační vzor bude promítán společně se singulárním vektorem a věta, která nejlépe reprezentuje tento vzorec, bude mít největší indexovou hodnotu s tímto vektorem. Předpokládejme, že každé jednotlivé slovo kombinovaného vzoru popisuje určité téma v dokumentu, na každý singulární vektor se můžeme dívat jako na reprezentaci tématu, jeho hodnota reprezentuje stupeň důležitosti tohoto tématu. Metoda vybere pro souhrn větu, jejíž vektorová reprezentace v matici $\Sigma^2 * V^T$ má nejlepší „délku“. Intuitivně je nápad vybrat větu, která má největší kombinovanou hodnotu všech důležitých témat.

Rozšíření na více dokumentů:

V této části se rozebere rozšíření na shluk dokumentů. Více dokumentová sumarizace je více komplexnější, než jedno-dokumentová sumarizace. V tomto případě je matice tvořená ze všech vět shluku dokumentů. Pak jsou věty ohodnoceny. Každá věta dostane skóre vypočítané stejně jako u jedno-dokumentového přístupu pomocí vektoru délky matic $\Sigma^2 * V^T$. Nyní lze vybrat větu (s nejvyšší hodnotou) pro souhrn. Nicméně, dva dokumenty napsané o stejném tématu nebo události mohou obsahovat podobné věty a tedy musí být řešena redundance. Je navrhnut následující proces: Před přidáním věty do souhrnu se zkontroluje, jestli již v souhrnu není nějaká podobná věta. Podobnost je měřena pomocí kosinové podobnosti v původním prostoru termů. Zde je definován práh (Z experimentů vyplývá, že odpovídající prahová hodnota cosínového úhlu mezi vektory dvou dokumentů je 0,6). Získaná věta by měla přibližně odpovídat požadavku uživatele. Pro zajištění tohoto požadavku, dostávají termy vyšší hodnotu ve vstupu do matice (z experimentů vyplývá, že nejlepší váha je 2. Takže požadovaný term dostane dvojnásobnou váhu než term, který není v požadavku). Další problém v tomto přístupu je, že upřednostňuje dlouhé věty. Je to přirozené, protože delší věta pravděpodobně obsahuje důležitější termy než věta krátká. Toto je řešeno dělením hodnoty věty pomocí $počet\ termů^{lk}$, kde lk je délkový koeficient (Nejlepší výsledky jsou s $lk = 0{,}4$). Experimenty ukazují dobré výsledky, když je nízká dimenze. Je to pro použití do 10-ti dimenzí (témat). Nicméně, témata

nejsou stejně důležité. Rozsah každého singulárního vektoru udržuje důležitost tématu. Aby to bylo více obecné, experimentuje se s jinými řídícími funkcemi v počítání finální matice určenou pro definování hodnot vět: $\Sigma^{\text{řídící}} * V^T$

5.3 Implementace statistické metody

Statistické metody většinou vycházejí z předpokladu, že důležitost termů se odráží ve frekvenci jejich výskytu. Následně se zaměříme na Luhnuv sumarizátor, jež je navržen podle frekvence termů a jejich výskytu.

Luhnuva metoda:

Důležitost každého slova je stanovena podle jeho výskytu v textu. Ovšem v případě příliš častého vyskytování slova v textu dochází ke klesání jeho důležitosti. Proto nelze měřit důležitost každého slova pouze podle jeho výskytu v textu. Významnost termu (term – slovo nebo i celá věta) t v dokumentu se tedy získává jako součin jeho frekvence výskytů (tf – term frequency) a převrácené frekvence dokumentu (idf – inverted dokument frequency). Věty, které obsahují důležité termy jsou následně zařazovány do výsledné sumarizace.

Tuto metodu můžeme shrnout a popsat v následujících bodech:

1) Vytvořte vektor frekvence termů tf_i pro každou větu *i* v textu.
   Frekvenci termů spočteme následujícím způsobem:
   a) Použijeme výpočet klasické frekvence termů $tf_{i,j}$, jež reprezentuje frekvenci termu (slova) *i* ve větě *j*, tedy počet výskytu termu ve větě.
   b) Použijeme výpočet normované frekvence termů $tf_{norm\ i,j}$:

$$tf_{norm\ i,j} = \frac{tf_{i,j}}{tf_{\max k,j}}$$

   Kde $tf_{i,j}$ je definováno v za a) a $tf_{\max k,j}$ je maximální frekvence přes všechny termy (slova) *k* ve větě *j*

2) Vytvořte vektor *D* (idf_i- inverzí frekvence termů) pro každou větu *i* v textu
   Inverzní frekvenci termu spočteme následujícím způsobem:

$$idf_j = \log\frac{N}{n_j}$$

   Kde *N* reprezentuje počet vět v textu a n_j reprezentuje počet vět článku, v nichž se vyskytuje term (slovo) *j*.

3) Vypočtěte významnost (váhu) w_i každé věty *i* v textu
   Tuto významnost spočteme za pomoci skalárního součinu vektorů frekvence termů a vektorů inverzní frekvence termů:

$$w_i = tf_i \times idf_i$$

4) Vytvoříme výslednou sumarizaci z *X* vět, jež mají nejvyšší významnost (váhu). *X* reprezentuje počet vět, z nichž je výsledná sumarizace tvořena.

Jelikož v takto vytvořené sumarizaci by byl nedostatek, jenž by preferoval jedno hlavní téma textu (toto téma by pak bylo ve výsledku zastoupeno redundantně). Proto je tato metoda rozšířena o další body a zároveň musí být bod 4 změněn. Může se tedy psát:

4) Do výsledné sumarizace je zařazena pouze jedna věta *v*, jež má nejvyšší významnost.
5) Pokud počet vět ve výsledné sumarizaci odpovídá požadovanému počtu vět, je algoritmus zastaven, pokud tomu tak není, pokračuje bodem 6
6) Z textu odstraníme větu *v* ,všechny její vektory a významnost této věty.
7) Opakujeme od bodu 1 (v textu již není věta *v*).

Po vybrání těchto *X* vět, jsou tyto věty seřazeny podle původního umístění v článku a až poté je vytvořen výstup z programu v podobě textového souboru s výsledky.

Jelikož tato metoda byla spíše vytvořena pro sumarizaci více dokumentů, byla upravena do této podoby, aby vyhovovala pro sumarizaci jediného článku (textu).

Tato metoda byla naprogramována podle popisu výše v programovacím jazyce Python, zdrojový kód je přiložen v sekci příloha.

Z článku [3] byly čerpány podklady pro vytvoření této metody.

5.4 Modifikace Luhnovy metody

Byla vybrána modifikace Luhnovy metody, jež je popsána v článku [12]:

Luhnova metoda jako taková, je založena na systematickém přístupu sumarizace, jež v dnešní době tvoří jádro tohoto oboru. V této extrakční metodě je každé větě přiřazen faktor důležitosti a věty s nejvyšší hodnotou tohoto faktoru jsou vybírány do konečné sumarizace. K tomu, abychom mohli vypočítat tento důležitý faktor věty, musíme vytvořit takzvaný „bazén významných slov“, který se dá definovat jako dvě slova, jejichž frekvence je mezi odříznutými vysokými a nízkými frekvencemi, jež se dají pozměňovat a tím měnit vlastnosti sumarizačního systému. Po dokončení tohoto kroku můžeme důležitý faktor věty vypočítat podle Luhnovy metody následovně:

1) Nastavíme omezení L, jež reprezentuje vzdálenost v které by mohla ležet nějaká dvě významná slova, která by se mohla považovat za významně související.
2) Najdeme část věty, jež je ohraničena významnými slovy, která nejsou od sebe vzdálené více než L nevýznamných slov.
3) Spočteme počet významných slov v této části a tento počet vydělíme celkovým počtem slov v této části. Výsledkem je významný faktor, jež se vztahuje k S.

S cílem přizpůsobit tento postup pro webové stránky, musíme provést úpravu Luhnova algoritmu. V tomto úkolu klasifikovat web je kategoriální informace každé stránky známá v trénovacích datech, takže výběr důležitých slov může být zpracován v rámci každé kategorie. Tedy touto cestou vytvoříme bazén důležitých slov pro každou kategorii tak, že vybereme slova s vysokou frekvencí, ale až po odstranění stop slov v dané kategorii. Po výběru těchto slov použijeme Luhnovu metodu pro výpočet významného faktoru.

Uvedeme dvě výhody této modifikace. Za prvé, v sumarizaci je využívána předchozí znalost kategorií. Za druhé, některá hlučná slova, která mohou být na určité stránce poměrně častá, budou odstraněna za pomoci více-dokumentové statistiky. Při sumarizaci webových stránek v trénovací sadě je významné skóre každé věty počítáno podle bazénu významných slov, odpovídající jeho kategorii štítku. Pro webovou stránku v nejlepší sadě nemáme informaci o kategorii. V tomto případě budeme počítat významný faktor podle jiným bazénům významných slov, přes všechny kategorie zvlášť. Významné skóre dané věty bude průměrné přes všechny kategorie a bude podle S_{luhn}. Sumarizace této stránky bude vytvořena z vět s nejvyššími skóre.

Kapitola 6

Grafové metody

6.1 Popis grafových metod

Graf je podle článku [13] nejefektivnější reprezentace vztahů mezi dvěma proměnnými. Nejčastěji se umisťuje na osu x nezávislá proměnná (může být změněna) a na osu y závislá proměnná (reaguje na změny v nezávislé proměnné).

Základní popis grafových metod je popsán v článku [14] následovně:

Iterační grafové algoritmy, jako například HITS nebo PageRank (používaný googlem), byly původně navrženy jako prohledávací nástroj spojovací struktury pro ohodnocení webových stránek. Později se též úspěšně aplikovaly v jiných oblastech jakož například v citační analýze, v sociálních sítích atd. V grafovém ohodnovacím algoritmu je důležitost vrcholu v grafu iteračně vypočítána z celého grafu. Model, který byl založen na grafové bázi, byl aplikován na zpracování přirozeného jazyka, jež vyústilo v algoritmus pojmenovaný TextRank. Stejné ohodnující principy založené na grafové bázi byly aplikovány na sumarizaci. Graf je vytvářen přidáváním vrcholu každé větě v textu. Na vzájemném propojení vět jsou mezi vrcholy stanoveny hrany. Tato propojení jsou definována za pomoci podobnostních vztahů, kde tato podobnost je měřena jako funkce překrývání obsahu. Překrytí dvou vět může být definováno jako běžný počet hodnot mezi lexikální reprezentací těchto dvou vět. Následně můžeme tedy aplikovat iterační část algoritmu na vytvořený graf, reprezentující věty. Po dokončení procesu jsou vrcholy (věty) tříděny podle výsledných ohodnoceních. Nejlépe ohodnocené věty jsou následně zahrnuty do výsledků.

6.2 Příklady grafových metod

6.2.1 Metoda HITS

Hyperodkazové hledání indukovaného tématu (HITS - Hyperlinked Induced Topic Search) [15], je interační algoritmus vytvořen pro ohodnocování webových stránek podle jejich stupně „autority“. HITS algoritmus rozhoduje mezi „autoritami“ (webové stránky s velkým počtem příchozích odkazů) a „hubů“ (webové stránky s velkým počtem odchozích odkazů). Pro každý vrchol, HITS vytváří dva sety skóre („autority“ skóre a „hub“ skóre).

6.2.2 PageRank

PageRank je pravděpodobně jeden z nejpopulárnějších ohodnovacích algoritmů podle článku [15], a byl navržen jako metoda pro analýzu webových odkazů. Na rozdíl od jiných ohodnovacích algoritmů se PageRank integruje na dopad jak příchozích tak odchozích odkazů do jednoho modelu a proto je produkován jen jeden set skóre:

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

kde d je parametr, který je mezi 0 a 1.

Pro každý tento algoritmus, začínající od libovolných hodnot přiřazených ke každému uzlu v grafu, výpočet se opakuje tak dlouho, až dokonverguje pod danou hranici. Po doběhnutí algoritmu je skóre spojeno s každým vrcholem, který reprezentuje „důležitost“ nebo „sílu“ tohoto daného vrcholu v grafu. Mělo by se brát v potaz upozornění, že konečné hodnoty nejsou ovlivněny volbou počáteční hodnoty, ale pouze počet iterací do dané hodnoty mohou být různé.

6.3 Implementace grafové metody

Pro implementaci v programovacím jazyce Python byla vybrána grafová metoda LexRank, čerpáno z článku [16]. Tato metoda je extraktivní a byla upravena pro vytvoření sumarizace článku. Můžeme tedy napsat tento algoritmus následovně:

1) Načteme článek (text), který chceme sumarizovat. Tento text je načten do pole po větách (každý prvek pole obsahuje jednu větu), prvky tohoto pole jsou vytvořeny jako pole slov (každý prvek pole obsahuje slovo nebo znak dané věty).

2) Vypočteme inverzní frekvenci termů (idf) pro každý term (slovo nebo znak) následujícím způsobem:

$$idf_j = \log \frac{N}{n_j}$$

Kde *N* reprezentuje počet vět v článku a n_j reprezentuje počet vět článku, v nichž se vyskytuje term (slovo) *j*.

3) Vypočteme normovanou frekvenci termů (tf_{norm}) pro každý term (slovo nebo znak) následujícím způsobem:

$$tf_{norm\ i,j} = \frac{tf_{i,j}}{tf_{\max k,j}}$$

Kde $tf_{i,j}$ je reprezentuje frekvenci termu i ve větě j, tedy udává počet (výskytů) termu ve větě.
$tf_{\max k,j}$ je maximální frekvence přes všechny termy *k* ve větě *j*, neboli term s nejvyšší hodnotou tf.

4) První krok k vytvoření kosinové matice o rozměrech [n,n].
Modifikovaná kosinová věta (potřebná k vytvoření kosinové matice):

$$cosMat[i,j] = \frac{\sum_{\omega \in i,j} tf_{\omega,i}\, tf_{\omega,j} (idf_\omega)^2}{\sqrt{\sum_{x_i \in x} \left(tf_{x_i,x} idf_{x_i}\right)^2} \times \sqrt{\sum_{y_i \in y} \left(tf_{y_i,y} idf_{y_i}\right)^2}}$$

ω reprezentuje slovo, jež se vyskytuje jak ve větě *i*, tak ve větě *j*.
Kde *i* a *j* reprezentují indexy vět v poli (tvořeného větami článku).

Před vytvořením kosinové matice stanovíme takzvanou „prahovou hodnotu“ *t*.
Podle experimentů v článku [16] se nejlepších výsledků dostávalo při nastavení této „prahové hodnoty“ na hodnotu: *t = 0.1* (použito v algoritmu).
Nyní můžeme dosazovat hodnoty do kosinové matice následovně:

Algoritmický zápis:

```
if kosMat[i,j] > t then
        kosMat[i,j] = 1;
        stupne[i] ++;
end
else
        kosMat[i,j] = 0;
end
```

Slovní zápis:

a) Pokud na pozici [i,j] je vypočítaná hodnota pomocí modifikované kosinové věty větší než prahová hodnota, přiřadíme do matice na pozici [i,j] hodnotu 1 a do pole stupňů na pozici *i* přičteme 1.

b) Do kosinové matice přiřadíme na pozici [i,j] hodnotu 0.
Tento bod je proveden pouze v případě, že není splněna podmínka v bodě a).

V tomto bodě jsme tedy získali kosinovou matici (před konečnou úpravou) a pole stupňů této matice.

5) Druhý a konečný krok pro získání kosinové matice:
V tomto bodě v podstatě doopravíme kosinovou matici z minulého bodu do následujícím způsobem:

Algoritmický zápis:

```
for i = 1 to n do
    for j = 1 to n do
        kosMat[i,j] = kosMat[i,j] / supne[i]
    end
end
```

Slovní zápis:
Hodnotu kosinové matice na pozici [i,j] nahradíme hodnotou spočtenou vydělením původní hodnoty kosinové matice na pozici [i,j] hodnotou z pole stupňů na pozici [i].

V tomto bodě jsme tedy dostaly výslednou kosinovou matici o rozměrech [n,n], kde n reprezentuje počet vět článku.

6) Vytvoříme konečný vektor *L*, jež obsahuje konečné LexRank váhy.
Tento vektor vytvoříme pomocí „Silové metody“ (Power Method). Vstupní hodnoty do této metody tvoří, kosinová matice, její velikost *n* a tolerance chyby ε.
Výstupem je vektor s konečnými výsledky. Metoda pracuje následovně:

V metodě se pracuje s transponovanou kosinovou maticí, tak si můžeme původní kosinovou matici transponovat a uložit pod označením M^T.
Vytvoříme vektor p_0, jenž bude mít velikost *n* (velikost kosinové matice) a každý prvek tohoto vektoru bude mít hodnotu: $\frac{1}{n}$
Vytvoříme prvek *t* a nastavíme jeho hodnotu na nulu (t = 0).
Tolerance chyby ε, jež vstupuje do této metody je nastavena na: $\varepsilon = 0.1$
Hodnota δ je nastavena na hodnotu: $\delta = 1$

Nyní v cyklu opakujeme následné kroky, dokud bude platit podmínka $\delta > \delta$:
Prvek *t* změníme následovně: *t = t + 1*
Spočteme vektor: $p_t = M^T p_{t-1}$
Spočteme δ následovně: $\delta = \|p_t - p_{t-1}\|$

Po vyskočení z tohoto cyklu tato metoda vrací poslední spočtený vektor p_t, jež reprezentuje výsledky (váhy vět).

Metoda nám tedy vypočetla vektor *L* jenž obsahuje výsledné váhy důležitosti vět.

7) Vybereme *x* (reprezentuje číslo počtu vět, jež uživatel požaduje extrahovat) vět původního článku s nejvyššími vahami LexRank, jež jsou uvedené ve vektoru *L*. Tyto věty následně seřadíme podle původního umístění v článku.

Na konci programu je vytvořen výstup do textového souboru obsahujícího výsledky (výslednou sumarizaci) této metody z daného vstupního textu.

Kapitola 7

Testování a porovnávání implementovaných metod

7.1 Originální články

V této části jsou uvedeny originální texty, na nichž byla aplikována automatická sumarizace. Implementované metody: Luhnova metoda a LexRank metoda.

7.1.1 Článek č.1

Na kocovinu je nejlepším lékem velké množství tekutin , nejraději čisté vody , případně ovocného čaje , a dostatek spánku . Jako radu k oslavám konce roku a vítání nového to řekl Pavel Suchánek z Fóra zdravé výživy . Nejlepší podle něj ale je , aby člověk ke kocovině vůbec nedospěl , aby tedy pil jen rozumné množství alkoholu . Nebude pak mít starost , jak se s kocovinou vypořádat . " Ale pokud už člověk ví , že překročil míru , která by mohla ukazovat - hodně silná káva - rozpustit více kostek cukru či med ve sklenici vody a vypít - masový vývar - velká dávka vitaminu C - kyselé okurky nebo zavináče , ty vyrovnají kyselost v žaludku a udělá se vám po nich dobře - hrnek zeleného čaje urychlující regeneraci , pomáhá také čaj z heřmánku či máty zklidní váš žaludek a zmírní bolest hlavy Zdroj : www.ulekare.cz na možnou kocovinu , tak jednoznačně velké množství tekutin a dostatečné množství spánku , " uvedl . Někomu se podle Suchánka osvědčilo také malé množství nízkoalkoholového piva , to on ale nedoporučuje . Vystřízlivění urychlí silná káva , voda s cukrem , velká porce masového vývaru , pořádná dávka vitaminu C , kyselé okurky nebo zavináče , které vyrovnají kyselost v žaludku . Regeneraci urychlí zelený čaj . Odvar z heřmánku nebo máty zklidní žaludek a zmírní bolest hlavy . Rychle na nohy postaví směs syrového vajíčka s octem a solí doplněná ostrým kečupem . Před kocovinou chrání jídlo , není

dobré pít nalačno . Výhodné je dát si něco tučného , tuk na sebe váže alkohol , a tak žaludek a játra mají méně práce s jeho zpracováním . Nejlepší je vůbec nepít . Pomůže i výmluva Kocovině se podle lékařů úspěšně vyhne také ten , kdo alkohol vůbec nepije . Není to snadné , člověk se musí naučit říkat ne . Komu se to zdá obtížné , může si připravit výmluvu , například že bere antibiotika , bude řídit , jde ráno do práce nebo měl nedávno žloutenku a má zakázáno pít . " Pokud jste zdatný herec , můžete si koupit rybízový džus a vydávat ho za víno , popíjet nealkoholické pivo a o půlnoci si připít dětským šampaňským , " radí lékaři na webu . Pro větší efekt si člověk může nealkoholický nápoj přelít do láhve od alkoholické varianty . Dobré je připravit program , aby se hosté bavili . Pití alkoholu totiž často plní jakousi náhradní roli zábavy . Velká legrace se dá zažít například u Twisteru , kdy soutěžící pokládají ruce a nohy na určitá místa nakreslená na herním plánu a snaží se nespadnout .

7.1.2 Článek č.2

Článek o životě bezdomovců v tunelech Las Vegas se stal nejčtenější Návštěvou roku 2010. V závěsu boduje rodinný dům s vlastní elektrárnou , který má i nejvíce diskusních příspěvků . Na třetí příčku dosedl Jaroslav Dušek se svým hliněným domkem . Podívejte se na sedm čtenářsky nejúspěšnějších návštěv letošního roku . Návštěvy patří mezi čtenáři mezi nejvyhledávanější , v roce 2010 je však " porazila " rubrika Stavba , a to hned dvakrát . Suverénním vítězem v zájmu čtenářů se stal Domek za 700 tisíc i s nábytkem navrhl český architekt . Dřevěný modulový dům architekta Štěpána zaujal výrazně kolemjdoucí na Václavském náměstí , kde byl v říjnu poprvé vystavený , tak více než 295 tisíc čtenářů iDNES . cz . Freedomek an Václavském náměstí zůstane do konce akce Architecture Week , která končí 17. 10. 2010 Na druhém místě skončil opět článek se stavební tematikou . Na dům měl mladý pár jen milion korun , architekti si s ním vystačili . I v tomto případě si jeho tvůrci zvolili jako základní materiál dřevo . Dům je zasazený do svahu - vizualizace Na třetím místě ve čtenosti (a nejčtenější z Návštěv) skončil článek o bydlení bezdomovců . 1. Americký sen v Las Vegas : v tunelech zde přežívají stovky bezdomovců Steven a Kathryn ve své provizorní ložnici v tunelu pod Las Vegas . Steven kvůli své drogové závislosti přišel o práci . Dnes se živí vybíráním zapomenutých mincí v automatech Las Vegas má problém . V centru amerického hazardu přibývají bezdomovci . Často jde o lidi , jimž banky zabavily dům kvůli nesplácené hypotéce . V pouštním městě pro ně není dostatek přístřeší , a tak nezřídka nacházejí úkryt v tunelech pod městem . Podle televizní stanice CBS News jich tak žijí stovky . 2. Nepotřebují ČEZ ani nikoho jiného . Ve svém domě mají vlastní elektrárnu Štítové stěny jsou obezděny kamenem pocházejícím z větší části ze starého stavení , které manželé koupili jako ruinu . Další kámen byl posbírán v okolních lesích Slunce se opírá do svahu , na jehož vršku vyrostla atypická stavba z kamene , dřeva a skla . Dům na pozemku v jižních Čechách poskytuje mladým manželům příjemné bydlení a navíc je energeticky soběstačný . 3. Herec Dušek vlastníma rukama postavil hliněný domek V Jindřichovicích pod Smrkem jsou dveře otevřené všem odvážným zájemcům o vyzkoušení přebývání v hliněném domku v zimě Ke stavbě hliněného domku stačí pár kamarádů , důkladná příprava a měsíc dovolené . Veškerá práce se dá zvládnout ručně , materiál na zdivo si nakopete přímo na svém pozemku .

Takovou hliněnku si postavil i herec Jaroslav Dušek , který zjistil , že bez stavební firmy je stavění radost . 4. Vyměnila Prahu za samotu u lesa . Dům s úžasným výhledem stál milion a půl Jednoduchý přízemní domek má rozlohu 80 metrů čtverečních Po dvanácti letech udělala Johana radikální řez . Dala výpověď , prodala pražský byt a koupila parcelu nedaleko Benešova . Do pozemku u lesa s nádherným výhledem se okamžitě zamilovala . Po byrokratickém kolečku se pustila do stavby vysněného rodinného domu . Od jara v něm žije s fenkou Lucy . 5. Režisérka Třeštíkova bydlí v úžasném půdním bytě na Letné . Splnil se jí sen Černobílá kombinace mívá zaručen úspěch Vynikající dokumentaristka Helena Třeštíková a její muž Michael , vášnivý sběratel umění a architekt , si postavili skvělý půdní byt . Hlavním aktérem , který měl volnou ruku v rekonstrukci obrovského prostoru , byl manžel Michael . K modernímu pojetí přistoupil s nadhledem . 6. Dřevostavbu , která vypadá jako staré venkovské stavení , smontovali za 14 dní Pohled ze dvora . Tvar domu , velikost , rozmístění a členění oken odpovídají typické venkovské zástavbě . Nová dřevostavba , která vyrostla na návsi malé pošumavské vesničky , je moderní od sklepa až na půdu . Přitom na první pohled vypadá jako tradiční venkovské stavení , kterých je v okolí plno . Málokdo by věřil , že se tento dům stavěl moderní montovanou technologií . 7. Neuvěřitelný byt v Praze má vířivku i vlastní kino . Je řízený iPhonem Obývacímu prostoru s olejovanými dubovými parketami vévodí sedací souprava Lava (Cor) sestavená z více modulů , kterou doplňují stolky Ameo (Walter Knoll) . Prostor osvětlují svítidla značek Dark , Ingo Maurer a Artemide . Po návratu z dlouhodobého pobytu v zahraničí hledali Honza s Radkem místo , které by představovalo jejich skutečný domov . Našli ho v jedné z pražských rezidenčních vil . Obývají tam velkolepý dvoupodlažní interiér o výměře 400 m .

7.2 Výsledky z testovaných článků

7.2.1 Úvod

- Výsledky statistické metody

Na pět článků aplikujeme Luhnuv algoritmus, jenž byl naprogramován v jazyce Python. Výstup z tohoto algoritmu bude reprezentován sumarizací složenou z pěti a dále z deseti vět. Budou zde uvedeny tedy i váhy jak pro verzi sumarizace s pěti, tak i s deseti větami. Pro větší přehlednost jsou věty seřazeny podle výskytu v původním dokumentu.

- Výsledky grafické metody

Na pět článků aplikujeme LexRank algoritmus, jenž byl naprogramován v jazyce Python. Výstup z tohoto algoritmu bude reprezentován sumarizací složenou z pěti a dále z deseti. Budou zde uvedeny tedy i váhy jak pro verzi sumarizace s pěti, tak i s deseti větami. Pro větší přehlednost jsou věty seřazeny podle výskytu v původním dokumentu.

7.2.2 Výsledky

- Článek č.1

  - **Výsledky statistické metody (reprezentované sumarizací z pěti vět)**

Jako radu k oslavám konce roku a vítání nového to řekl Pavel Suchánek z Fóra zdravé výživy . (20.3113299523)

" Ale pokud už člověk ví , že překročil míru , která by mohla ukazovat - hodně silná káva - rozpustit více kostek cukru či med ve sklenici vody a vypít - masový vývar - velká dávka vitaminu C - kyselé okurky nebo zavináče , ty vyrovnají kyselost v žaludku a udělá se vám po nich dobře - hrnek zeleného čaje urychlující regeneraci , pomáhá také čaj z heřmánku či máty zklidní váš žaludek a zmírní bolest hlavy Zdroj : www.ulekare.cz na možnou kocovinu , tak jednoznačně velké množství tekutin a dostatečné množství spánku , " uvedl . (26.4301905964)

Někomu se podle Suchánka osvědčilo také malé množství nízkoalkoholového piva , to on ale nedoporučuje . (15.0430988235)

Pomůže i výmluva Kocovině se podle lékařů úspěšně vyhne také ten , kdo alkohol vůbec nepije . (16.9447137222)

Pro větší efekt si člověk může nealkoholický nápoj přelít do láhve od alkoholické varianty . (15.8925738341)

  - **Výsledky grafové metody (reprezentované sumarizací z pěti vět)**

Nejlepší podle něj ale je , aby člověk ke kocovině vůbec nedospěl , aby tedy pil jen rozumné množství alkoholu . (0.0877551020408)

" Ale pokud už člověk ví , že překročil míru , která by mohla ukazovat - hodně silná káva - rozpustit více kostek cukru či med ve sklenici vody a vypít - masový vývar - velká dávka vitaminu C - kyselé okurky nebo zavináče , ty vyrovnají kyselost v žaludku a udělá se vám po nich dobře - hrnek zeleného čaje urychlující regeneraci , pomáhá také čaj z heřmánku či máty zklidní váš žaludek a zmírní bolest hlavy Zdroj : www.ulekare.cz na možnou kocovinu , tak jednoznačně velké množství tekutin a dostatečné množství spánku , " uvedl . (0.0711734693878)

Vystřízlivění urychlí silná káva , voda s cukrem , velká porce masového vývaru , pořádná dávka vitaminu C , kyselé okurky nebo zavináče , které vyrovnají kyselost v žaludku . (0.0682798833819)

Výhodné je dát si něco tučného , tuk na sebe váže alkohol , a tak žaludek a játra mají méně práce s jeho zpracováním . (0.0611917488084)

Komu se to zdá obtížné , může si připravit výmluvu , například že bere antibiotika , bude řídit , jde ráno do práce nebo měl nedávno žloutenku a má zakázáno pít . (0.0559425278048)

- **Výsledky statistické metody (reprezentované sumarizací z deseti vět)**

Jako radu k oslavám konce roku a vítání nového to řekl Pavel Suchánek z Fóra zdravé výživy . (20.3113299523)

" Ale pokud už člověk ví , že překročil míru , která by mohla ukazovat - hodně silná káva - rozpustit více kostek cukru či med ve sklenici vody a vypít - masový vývar - velká dávka vitaminu C - kyselé okurky nebo zavináče , ty vyrovnají kyselost v žaludku a udělá se vám po nich dobře - hrnek zeleného čaje urychlující regeneraci , pomáhá také čaj z heřmánku či máty zklidní váš žaludek a zmírní bolest hlavy Zdroj : www.ulekare.cz na možnou kocovinu , tak jednoznačně velké množství tekutin a dostatečné množství spánku , " uvedl . (26.4301905964)

Někomu se podle Suchánka osvědčilo také malé množství nízkoalkoholového piva , to on ale nedoporučuje . (15.0430988235)

Odvar z heřmánku nebo máty zklidní žaludek a zmírní bolest hlavy . (10.9167511971)

Rychle na nohy postaví směs syrového vajíčka s octem a solí doplněná ostrým kečupem (14.5074317388)

Výhodné je dát si něco tučného , tuk na sebe váže alkohol , a tak žaludek a játra mají méně práce s jeho zpracováním . (11.4430113077)

Pomůže i výmluva Kocovině se podle lékařů úspěšně vyhne také ten , kdo alkohol vůbec nepije . (16.9447137222)

" Pokud jste zdatný herec , můžete si koupit rybízový džus a vydávat ho za víno , popíjet nealkoholické pivo a o půlnoci si připít dětským šampaňským , " radí lékaři na webu . (12.5234555245)

Pro větší efekt si člověk může nealkoholický nápoj přelít do láhve od alkoholické varianty . (15.8925738341)

Velká legrace se dá zažít například u Twisteru , kdy soutěžící pokládají ruce a nohy na určitá místa nakreslená na herním plánu a snaží se nespadnout . (14.2377637856)

- **Výsledky grafové metody (reprezentované sumarizací z deseti vět)**

Jako radu k oslavám konce roku a vítání nového to řekl Pavel Suchánek z Fóra zdravé výživy . (0.047619047619)

Nejlepší podle něj ale je , aby člověk ke kocovině vůbec nedospěl , aby tedy pil jen rozumné množství alkoholu . (0.0877551020408)

Nebude pak mít starost , jak se s kocovinou vypořádat . (0.047619047619)

" Ale pokud už člověk ví , že překročil míru , která by mohla ukazovat - hodně silná káva - rozpustit více kostek cukru či med ve sklenici vody a vypít - masový vývar - velká dávka vitaminu C - kyselé okurky nebo zavináče , ty vyrovnají kyselost v žaludku a udělá se vám po nich dobře - hrnek zeleného čaje urychlující regeneraci , pomáhá také čaj z heřmánku či máty zklidní váš žaludek a zmírní bolest hlavy Zdroj : www.ulekare.cz na možnou kocovinu , tak jednoznačně velké množství tekutin a dostatečné množství spánku , " uvedl . (0.0711734693878)

Vystřízlivění urychlí silná káva , voda s cukrem , velká porce masového vývaru , pořádná dávka vitaminu C , kyselé okurky nebo zavináče , které vyrovnají kyselost v žaludku . (0.0682798833819)

Regeneraci urychlí zelený čaj . (0.047619047619)

Před kocovinou chrání jídlo , není dobré pít nalačno . (0.047619047619)

Výhodné je dát si něco tučného , tuk na sebe váže alkohol , a tak žaludek a játra mají méně práce s jeho zpracováním . (0.0611917488084)

Komu se to zdá obtížné , může si připravit výmluvu , například že bere antibiotika , bude řídit , jde ráno do práce nebo měl nedávno žloutenku a má zakázáno pít . (0.0559425278048)

" Pokud jste zdatný herec , můžete si koupit rybízový džus a vydávat ho za víno , popíjet nealkoholické pivo a o půlnoci si připít dětským šampaňským , " radí lékaři na webu . (0.0489680964979)

- Článek č.2

- **Výsledky statistické metody (reprezentované sumarizací z pěti vět)**

Článek o životě bezdomovců v tunelech Las Vegas se stal nejčtenější Návštěvou roku 2010. V závěsu boduje rodinný dům s vlastní elektrárnou , který má i nejvíce diskusních příspěvků . (36.5204342781)

Freedomek an Václavském náměstí zůstane do konce akce Architecture Week , která končí 17. 10. 2010 Na druhém místě skončil opět článek se stavební tematikou . (35.1580462714)

Dům je zasazený do svahu - vizualizace Na třetím místě ve čtenosti (a nejčtenější z Návštěv) skončil článek o bydlení bezdomovců . (31.0557972559)

Ve svém domě mají vlastní elektrárnu Štítové stěny jsou obezděny kamenem pocházejícím z větší části ze starého stavení , které manželé koupili jako ruinu . (34.7581624662)

Dům s úžasným výhledem stál milion a půl Jednoduchý přízemní domek má rozlohu 80 metrů čtverečních Po dvanácti letech udělala Johana radikální řez . (33.2530494941)

- **Výsledky grafové metody (reprezentované sumarizací z pěti vět)**

Článek o životě bezdomovců v tunelech Las Vegas se stal nejčtenější Návštěvou roku 2010. V závěsu boduje rodinný dům s vlastní elektrárnou , který má i nejvíce diskusních příspěvků . (0.0289855072464)

Dům je zasazený do svahu - vizualizace Na třetím místě ve čtenosti (a nejčtenější z Návštěv) skončil článek o bydlení bezdomovců . (0.0317028985507)

1. Americký sen v Las Vegas : v tunelech zde přežívají stovky bezdomovců Steven a Kathryn ve své provizorní ložnici v tunelu pod Las Vegas . (0.0271739130435)

Veškerá práce se dá zvládnout ručně , materiál na zdivo si nakopete přímo na svém pozemku . (0.0258907004831)

Nová dřevostavba , která vyrostla na návsi malé pošumavské vesničky , je moderní od sklepa až na půdu . (0.0273060084541)

- **Výsledky statistické metody (reprezentované sumarizací z deseti vět)**

Článek o životě bezdomovců v tunelech Las Vegas se stal nejčtenější Návštěvou roku 2010. V závěsu boduje rodinný dům s vlastní elektrárnou , který má i nejvíce diskusních příspěvků . (36.5204342781)

Suverénním vítězem v zájmu čtenářů se stal Domek za 700 tisíc i s nábytkem navrhl český architekt . (21.9569987727)

Freedomek an Václavském náměstí zůstane do konce akce Architecture Week , která končí 17. 10. 2010 Na druhém místě skončil opět článek se stavební tematikou . (35.1580462714)

Dům je zasazený do svahu - vizualizace Na třetím místě ve čtenosti (a nejčtenější z Návštěv) skončil článek o bydlení bezdomovců . (31.0557972559)

V pouštním městě pro ně není dostatek přístřeší , a tak nezřídka nacházejí úkryt v tunelech pod městem . (23.6951108075)

Ve svém domě mají vlastní elektrárnu Štítové stěny jsou obezděny kamenem pocházejícím z větší části ze starého stavení , které manželé koupili jako ruinu . (34.7581624662)

3. Herec Dušek vlastníma rukama postavil hliněný domek V Jindřichovicích pod Smrkem jsou dveře otevřené všem odvážným zájemcům o vyzkoušení přebývání v hliněném domku v zimě Ke stavbě hliněného domku stačí pár kamarádů , důkladná příprava a měsíc dovolené . (29.7456982055)

Dům s úžasným výhledem stál milion a půl Jednoduchý přízemní domek má rozlohu 80 metrů čtverečních Po dvanácti letech udělala Johana radikální řez . (33.2530494941)

Je řízený iPhonem Obývacímu prostoru s olejovanými dubovými parketami vévodí sedací souprava Lava (Cor) sestavená z více modulů , kterou doplňují stolky Ameo (Walter Knoll) . (24.6962635152)

Po návratu z dlouhodobého pobytu v zahraničí hledali Honza s Radkem místo , které by představovalo jejich skutečný domov . (25.907048229)

- **Výsledky grafové metody (reprezentované sumarizací z deseti vět)**

Článek o životě bezdomovců v tunelech Las Vegas se stal nejčtenější Návštěvou roku 2010. V závěsu boduje rodinný dům s vlastní elektrárnou , který má i nejvíce diskusních příspěvků . (0.0289855072464)

Na třetí příčku dosedl Jaroslav Dušek se svým hliněným domkem . (0.0217391304348)

Návštěvy patří mezi čtenáři mezi nejvyhledávanější , v roce 2010 je však " porazila " rubrika Stavba , a to hned dvakrát . (0.0217391304348)

Dům je zasazený do svahu - vizualizace Na třetím místě ve čtenosti (a nejčtenější z Návštěv) skončil článek o bydlení bezdomovců . (0.0317028985507)

1. Americký sen v Las Vegas : v tunelech zde přežívají stovky bezdomovců Steven a Kathryn ve své provizorní ložnici v tunelu pod Las Vegas . (0.0271739130435)

Steven kvůli své drogové závislosti přišel o práci . (0.0249094202899)

Veškerá práce se dá zvládnout ručně , materiál na zdivo si nakopete přímo na svém pozemku . (0.0258907004831)

6. Dřevostavbu , která vypadá jako staré venkovské stavení , smontovali za 14 dní Pohled ze dvora . (0.0235507246377)

Nová dřevostavba , která vyrostla na návsi malé pošumavské vesničky , je moderní od sklepa až na půdu . (0.0273060084541)

Přitom na první pohled vypadá jako tradiční venkovské stavení , kterých je v okolí plno . (0.021923120471)

7.4 Porovnávání a ohodnocování výsledků obou metod

Ohodnocení výsledků a následné porovnání je provedeno subjektivně. Tedy můžeme říci, že ohodnocení záleží na tom, kdo provádí ohodnocení těchto metod. Proto také bylo vybráno pouze pět článků, aby subjektivní ohodnocení nebylo tak náročné.

7.4.1 Ohodnocení a porovnání článku č.1

Z Luhnovy metody dostáváme sumarizace tvořené extrakcí pěti a deseti vět z originálního textu. Sumarizace z deseti vět dává větší smysl nežli sumarizace z pěti vět, jež je podle mého názoru nedostačující.

Při použití LexRank metody dostáváme srozumitelnou sumarizaci již při extrakci pěti vět, jež je dostačující.

Lze říci, že pro tento článek je lepší použít grafovou metodu.

7.4.2 Ohodnocení a porovnání článku č.2

Luhnova metoda při extrakci pěti vět z originálu nedává velký smysl. Při extrakci deseti vět lze říci, že sumarizace je vcelku srozumitelná, ale nedostačující.

LexRank metoda nedosahuje dobrých výsledků ani při extrakci pěti ani deseti vět z originálu, výsledná sumarizace není příliš srozumitelná a není v tomto případě moc vhodná.

Pro tento typ článku je lepší použít Luhnovu metodu.

7.4.3 Ohodnocení a porovnání článku č.3

Luhnova metoda v případě extrakce pěti vět je nedostačující, není zachycená základní podstata článku. Při extrakci deseti vět je srozumitelná a pro tento článek dostačující.

LexRank metoda je dostačující už při extrakci pěti vět. V případě extrakce deseti vět není tak kvalitní jako Luhnova metoda.

Výběr metody pro kvalitnější sumarizaci u tohoto článku závisí na délce výsledné sumarizace.

7.4.4 Ohodnocení a porovnání článku č.4

V tomto sportovním článku dosahovala Luhnova metoda lepších výsledků v případě extrakce deseti vět. V kratším verzi sumarizace není dostatečně přesná, což se projevuje v částečné nesrozumitelnosti po sobě jdoucích vět.

Při extrakci pěti vět za použití LexRank metody je sumarizace nesrozumitelná a málo kvalitní. V případě extrakce deseti vět dostáváme kvalitní sumarizaci, jež je v tomto případě tvořena hlavně výsledky, jenž jsou podstatou tohoto článku.

Tento článek má téma sport, tedy obsahuje také věty udávající výsledky, jež jsou hlavní podstatou těchto článků. Sumarizace pomocí LexRank metody extrahuje především tyto výsledky, ovšem podle mého názoru je lepší metoda Luhnova, jež extrahuje také věty částečně nahrazující tyto detailní výsledné hodnoty uvedené v článku.

7.4.5 Ohodnocení a porovnání článku č.5

Luhnova metoda nepodává v tomto článku kvalitní sumarizace v případě extrakce pěti ani deseti vět. Nesrozumitelné a nedostačující.

Ani metoda LexRank není kvalitní při extrahování pěti nebo deseti vět z tohoto článku. Též nesrozumitelné a nedostačující.

Špatné výsledky obou metod jsou u tohoto článku způsobeny délkou (počtem vět). Zkvalitnění výsledné sumarizace by mohlo být provedeno extrakcí většího počtu vět.

7.4.6 Závěr

Každý článek je originální a výsledné sumarizace tomu odpovídají. U některých článků dostáváme výsledky lepší než u jiných. Nelze jednoznačně říci, jaká z těchto dvou metod je lepší. Výsledky závisí na spoustě faktorů, jakožto délka originálního článku, délka výsledné sumarizace, hlavní téma článku, délka vět a jistě bychom mohli uvést i další. Lze konstatovat, že při rozumné délce článku se lépe osvědčila LexRank metoda již při extrahování pěti vět. Luhnova metoda byla povětšinou dostačující až při výším počtu extrahovaných vět. Nejlepší výsledky dosáhly obě metody v článku č.3, jenž dostatečně nahrazují.

Pro dosažení lepších výsledků by bylo dobré zahrnout do obou metod algoritmus, jenž by podle počtu vět originálního textu stanovil vhodnou délku výsledné sumarizace. Tento návrh ztěžuje skutečnost, že sumarizace by měla být co nejkratší. Tudíž před použitím tohoto algoritmu v praxi by bylo nutné důkladně vyzkoušet vhodné délky pro obě metody.

Kapitola 8

Závěr

V této práci je popsán pojem sumarizace, následně je také uvedená definice sumarizace. Velká část je zaměřená na metody automatické sumarizace, jejich dělení a příklady těchto metod. Věnuje se také sumarizaci novinových článků a metodám, jenž se používají při tomto druhu sumarizace.

Jsou zde uvedeny hodnotící metody, jejich dělení a příklady těchto metod.

V další části se věnujeme hlavně statistickým a grafovým metodám automatické sumarizace. Ze statistických metod byla vybrána Luhnova metoda a aplikována. Z grafových metod byla vybrána metoda LexRank. Aplikace obou těchto metod byla provedena v programovacím jazyce Python.

Z aplikace těchto metod jsou uvedené v další části této práce výsledky z testovaných textů a následně jsou tyto výsledky ohodnoceny.

Je patrné, že automatická sumarizace textů je na vzestupu, protože se stále zvětšuje zájem o sumarizaci. Důležité je se zabývat v tomto směru zdokonalováním již existujících metod a vývojem nových metod, jež podle mého názoru převýší již používané metody. Tím narážíme na abstraktní sumarizaci, jež není v této době rozšířená (jako extrakční sumarizace). Metody abstraktní sumarizace mají své „mouchy“ a je důležitý další výzkum v této oblasti.

Na ohodnocování výsledků sumarizace sice metody existují, ale většina metod používá pro ohodnocení porovnávací kolekce. Jelikož tyto kolekce nejsou dostupné na fakultě, nemohlo být provedeno automatické ohodnocení výsledků.

Sumarizace textů mě zaujala a jsem rád, že jsem si v tomto směru rozšířil obzory. Tato práce je hlavně o extrakčních metodách, které se podle mého názoru nebudou rapidně vylepšovat a vyvíjet v nejbližší době. Bylo nicméně zajímavé tyto metody vyzkoušet. Kdybych měl možnost se dále zaměřit na automatickou sumarizaci, pravděpodobně bych se v další práci věnoval spíše metodám abstraktním, které mají velký potenciál a jejich vývoj je na vzestupu.

Literatura

[1] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed),
The Ox-ford Handbook of Computational Linguistics
chapter 32, pages 583-598.Oxford University Press, 2005.

[2] Mani, I., House, D., Klein, G., et al .
The TIPSTER SUMMAC Text Summarization Evaluation.
In Proceedings of EACL, 1999.

[3] Karel JEŽEK , Josef STEINBERGER
Sumarizace textů
Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni , Univerzitní 8, 306 14 Plzeň
European Commission Joint Research Centre, IPSC Ispra T.P. 267, 21027 Ispra (VA), Italy

[4] Shannon Johnson
How to Summarize a Newspaper Article
eHow Contributor, 2011

[5] Chia-Wei Wu and Chao-Lin Liu
Ontology-based Text Summarization for Business News Articles
Department of Computer Science, National Chengchi University
Taipei 11605, Taiwan
http://www.cs.nccu.edu.tw/~chaolin/papers/wu03.pdf

[6] T. Gruber,
Ontology Definition
www-ksl.Stanford.edu/kst/what-is-an-ontology.htm

[7] Masaharu Yoshioka, Makoto Haraguchi
Multiple News Articles Summarization Based on Event Reference Information,
Graduate School of Information Science and Technology, Hokkaido University
N14 W9, Kita-ku, Sapporo-shi, Hokkaido, JAPAN
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-YoshiokaM.pdf

[8] T. Kudo and Y. Matsumoto.
Japanese dependency analysis using cascaded chunking.
In CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COOLING 2002 Post-Conference Workshops).

[9] Megumi Makino and Kazuhide Yamamoto
Summarization by Analogy: An Example-based Approach for News Articles
Dept. of Electrical Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata 940-2188 Japan

[10] H. Gregory Silber, Kathleen F. McCoy
Efficient Text Summarization Using Lexical Chains
Computer and Information Sciences, University of Delaware
Newark, DE 19711
http://web.media.mit.edu/~lieber/IUI/Silber/Silber.pdf

[11] J. Steinberger, M. Křišťan
LSA-Based Multi-Document Summarization
Text Mining Group, Dept. of Computer Science and Engineering,
University of West Bohemia in Plzeň, Czech Republic

[12] Dou Shen, Qiang Yang, Zheng Chen
Noise reduction through summarization for Web-page classification
Department of Computer Science and Technology, Hong Kong University of Science and Technology, Hong Kong, PR China
Microsoft Research Asia, Beijing, PR China
Received 17 July 2006; received in revised form 4 January 2007; accepted 8 January 2007, Available online 26 March 2007

[13] Modeling Workshop Project 2002
Graphical Methods-Summary
Unit I Reading - GM summary v2.0
http://modeling.asu.edu/Modeling-pub/Mechanics_curriculum/1-Sci%20Thinking/Resources/U1-GMsummary.pdf

[14] Karel Ježek, Josef Steinberger
Automatic Text Summarization
(The state of the art 2007 and new challenges)
Katedra informatiky a výpočetní techniky, FAV,
ZČU - Západočeská Univerzita v Plzni, Univerzitní 22, 306 14 Plzeň
http://znalosti2008.fiit.stuba.sk/download/articles/znalosti2008-Jezek.pdf

[15] Rada Mihalcea
Graph-based Ranking Algorithms for Sentence Extraction,
Applied to Text Summarization
Department of Computer Science
University of North Texas
http://acl.ldc.upenn.edu/P/P04/P04-3020.pdf

[16] Günes Erkan , Dragomir R. Radev
LexRank: Graph-based Lexical Centrality as Salience in Text Summarization
Department of EECS, University of Michigan, Ann Arbor, MI 48109 USA
School of Information & Department of EECS, University of Michigan, Ann Arbor, MI 48109 USA
http://www.aaai.org/Papers/JAIR/Vol22/JAIR-2214.pdf

[17] Josef Steinberger, Karel Ježek
EVALUATION MEASURES FOR TEXT SUMMARIZATION
Department of Computer Science and Engineering
University of West Bohemia in Pilsen
Univerzitní 8, 306 14 Plzeň, Czech Republic
Revised manuscript received 20 March 2007
http://www.cai.sk/ojs/index.php/cai/article/download/37/24

[18] Morris, A.—Kasper, G.—Adams, D.
The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. In Information Systems
Research, Vol. 3, 1992, No. 1, pp. 17–35.

[19] Dipanjan Das, André F.T. Martins
A Survey on Automatic Text Summarization
Language Technologies Institute, Carnegie Mellon University
November 21, 2007
http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf

[20] Helena Ahonen-Myka
Processing of large dokument collections
Part 5 (Text summarization)
https://www.cs.helsinki.fi/u/hahonen/lado06/material/lado06_5.pdf

Dodatky

Obsah přiloženého CD

- Elektronická podoba bakalářské práce
- Zdrojové kódy obou implementovaných metod v jazyce Python

Příloha

- Originální články

Zde jsou uvedeny další články, jež byly otestované implementovanými metodami.

o Článek č.3

Praha - Současný místopředseda Senátu Petr Pithart (KDU-ČSL) je jedním z mála exponentů sametové revoluce z listopadu 1989 , kteří se i v současnosti udrželi v nejvyšší politice . Vystudovaný právník , který 2. ledna oslaví 70. narozeniny , prošel řadou funkcí - mimo jiné byl dva roky premiérem české vlády a šest let předsedal horní parlamentní komoře . Petr Pithart se narodil v roce 1941. Po studiu práv působil jako asistent na právnické fakultě Univerzity Karlovy . V letech 1960 až 1968 byl členem komunistické strany , což mu bylo během jeho polistopadové politické kariéry čas od času připomínáno . " Dnes si troufnu říci , že už to nepokládám za svůj handicap . Já jsem to odčinil , odpykal , odtrpěl a už se k tomu nebudu vracet kajícnicky , " poznamenal Pithart v roce 2002. Poté , co z komunistické strany vystoupil , musel z vědeckého pracoviště odejít . Pracoval jako dělník a v roce 1972 se vrátil k profesi právníka - pracoval v podniku Ředitelství výstavby pracovišť ČSAV . V roce 1977 však podepsal Chartu 77 a byl propuštěn . Poté pracoval jako zahradní dělník a odborný referent . V disentu publikoval eseje , statě a politologické studie o moderních českých dějinách . Byl členem redakční rady samizdatových Lidových novin . Po listopadu 1989 se stal jedním z předních mluvčích Občanského fóra (OF) a v roce 1990 zasedl do poslaneckých lavic Federálního shromáždění ČSFR a později ČNR . V letech 1990 až 1992 předsedal české vládě a byl jedním z politiků , kteří za českou stranu neúspěšně vyjednávali o dalším uspořádání Československa . Jeho koncepce " dvojdomku " , což mělo být volnější uspořádání společného státu , však neuspěla . Po rozpadu OF v roce 1991 působil v neúspěšném Občanském hnutí , později přejmenovaném na Svobodné demokraty (SD) . Do vysoké politiky se vrátil na podzim 1996 , kdy získal senátorský mandát jako nestraník na kandidátce KDU-ČSL (v roce 1999 se pak stal členem lidové strany) a vzápětí se stal prvním předsedou obnovené horní parlamentní komory . Volbu však provázely rozpaky , neboť lidovci prosadili jeho nominaci i přes nesouhlas ostatních členů tehdejší vládní

koalice - ODS a ODA . Pithart nakonec nejtěsnějším poměrem prošel až ve druhém kole , a navíc díky tomu , že odevzdal hlas sám sobě . Jeho dny v předsednickém křesle byly prakticky sečteny již v červenci 1998 , kdy ČSSD a ODS podepsaly opoziční smlouvu , která mimo jiné deklarovala právo ODS na místa předsedů obou parlamentních komor . Pithart však zůstal ve vedení Senátu jako místopředseda s odpovědností za zahraniční styky . Svůj senátorský mandát obhájil na Chrudimsku v listopadu 2000. Strany opoziční smlouvy v těchto volbách ztratily v Senátu většinu a Pithart se tak mohl vrátit do předsednického křesla . Po čtyřech letech však získala v Senátu rozhodující slovo ODS a Pithart se opět přesunul na místopředsednickou pozici , kde zůstal i po letošních volbách . Mezitím stihl v roce 2006 na opět obhájit svůj mandát . V prosinci 2002 se stal kandidátem KDU-ČSL na úřad prezidenta po Václavu Havlovi . V lednové volbě postoupil až do třetího kola , v němž se utkal s kandidátem ODS Václavem Klausem . Klause ve třetím kole podpořilo 113 senátorů a poslanců , pro Pitharta hlasovalo o 29 zákonodárců méně . Zvolen tedy nebyl nikdo , v dalších volbách pak Pithart již nekandidoval a na Hrad byl nakonec zvolen právě Václav Klaus . Začátkem roku 2001 na sebe upozornil cestou na Kubu , kde jednal s Fidelem Castrem o propuštění tehdejšího poslance Ivana Pilipa (US-DEU) a bývalého studentského aktivisty Jana Bubeníka , které zadržely kubánské úřady kvůli stykům s disidenty . Pithart se vrátil bez obou zadržených , propuštěni ale byli dva dny poté . Petr Pithart je od roku 1964 ženatý , s manželkou Drahomírou mají syna Davida a dceru Kláru .

o Článek č.4

Buffalo - Hráči Švédska zdolali na mistrovství světa hokejistů do 20 let v duelu dvou favoritů celého turnaje v základní skupině B Kanadu 6 : 5 po samostatných nájezdech . Seveřané si tak zajistili přímý postup do semifinále šampionátu , kam postupují vítězové základních skupin . Ve skupině A utrpělo Slovensko debakl s Finy 0 : 6. Kanaďané sice začali lépe a šli již po 58 vteřinách do vedení zásluhou gólu Couturiera , ale Švédové v přesilovce rychle vyrovnali a na konci 15. minuty je Klingberg poslal do vedení . Ovšem za dalších 43 sekund bylo po Howdenově trefě vyrovnáno na 2 : 2 a pouhou vteřinu před vypršením hracího času úvodní třetiny vrátil vedení zpět na stranu Kanady Hamilton . Skóre se však přelévalo ze strany na stranu i nadále . Do 23. minuty zajistili další obrat Klingberg s Thörnbergem . Jenže necelé dvě minuty poté napodobil Klingberga s druhým zápisem do střelecké listiny také Kanaďan Hamilton a při vlastním oslabení vyrovnal na 4 : 4. Ve třetím dějství souboje stříbrného mužstva (Kanada) s bronzovým týmem (Švédsko) z minulého šampionátu šel znovu do vedení zámořský výběr , když se ve 44. minutě trefil v početní výhodě Schenn . V 52. minutě ovšem srovnal Cehlin a zápas díky tomu dospěl nejen k prodloužení , ale nakonec i k samostatným nájezdům . V nich se ze tří švédských exekutorů trefili Lindberg a Lander , naproti tomu Kanaďané Ellis a Schenn neuspěli . Odměnou za lépe zvládnutou bitvu před téměř 18 tisíci diváků je pro Švédy přímý postup do semifinále . " Ještě nikdy jsem dosud nevyhrál zápas proti Kanadě , " hlásil po utkání šťastný švédský forvard Calle Järnkrok pro webové stránky IIHF . " Nyní jsme vyhráli skupinu , máme výhodu dalšího dne na trénink a budeme hrát o jeden zápas

méně , " dodal Järnkrok . V utkání skupiny A vyprovodili Finové juniory Slovenska do skupiny o udržení debaklem 6 : 0 , přičemž tento výsledek zářil na světelné tabuli už po druhé třetině . Tři branky vsítil tým Suomi v přesilovkách , jednou uspěl také v oslabení . MS hokejistů do dvaceti let v Buffalu a Lewistonu (USA) : Skupina A : Slovensko - Finsko 0 : 6 (0 : 3 , 0 : 3 , 0 : 0) Branky a nahrávky : 2. Salomäki (Junttila) , 7. Jokipakka (Rajala , Virtanen) , 11. Haula (Vatanen , Pulkkinen) , 27. Donskoi (Pulkkinen , Armia) , 29. Haula , 38. Virkkunen (Tallberg , Turtiainen) . Rozhodčí : Jablukov (Něm .) , Kaval - Morrison (oba USA) , Kaspar (Rak .) . Vyloučení : 4 : 6. Využití : 0 : 3. V oslabení : 0 : 1. Diváci : 13371. Skupina B : Kanada - Švédsko 5 : 6 po sam . nájezdech (3 : 2 , 1 : 2 , 1 : 1 - 0 : 0) Branky a nahrávky : 1. Couturier , 16. Howden (Johansen , de Haan) , 20. Hamilton (Johansen) , 25. Hamilton (Schenn , Després) , 44. Schenn (Johansen , de Haan) - 3. Friberg (Rakell) , 15. Klingberg (Wännström) , 21. Klingberg , 23. Thörnberg (Nemeth) , 52. Cehlin (Erixon , Styrman) , rozhodující sam . nájezd Lindberg . Rozhodčí : Kadyrov (Rus .) , Fraňo - Bláha (oba ČR) , Brown (USA) . Vyloučení : 4 : 3. Využití : 1 : 1. V oslabení : 1 : 0. Diváci : 17761. .

o Článek č.5

Liberec - V Libereckém kraji se letos první dítě nového roku narodilo v Liberci . Malý Matyáš přišel na svět pouhých 43 sekund po půlnoci , a je tak zřejmě prvním dítětem narozeným na Nový rok v Česku . Chlapeček měří 49 centimetrů a váží 2720 gramů . Jeho matka je z Hrádku nad Nisou , řekla dnes ČTK mluvčí Krajské nemocnice v Liberci Alexandra Kittnerová . " Je to už čtvrté dítě , porod byl opravdu rychlý , během dvaceti minut byl Matyáš na světě , " řekla ČTK šťastná maminka Eva Mottlová . Má už dva dospělé syny a dvanáctiletou dceru . " Ta se na miminko moc těší , " dodala žena . Jen jí bylo trochu líto , že vzhledem k porodu po půlnoci přišla o porodné . " Ty peníze by se nám určitě hodily , ale co se dá dělat , hlavně , že je syn zdravý a má se k světu , " dodala . O porodné připravila ženu novela zákona o sociální podpoře , která od Nového roku podmínky k jeho přiznávání výrazně změnila . Zatímco ještě na Silvestra na něj měly nárok všechny ženy , které porodily dítě , od Nového roku ho dostávají pouze prvorodičky . Pro nárok se dále bude posuzovat příjem rodiny , který nesmí přesáhnout 2 , 4násobek životního minima . Zároveň se mění i podmínky pro rodičovské příspěvky . Liberecká nemocnice letos konečně porazila porodnici v Jablonci , kde měli první miminka v posledních dvou letech . Vždy to byly holčičky a v obou případech dostaly jméno Eliška . I letos se tam jako první narodila holčička . Přišla na svět dvě minuty po půlnoci , na prvenství to ale nestačilo . " Holčička dostala jméno Sofie , měří 48 centimetrů a váží 3050 gramů , " řekla mluvčí tamní nemocnice Petra Krajinová . Královéhradecký kraj Prvním miminkem letošního roku v hradeckém kraji je Vanesa Žočková z Křinic na Náchodsku . Narodila se dnes v 01 : 25 v nemocnici v Náchodě . Na svět přišla s váhou 3,6 kilogramu a mírou 49,5 centimetru . ČTK to řekl mluvčí kraje Imrich Dioszegi . Druhá se v kraji narodila Nikola Nováková v trutnovské nemocnici v 04 : 03. Holčička je z Trutnova . Třetím dítětem v pořadí a zároveň prvním chlapcem letošního roku je v

hradeckém kraji Jakub Chmelík z Radvanic u Trutnova . Narodil se v trutnovské nemocnici v 06 : 55 a vážil 3,5 kilogramu a měřil 50 centimetrů . Prvním dítětem ve stotisícovém Hradci Králové se stala Karolína Černá . Narodila se v hradecké fakultní nemocnici v 05 : 17. Vážila téměř čtyři kilogramy a měřila 53 centimetrů . První děti roku dostanou od kraje i od některých radnic dárky . Například krajský úřad dá rodičům prvního chlapce a děvčete narozeného v krajské nemocnici dárkový poukaz v hodnotě 5000 korun . Letos dar dostanou rodiče Vanesy a Jakuba . Vedení hradecké radnice stejně jako v předchozích letech věnuje rodičům novorozence pět tisíc korun . Prvním novorozencem loňského roku se v hradeckém kraji stal Mário Matuš , který se v Oblastní nemocnici v Trutnově narodil na Nový rok v 04 : 59 rodičům z Libče u Trutnova . Pardubický kraj Prvním miminkem letošního roku v Pardubickém kraji je Richard Drahoš z Holic na Pardubicku . Narodil se v 08 : 30 v nemocnici v Pardubicích . Vážil 3,78 kilogramu a měřil 52 centimetrů . Chlapec i maminka jsou v pořádku . ČTK to řekla mluvčí nemocnice Alice Štrajtová Štefková . Jiné děti se do dnešního dopoledne v nemocnicích v Pardubickém kraji nenarodily . První letos narozené děti dostanou od pardubického hejtmanství a některých radnic dárky . První děvčátko a první chlapeček narození v roce 2011 v Pardubickém kraji obdrží od hejtmanství 5000 korun a drobný dárek , řekla ČTK mluvčí kraje Magdalena Navrátilová . Prvním novorozencem loňského roku se v Pardubickém kraji stal Jindřich Sádovský z Filipova u Hlinska . Narodil se 1. ledna 2010 v 05 : 55 v nemocnici v Chrudimi . Olomoucký kraj Prvním dítětem , které letos přišlo na svět v Olomouckém kraji , je holčička jménem Terezka . Narodila se v 4 : 38 v porodnici v Šumperku . Vážila 2,61 kilogramu a měřila 47 centimetrů . Maminka i dítě jsou v pořádku , řekla ČTK Hana Szotkowská , mluvčí skupiny Agel , pod kterou nemocnice spadá . První letošní děti už hlásí i další zdravotnická zařízení v Olomouckém kraji . Ve Šternberku se jako první narodil chlapec Jaroslav . Na svět přišel v 5 : 32 , vážil 3,4 kilogramu a měřil 50 centimetrů . V olomoucké porodnici se jako první narodilo děvčátko . Na svět přišlo v 6 : 28. V Přerově , Prostějově a v Jeseníku zatím na první letošní miminko stále čekají . V šumperské nemocnici se narodilo první dítě Olomouckého kraje i v loňském roce , tehdy ale přišlo na svět mnohem dříve . Chlapec jménem Martin se totiž narodil už 11 minut po půlnoci . Moravskoslezský kraj Prvním miminkem narozeným letos v Moravskoslezském kraji je Tobiáš z Ostravy . Na svět přišel ve Vítkovické nemocnici hodinu po půlnocí . " Chlapeček se narodil 0 : 58 s ukázkovými mírami . Váží 3,7 kilogramu a měří 52 centimetrů , " řekla ČTK mluvčí nemocnice Hana Szotkowská . Druhým letošním miminkem je opět chlapeček , kterému na svět pomohli lékaři Slezské nemocnice v Opavě . " Naše první miminko se narodilo ve 03 : 40 , " řekl mluvčí zdravotnického zařízení Daniel Svoboda . Dítě měří stejně jako Tobiáš a váží 3,75 kilogramu . První novorozeně už mají i v Havířově a Novém Jičíně . Havířovští lékaři přivedli na svět 04 : 20 Borise . Jejich kolegům v Novém Jičíně se šest minut před 07 : 00 narodila malá Eliška . Na svět přišla císařským řezem . Jednalo se o předčasný porod . Holčička měří 44 centimetrů a váží dva kilogramy . V dalších šesti porodnicích Moravskoslezského kraje se k 10 : 00 miminka nenarodila . Ústecký kraj Jako první dítě roku 2011 v Ústeckém kraji se narodil dnes ráno v 02 : 57 chlapec v porodnici v Chomutově . Váží 2000 gramů a měří 45 centimetrů . ČTK to dnes řekl mluvčí Krajské zdravotní , která nemocnici v Chomutově provozuje , Jiří Vondra . Do 10 : 00 se v pěti

porodnicích Krajské zdravotní narodilo celkem šest dětí . V menších nemocnicích v Ústeckém kraji zatím na první miminko čekají . V porodnicích v Rumburku , Kadani , Litoměřicích i Žatci budou mít první letošní porod většinou asi až odpoledne . Vloni se první dítě v Ústeckém kraji narodilo v děčínské porodnici . Deset minut po půlnoci tam přivítaly zdravotní sestry na světě holčičku vážící 2,8 kilogramu a měřící 48 centimetrů . V roce 2009 se jako první narodila minutu po půlnoci holčička v Rumburku , stala se tak prvním narozeným dítětem v celém Česku . Vážila 3,46 kilogramu a měřila 52 centimetrů . Od Ústeckého kraje tehdy dostala šek na 20000 korun . Za první tři čtvrtletí roku 2010 se v Ústeckém kraji podle údajů Českého statistického úřadu narodilo 6961 dětí , údaje za poslední čtvrtletí minulého roku ještě nejsou k dispozici . Nejvíce dětí se od ledna do září narodilo na Děčínsku (1115) , Teplicku (1076) a Chomutovsku (1057) . V roce 2009 přišlo na svět v Ústeckém kraji 9626 dětí . Praha Prvním pražským miminkem letošního roku je chlapeček Vojta , který se devět minut po půlnoci narodil ve Všeobecné fakultní nemocnici (VFN) na Karlově náměstí . " Chlapeček je v pořádku a je zdravý , " řekla ČTK sloužící sestra ve VFN . ČTK má údaje ze všech šesti pražských porodnic . V motolské nemocnici se první miminko chystalo na svět až po první hodině ranní , v Podolí se narodilo zhruba ve tři hodiny ráno . V dalších nemocnicích se k porodu připravovali až v dalších hodinách , například na Vinohradech se rodilo až kolem sedmé hodiny . Podle oslovených pražských nemocnic ale kvůli tomu žádné porody nebyly urychleny . " Předčasné vyvolávání porodu je nebezpečné , " uvedla mluvčí motolské nemocnice Eva Jurinová . Pražské maminky , které porodí na Nový rok , dostanou stejně jako v minulých letech finanční dar od primátora hlavního města . Loni činil 15000 korun . V únoru pak budou také pozvány na slavnostní přivítání do primátorské rezidence . Jihočeský kraj Prvním miminkem narozeným v roce 2011 na jihu Čech se je Jolanka Smolíková z Oldřichova u Písku . Narodila se v 02 : 32 v porodnici písecké nemocnice . Váží 2,90 kilogramu a měří 50 centimetrů . Ze zbývajících šesti jihočeských porodnic přivítali do rána přírůstek jen v Jindřichově Hradci . Holčička Zuzanka se tam narodila v 04 : 55. Nezvyklý klid přičítají porodníci tomu , že se ode dneska omezuje vyplácení porodného . První letošní Jihočeška Jolanka už má doma staršího bratříčka . To , že rodina přijde o porodné , maminku Janu Smolíkovou moc nemrzí . " Už jsme v životě přišli o víc , " poznamenala . Přesto by byla raději , kdyby holčička přišla na svět už na Silvestra . " Manžel má zrovna na Silvestra narozeniny , " vysvětlila . Rodiče prvního občánka se v nejbližších dnech dočkají gratulace i z hejtmanství . " Vítání do rodiny Jihočechů probíhá až po nějakém čase , přibližně po šesti nedělích , " uvedla krajská mluvčí Kateřina Koželuhová . Setkání s hejtmanem se uskutečňují většinou v místě , kde rodina žije ; po dohodě s rodiči to může být v obřadní síni , ale i v domácnosti . Mezi dárky bývá i peněžní poukázka v hodnotě několika tisíc korun . Prvním loňským občánkem Jihočeského kraje byl Štěpán Fousek z Českého Krumlova . Na svět přišel pět minut po půlnoci v českokrumlovské porodnici . Jihomoravský kraj Prvním novorozencem letošního roku se na jihu Moravy stal David Klíma . Žena z Říčan u Brna ho porodila v bohunické porodnici Fakultní nemocnice Brno . Chlapec vážil 3,08 kilogramu a měřil 52 centimetrů , řekl ČTK vedoucí odboru zdravotnictví na krajském úřadu Josef Drbal . Do 08 : 00 přišlo v kraji na svět dalších sedm dětí . Na prvního Brňana město zatím čeká , dodal Drbal . David Klíma jako první miminko narozené v novém roce dostane dary od

kraje . Jejich letošní podobu budou teprve radní schvalovat . Loni obdrželi rodiče prvního dítěte 10000 korun a elektronickou chůvičku . Matka dostala kytici , otec balení jihomoravských vín . Loni kraj ocenil i první dítě narozené v každé z jihomoravských porodnic . Zda se letos v novince bude pokračovat , bude záležet na rozhodnutí radních . Novoroční děti se letos zatím narodily v 01 : 20 v Břeclavi , ve 02 : 23 v Kyjově , ve 02 : 35 v Boskovicích , ve 03 : 04 ve Znojmě , v brněnské Nemocnici Milosrdných bratří ve 04 : 42 a ve Vyškově v 07 : 04. Kraj Vysočina Prvním občánkem Vysočiny roku 2011 je Simonka Svobodová z Vladislavi . Narodila se v 02 : 03 v porodnici třebíčské nemocnice . Vážila 3,65 kilogramu a měřila 50 centimetrů . Holčička i její maminka jsou podle zdravotníků v pořádku . Do rána přibyli v kraji ještě dva chlapci . V Havlíčkově Brodě přišel na svět krátce po 04 : 00 Jakub a asi o půl hodiny později přivítali v pelhřimovské porodnici Lukáška . V porodnicích nemocnic v Jihlavě a v Novém Městě na Moravě se do 08 : 00 žádný porod nekonal . Prvního občánka kraje Vysočina navštíví hned v pondělí hejtman Jiří Běhounek . Stejně jako v předcházejících letech daruje miminku zlatý šperk v podobě znamení zvěrokruhu a jeho rodičům šek na deset tisíc korun . Vedení kraje dodržuje tradici vítání prvního občánka od roku 2002 , uvedl mluvčí kraje Jan Nechvátal . V havlíčkobrodské nemocnici se tehdy vteřinu po půlnoci narodil Vojtěch Ostatnický . Byl prvním miminkem Vysočiny i celého Česka . " Prvenství v rámci republiky jsme si na Vysočině , podle krajského archivu , připsali ještě v roce 2004 , kdy se na Nový rok v Nemocnici Nové Město na Moravě šťastným rodičům narodil Vojtíšek Pavliš , " připomněl mluvčí . Před rokem se na Vysočině jako první narodilo miminko z Pardubického kraje . Elen Mocková z Chocně přišla na svět ve 03 : 40 v havlíčkobrodské nemocnici . Prvním občánkem kraje roku 2010 se pak stala Ema Schadová z Třebíče , která se narodila v 05 : 05 rovněž v Havlíčkově Brodě . Plzeňský kraj Prvním miminkem v Plzeňském kraji narozeným v roce 2011 je Terezka Vilímová z Mlázov u Klatov . Na svět přišla v klatovské nemocnici dnes ráno v 05 : 09. Holčička váží 2,75 kilogramu a měří 48 centimetrů . Je zatím jediným miminkem v regionu , které se do dnešních 07 : 30 narodilo . Z ostatních pěti nemocnic v kraji hlásí první den v roce 2011 zatím klid , zjistila ČTK . " Terezka se narodila přirozenou cestou , " uvedla sestra gynekologicko-porodnického oddělení nemocnice v Klatovech . Děvčátko i maminka jsou v pořádku . Podle sestry byl konec roku klidný , naopak velký " fofr " zaznamenala porodnice před Vánoci , kdy se narodilo hodně dětí císařským řezem . Další miminka by se měla dnes narodit v nemocnici v Rokycanech , kde ráno na porod čekaly dvě ženy . Poslední dítě roku 2010 tam přišlo na svět 31. prosince v 17 : 40. " Jednalo se o vyvolávaný porod , " uvedla sestra . Oproti roku 2009 , kdy se v Rokycanech narodilo kolem 520 dětí , byl loňský rok slabší . Nemocnice eviduje kolem 420 novorozeňat . Na první letošní miminko čekají i další zařízení jako nemocnice v Domažlicích , ve Stodě a také obě nemocnice v Plzni , tedy Mulačova i fakultní . " Hodně porodů jsme měli před Vánoci , maminky se možná víc snažily , aby byly na svátky doma . Nyní máme klid , " uvedla sestra porodnice v Domažlicích , kde za loňský rok hlásí kolem 700 novorozeňat . Celkem 634 porodů eviduje za uplynulý rok plzeňská Mulačova nemocnice . Číslo je ve srovnání s rokem 2009 slabší . Naopak více dětí se narodilo vloni ve Stodě na jihu Plzeňska , kolem 430. Poslední tam přišlo na svět 31. prosince v 05 : 15. Celkem 3396 dětí přivedli na svět v největším zařízení v kraji , na Gynekologicko-porodnické klinice Fakultní nemocnice v

Plzni , která řeší nejkomplikovanější případy z celých západních Čech . Zatímco například letošní listopad byl podle tamní sestry slabší , prosinec naopak na množství porodů silnější . První lednový den roku 2011 je však klid , poslední miminko se narodilo na Silvestra v 19 : 25. Plzeňský kraj obdaruje první holčičku i chlapečka v regionu částkou 10000 korun . Středočeský kraj Ve středních Čechách se letos narodilo první dítě v kolínské nemocnici . Je jím chlapeček Jakub , který přišel na svět osm minut po půlnoci . ČTK to potvrdila mluvčí nemocnice Gabriela Uriková . Bližší informace o novorozeněti zatím neměla k dispozici . Druhým letošním Středočechem je Václav , který se narodil v 0 : 30 v Rakovníku . " Je to pěkný kousek . Váží 4,2 kilogramu a měří 54 centimetrů , " uvedla pracovnice gynekologicko-porodnického oddělení . Třetí v pořadí je podle zjištění ČTK holčička Nelly , kterou přivítali v 01 : 05 v Mělníku . Po narození vážila 3,26 kilogramu a měřila 49 centimetrů . Ve 02 : 02 ji následovala Karolínka , která přišla na svět v brandýské nemocnici . O necelou hodinu později se ještě v Kladně narodil Josef . Maminka letošního prvního Středočecha dostane podobně jako loni 10000 korun . Dítě přijede podle mluvčí kraje Berill Maschekové přivítat hejtman nebo jeho náměstek , a to zřejmě v neděli kolem poledne . Loni získal pomyslný titul prvního dítěte ve středních Čechách chlapeček jménem Chasan , který přišel na svět 13 minut po půlnoci v brandýské nemocnici . Jeho maminkou byla cizinka , která má v České republice dlouhodobý pobyt .

- Výsledky

Zde jsou uvedeny výsledky zbylých testovacích článků.

o Článek č.3

- **Výsledky statistické metody (reprezentované sumarizací z pěti vět)**

V letech 1960 až 1968 byl členem komunistické strany , což mu bylo během jeho polistopadové politické kariéry čas od času připomínáno . (26.4803673556)

V letech 1990 až 1992 předsedal české vládě a byl jedním z politiků , kteří za českou stranu neúspěšně vyjednávali o dalším uspořádání Československa . (26.5033776093)

Volbu však provázely rozpaky , neboť lidovci prosadili jeho nominaci i přes nesouhlas ostatních členů tehdejší vládní koalice - ODS a ODA . (27.2591238657)

Po čtyřech letech však získala v Senátu rozhodující slovo ODS a Pithart se opět přesunul na místopředsednickou pozici , kde zůstal i po letošních volbách . (24.9286167007)

Začátkem roku 2001 na sebe upozornil cestou na Kubu , kde jednal s Fidelem Castrem o propuštění tehdejšího poslance Ivana Pilipa (US-DEU) a bývalého studentského

aktivisty Jana Bubeníka , které zadržely kubánské úřady kvůli stykům s disidenty . (24.6399641034)

- **Výsledky grafové metody (reprezentované sumarizací z pěti vět)**

Praha - Současný místopředseda Senátu Petr Pithart (KDU-ČSL) je jedním z mála exponentů sametové revoluce z listopadu 1989 , kteří se i v současnosti udrželi v nejvyšší politice . (0.0425)

Já jsem to odčinil , odpykal , odtrpěl a už se k tomu nebudu vracet kajícnicky , " poznamenal Pithart v roce 2002. Poté , co z komunistické strany vystoupil , musel z vědeckého pracoviště odejít . (0.0525740740741)

Pracoval jako dělník a v roce 1972 se vrátil k profesi právníka - pracoval v podniku Ředitelství výstavby pracovišť ČSAV . (0.0358333333333)

Do vysoké politiky se vrátil na podzim 1996 , kdy získal senátorský mandát jako nestraník na kandidátce KDU-ČSL (v roce 1999 se pak stal členem lidové strany) a vzápětí se stal prvním předsedou obnovené horní parlamentní komory . (0.0444166666667)

Svůj senátorský mandát obhájil na Chrudimsku v listopadu 2000. Strany opoziční smlouvy v těchto volbách ztratily v Senátu většinu a Pithart se tak mohl vrátit do předsednického křesla . (0.0769423611111)

- **Výsledky statistické metody (reprezentované sumarizací z deseti vět)**

V letech 1960 až 1968 byl členem komunistické strany , což mu bylo během jeho polistopadové politické kariéry čas od času připomínáno . (26.4803673556)

V letech 1990 až 1992 předsedal české vládě a byl jedním z politiků , kteří za českou stranu neúspěšně vyjednávali o dalším uspořádání Československa . (26.5033776093)

Volbu však provázely rozpaky , neboť lidovci prosadili jeho nominaci i přes nesouhlas ostatních členů tehdejší vládní koalice - ODS a ODA . (27.2591238657)

Jeho dny v předsednickém křesle byly prakticky sečteny již v červenci 1998 , kdy ČSSD a ODS podepsaly opoziční smlouvu , která mimo jiné deklarovala právo ODS na místa předsedů obou parlamentních komor . (19.6252161082)

Po čtyřech letech však získala v Senátu rozhodující slovo ODS a Pithart se opět přesunul na místopředsednickou pozici , kde zůstal i po letošních volbách . (24.9286167007)

V lednové volbě postoupil až do třetího kola , v němž se utkal s kandidátem ODS Václavem Klausem . (18.9260645167)

Klause ve třetím kole podpořilo 113 senátorů a poslanců , pro Pitharta hlasovalo o 29 zákonodárců méně . (20.1243384751)

Zvolen tedy nebyl nikdo , v dalších volbách pak Pithart již nekandidoval a na Hrad byl nakonec zvolen právě Václav Klaus . (22.4167659075)

Začátkem roku 2001 na sebe upozornil cestou na Kubu , kde jednal s Fidelem Castrem o propuštění tehdejšího poslance Ivana Pilipa (US-DEU) a bývalého studentského aktivisty Jana Bubeníka , které zadržely kubánské úřady kvůli stykům s disidenty . (24.6399641034)

Petr Pithart je od roku 1964 ženatý , s manželkou Drahomírou mají syna Davida a dceru Kláru . (18.4126714071)

- **Výsledky grafové metody (reprezentované sumarizací z deseti vět)**

Praha - Současný místopředseda Senátu Petr Pithart (KDU-ČSL) je jedním z mála exponentů sametové revoluce z listopadu 1989 , kteří se i v současnosti udrželi v nejvyšší politice . (0.0425)

Vystudovaný právník , který 2. ledna oslaví 70. narozeniny , prošel řadou funkcí - mimo jiné byl dva roky premiérem české vlády a šest let předsedal horní parlamentní komoře . (0.0333333333333)

V letech 1960 až 1968 byl členem komunistické strany , což mu bylo během jeho polistopadové politické kariéry čas od času připomínáno . (0.0333333333333)

Já jsem to odčinil , odpykal , odtrpěl a už se k tomu nebudu vracet kajícnicky , " poznamenal Pithart v roce 2002. Poté , co z komunistické strany vystoupil , musel z vědeckého pracoviště odejít . (0.0525740740741)

Pracoval jako dělník a v roce 1972 se vrátil k profesi právníka - pracoval v podniku Ředitelství výstavby pracovišť ČSAV . (0.0358333333333)

V roce 1977 však podepsal Chartu 77 a byl propuštěn . (0.0333333333333)

V disentu publikoval eseje , statě a politologické studie o moderních českých dějinách . (0.0333333333333)

Do vysoké politiky se vrátil na podzim 1996 , kdy získal senátorský mandát jako nestraník na kandidátce KDU-ČSL (v roce 1999 se pak stal členem lidové strany) a vzápětí se stal prvním předsedou obnovené horní parlamentní komory . (0.0444166666667)

Svůj senátorský mandát obhájil na Chrudimsku v listopadu 2000. Strany opoziční smlouvy v těchto volbách ztratily v Senátu většinu a Pithart se tak mohl vrátit do předsednického křesla . (0.0769423611111)

Po čtyřech letech však získala v Senátu rozhodující slovo ODS a Pithart se opět přesunul na místopředsednickou pozici , kde zůstal i po letošních volbách . (0.0346177951389)

o Článek č.4

- **Výsledky statistické metody (reprezentované sumarizací z pěti vět)**

Seveřané si tak zajistili přímý postup do semifinále šampionátu , kam postupují vítězové základních skupin . (17.5709672628)

Ve skupině A utrpělo Slovensko debakl s Finy 0 : 6. Kanaďané sice začali lépe a šli již po 58 vteřinách do vedení zásluhou gólu Couturiera , ale Švédové v přesilovce rychle vyrovnali a na konci 15. minuty je Klingberg poslal do vedení . (25.063245306)

Ovšem za dalších 43 sekund bylo po Howdenově trefě vyrovnáno na 2 : 2 a pouhou vteřinu před vypršením hracího času úvodní třetiny vrátil vedení zpět na stranu Kanady Hamilton . (18.9610161327)

Jenže necelé dvě minuty poté napodobil Klingberga s druhým zápisem do střelecké listiny také Kanaďan Hamilton a při vlastním oslabení vyrovnal na 4 : 4. Ve třetím dějství souboje stříbrného mužstva (Kanada) s bronzovým týmem (Švédsko) z minulého šampionátu šel znovu do vedení zámořský výběr , když se ve 44. minutě trefil v početní výhodě Schenn . (35.9489992689)

Odměnou za lépe zvládnutou bitvu před téměř 18 tisíci diváků je pro Švédy přímý postup do semifinále . (20.3160355957)

- **Výsledky grafové metody (reprezentované sumarizací z pěti vět)**

Jenže necelé dvě minuty poté napodobil Klingberga s druhým zápisem do střelecké listiny také Kanaďan Hamilton a při vlastním oslabení vyrovnal na 4 : 4. Ve třetím dějství souboje stříbrného mužstva (Kanada) s bronzovým týmem (Švédsko) z minulého šampionátu šel znovu do vedení zámořský výběr , když se ve 44. minutě trefil v početní výhodě Schenn . (0.048896275682)

V nich se ze tří švédských exekutorů trefili Lindberg a Lander , naproti tomu Kanaďané Ellis a Schenn neuspěli . (0.0520528413386)

MS hokejistů do dvaceti let v Buffalu a Lewistonu (USA) : Skupina A : Slovensko - Finsko 0 : 6 (0 : 3 , 0 : 3 , 0 : 0) Branky a nahrávky : 2. Salomäki (Junttila) , 7.

Jokipakka (Rajala , Virtanen) , 11. Haula (Vatanen , Pulkkinen) , 27. Donskoi (Pulkkinen , Armia) , 29. Haula , 38. Virkkunen (Tallberg , Turtiainen) . (0.0797265753243)

Vyloučení : 4 : 6. Využití : 0 : 3. V oslabení : 0 : 1. Diváci : 13371. Skupina B : Kanada - Švédsko 5 : 6 po sam . (0.0542160737489)

nájezdech (3 : 2 , 1 : 2 , 1 : 1 - 0 : 0) Branky a nahrávky : 1. Couturier , 16. Howden (Johansen , de Haan) , 20. Hamilton (Johansen) , 25. Hamilton (Schenn , Després) , 44. Schenn (Johansen , de Haan) - 3. Friberg (Rakell) , 15. Klingberg (Wännström) , 21. Klingberg , 23. Thörnberg (Nemeth) , 52. Cehlin (Erixon , Styrman) , rozhodující sam . (0.0927678801927)

- **Výsledky statistické metody (reprezentované sumarizací z deseti vět)**

Buffalo - Hráči Švédska zdolali na mistrovství světa hokejistů do 20 let v duelu dvou favoritů celého turnaje v základní skupině B Kanadu 6 : 5 po samostatných nájezdech . (16.487627057)

Seveřané si tak zajistili přímý postup do semifinále šampionátu , kam postupují vítězové základních skupin . (17.5709672628)

Ve skupině A utrpělo Slovensko debakl s Finy 0 : 6. Kanaďané sice začali lépe a šli již po 58 vteřinách do vedení zásluhou gólu Couturiera , ale Švédové v přesilovce rychle vyrovnali a na konci 15. minuty je Klingberg poslal do vedení . (25.063245306)

Ovšem za dalších 43 sekund bylo po Howdenově trefě vyrovnáno na 2 : 2 a pouhou vteřinu před vypršením hracího času úvodní třetiny vrátil vedení zpět na stranu Kanady Hamilton . (18.9610161327)

Jenže necelé dvě minuty poté napodobil Klingberga s druhým zápisem do střelecké listiny také Kanaďan Hamilton a při vlastním oslabení vyrovnal na 4 : 4. Ve třetím dějství souboje stříbrného mužstva (Kanada) s bronzovým týmem (Švédsko) z minulého šampionátu šel znovu do vedení zámořský výběr , když se ve 44. minutě trefil v početní výhodě Schenn . (35.9489992689)

Odměnou za lépe zvládnutou bitvu před téměř 18 tisíci diváků je pro Švédy přímý postup do semifinále . (20.3160355957)

" Ještě nikdy jsem dosud nevyhrál zápas proti Kanadě , " hlásil po utkání šťastný švédský forvard Calle Järnkrok pro webové stránky IIHF . (13.1369411038)

V utkání skupiny A vyprovodili Finové juniory Slovenska do skupiny o udržení debaklem 6 : 0 , přičemž tento výsledek zářil na světelné tabuli už po druhé třetině . (15.832103496)

MS hokejistů do dvaceti let v Buffalu a Lewistonu (USA) : Skupina A : Slovensko - Finsko 0 : 6 (0 : 3 , 0 : 3 , 0 : 0) Branky a nahrávky : 2. Salomäki (Junttila) , 7. Jokipakka (Rajala , Virtanen) , 11. Haula (Vatanen , Pulkkinen) , 27. Donskoi (Pulkkinen , Armia) , 29. Haula , 38. Virkkunen (Tallberg , Turtiainen) . (17.4238242531)

nájezdech (3 : 2 , 1 : 2 , 1 : 1 - 0 : 0) Branky a nahrávky : 1. Couturier , 16. Howden (Johansen , de Haan) , 20. Hamilton (Johansen) , 25. Hamilton (Schenn , Després) , 44. Schenn (Johansen , de Haan) - 3. Friberg (Rakell) , 15. Klingberg (Wännström) , 21. Klingberg , 23. Thörnberg (Nemeth) , 52. Cehlin (Erixon , Styrman) , rozhodující sam . (16.3110257192)

- **Výsledky grafové metody (reprezentované sumarizací z deseti vět)**

Jenže necelé dvě minuty poté napodobil Klingberga s druhým zápisem do střelecké listiny také Kanaďan Hamilton a při vlastním oslabení vyrovnal na 4 : 4. Ve třetím dějství souboje stříbrného mužstva (Kanada) s bronzovým týmem (Švédsko) z minulého šampionátu šel znovu do vedení zámořský výběr , když se ve 44. minutě trefil v početní výhodě Schenn . (0.048896275682)

V 52. minutě ovšem srovnal Cehlin a zápas díky tomu dospěl nejen k prodloužení , ale nakonec i k samostatným nájezdům . (0.0416666666667)

V nich se ze tří švédských exekutorů trefili Lindberg a Lander , naproti tomu Kanaďané Ellis a Schenn neuspěli . (0.0520528413386)

MS hokejistů do dvaceti let v Buffalu a Lewistonu (USA) : Skupina A : Slovensko - Finsko 0 : 6 (0 : 3 , 0 : 3 , 0 : 0) Branky a nahrávky : 2. Salomäki (Junttila) , 7. Jokipakka (Rajala , Virtanen) , 11. Haula (Vatanen , Pulkkinen) , 27. Donskoi (Pulkkinen , Armia) , 29. Haula , 38. Virkkunen (Tallberg , Turtiainen) . (0.0797265753243)

) , Kaval - Morrison (oba USA) , Kaspar (Rak . (0.0405678998328)

Vyloučení : 4 : 6. Využití : 0 : 3. V oslabení : 0 : 1. Diváci : 13371. Skupina B : Kanada - Švédsko 5 : 6 po sam . (0.0542160737489)

nájezdech (3 : 2 , 1 : 2 , 1 : 1 - 0 : 0) Branky a nahrávky : 1. Couturier , 16. Howden (Johansen , de Haan) , 20. Hamilton (Johansen) , 25. Hamilton (Schenn , Després) , 44. Schenn (Johansen , de Haan) - 3. Friberg (Rakell) , 15. Klingberg (Wännström) , 21. Klingberg , 23. Thörnberg (Nemeth) , 52. Cehlin (Erixon , Styrman) , rozhodující sam . (0.0927678801927)

Rozhodčí : Kadyrov (Rus . (0.0408053794661)

) , Fraňo - Bláha (oba ČR) , Brown (USA) . (0.0416682672237)

Vyloučení : 4 : 3. Využití : 1 : 1. V oslabení : 1 : 0. Diváci : 17761. . (0.0480059441358)

o Článek č.5

- **Výsledky statistické metody (reprezentované sumarizací z pěti vět)**

O porodné připravila ženu novela zákona o sociální podpoře , která od Nového roku podmínky k jeho přiznávání výrazně změnila . (36.005152626)

Praha Prvním pražským miminkem letošního roku je chlapeček Vojta , který se devět minut po půlnoci narodil ve Všeobecné fakultní nemocnici (VFN) na Karlově náměstí . (39.9392309873)

V Havlíčkově Brodě přišel na svět krátce po 04 : 00 Jakub a asi o půl hodiny později přivítali v pelhřimovské porodnici Lukáška . (33.7818981748)

Dítě přijede podle mluvčí kraje Berill Maschekové přivítat hejtman nebo jeho náměstek , a to zřejmě v neděli kolem poledne . (32.886421317)

Loni získal pomyslný titul prvního dítěte ve středních Čechách chlapeček jménem Chasan , který přišel na svět 13 minut po půlnoci v brandýské nemocnici . (35.8579511102)

- **Výsledky grafové metody (reprezentované sumarizací z pěti vět)**

" Ty peníze by se nám určitě hodily , ale co se dá dělat , hlavně , že je syn zdravý a má se k světu , " dodala . (0.016310107308)

Ústecký kraj Jako první dítě roku 2011 v Ústeckém kraji se narodil dnes ráno v 02 : 57 chlapec v porodnici v Chomutově . (0.0152178301364)

Novoroční děti se letos zatím narodily v 01 : 20 v Břeclavi , ve 02 : 23 v Kyjově , ve 02 : 35 v Boskovicích , ve 03 : 04 ve Znojmě , v brněnské Nemocnici Milosrdných bratří ve 04 : 42 a ve Vyškově v 07 : 04. Kraj Vysočina Prvním občánkem Vysočiny roku 2011 je Simonka Svobodová z Vladislavi . (0.0199186992994)

" Prvenství v rámci republiky jsme si na Vysočině , podle krajského archivu , připsali ještě v roce 2004 , kdy se na Nový rok v Nemocnici Nové Město na Moravě šťastným rodičům narodil Vojtíšek Pavliš , " připomněl mluvčí . (0.0173352060335)

Naopak více dětí se narodilo vloni ve Stodě na jihu Plzeňska , kolem 430. Poslední tam přišlo na svět 31. prosince v 05 : 15. Celkem 3396 dětí přivedli na svět v největším zařízení v kraji , na Gynekologicko-porodnické klinice Fakultní nemocnice v Plzni , která řeší nejkomplikovanější případy z celých západních Čech . (0.0206837874796)

- **Výsledky statistické metody (reprezentované sumarizací z deseti vět)**

Jeho matka je z Hrádku nad Nisou , řekla dnes ČTK mluvčí Krajské nemocnice v Liberci Alexandra Kittnerová . (27.1482154102)

O porodné připravila ženu novela zákona o sociální podpoře , která od Nového roku podmínky k jeho přiznávání výrazně změnila . (36.005152626)

Deset minut po půlnoci tam přivítaly zdravotní sestry na světě holčičku vážící 2,8 kilogramu a měřící 48 centimetrů . (28.634976852)

Praha Prvním pražským miminkem letošního roku je chlapeček Vojta , který se devět minut po půlnoci narodil ve Všeobecné fakultní nemocnici (VFN) na Karlově náměstí . (39.9392309873)

Chlapec vážil 3,08 kilogramu a měřil 52 centimetrů , řekl ČTK vedoucí odboru zdravotnictví na krajském úřadu Josef Drbal . (29.7746139698)

Novoroční děti se letos zatím narodily v 01 : 20 v Břeclavi , ve 02 : 23 v Kyjově , ve 02 : 35 v Boskovicích , ve 03 : 04 ve Znojmě , v brněnské Nemocnici Milosrdných bratří ve 04 : 42 a ve Vyškově v 07 : 04. Kraj Vysočina Prvním občánkem Vysočiny roku 2011 je Simonka Svobodová z Vladislavi . (26.7133879898)

V Havlíčkově Brodě přišel na svět krátce po 04 : 00 Jakub a asi o půl hodiny později přivítali v pelhřimovské porodnici Lukáška . (33.7818981748)

Zatímco například letošní listopad byl podle tamní sestry slabší , prosinec naopak na množství porodů silnější . (27.1985435148)

Dítě přijede podle mluvčí kraje Berill Maschekové přivítat hejtman nebo jeho náměstek , a to zřejmě v neděli kolem poledne . (32.886421317)

Loni získal pomyslný titul prvního dítěte ve středních Čechách chlapeček jménem Chasan , který přišel na svět 13 minut po půlnoci v brandýské nemocnici . (35.8579511102)

- **Výsledky grafové metody (reprezentované sumarizací z deseti vět)**

" Ty peníze by se nám určitě hodily , ale co se dá dělat , hlavně , že je syn zdravý a má se k světu , " dodala . (0.016310107308)

Prvním novorozencem loňského roku se v hradeckém kraji stal Mário Matuš , který se v Oblastní nemocnici v Trutnově narodil na Nový rok v 04 : 59 rodičům z Libče u Trutnova . (0.014214027901)

Váží 3,7 kilogramu a měří 52 centimetrů , " řekla ČTK mluvčí nemocnice Hana Szotkowská . (0.0126734914566)

Ústecký kraj Jako první dítě roku 2011 v Ústeckém kraji se narodil dnes ráno v 02 : 57 chlapec v porodnici v Chomutově . (0.0152178301364)

" Chlapeček je v pořádku a je zdravý , " řekla ČTK sloužící sestra ve VFN . (0.012337973729)

Novoroční děti se letos zatím narodily v 01 : 20 v Břeclavi , ve 02 : 23 v Kyjově , ve 02 : 35 v Boskovicích , ve 03 : 04 ve Znojmě , v brněnské Nemocnici Milosrdných bratří ve 04 : 42 a ve Vyškově v 07 : 04. Kraj Vysočina Prvním občánkem Vysočiny roku 2011 je Simonka Svobodová z Vladislavi . (0.0199186992994)

" Prvenství v rámci republiky jsme si na Vysočině , podle krajského archivu , připsali ještě v roce 2004 , kdy se na Nový rok v Nemocnici Nové Město na Moravě šťastným rodičům narodil Vojtíšek Pavliš , " připomněl mluvčí . (0.0173352060335)

Na svět přišla v klatovské nemocnici dnes ráno v 05 : 09. Holčička váží 2,75 kilogramu a měří 48 centimetrů . (0.0120707887217)

Na první letošní miminko čekají i další zařízení jako nemocnice v Domažlicích , ve Stodě a také obě nemocnice v Plzni , tedy Mulačova i fakultní . (0.0139955145303)

Naopak více dětí se narodilo vloni ve Stodě na jihu Plzeňska , kolem 430. Poslední tam přišlo na svět 31. prosince v 05 : 15. Celkem 3396 dětí přivedli na svět v největším zařízení v kraji , na Gynekologicko-porodnické klinice Fakultní nemocnice v Plzni , která řeší nejkomplikovanější případy z celých západních Čech . (0.0206837874796)