Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

DIPLOMOVÁ PRÁCE

Multi-dokumentová sumarizace novinových článků

PLZEŇ, 2014 **Jaromír Novotný**

P R O H L Á Š E N Í

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

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

V Plzni dne 14.5.2014 _____

Jaromír Novotný

Poděkování

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

Anotace

V této diplomové práci je popsán základ automatické sumarizace, možná identifikace a dělení získaných souhrnů. Také jsou nastíněné různé možnosti rozdělení automatických i manuálních sumarizačních metod a jejich stručný popis. Dále je práce zaměřená na ohodnocování výsledků automatických sumarizačních metod, k tomuto tématu jsou uvedeny různé metody, které jsou i teoreticky popsány. Hlavní část práce je více-dokumentová sumarizace a různé metody spojené s tímto typem sumarizace. To zahrnuje i implementaci tří automatických multi-dokumentových sumarizačních metod. Následně je provedeno ohodnocení výsledných souhrnů získaných z implementovaných metod za pomoci automatického ohodnocení (softwarový balíček ROUGE). Tyto výsledky jsou v konečné části porovnány a je vyhodnocena jejich použitelnost v praxi.

Klíčová slova: automatická sumarizace, sumarizace novinových článků, multi-dokumentová sumarizace, ohodnocování výsledných sumarizací, testování automatických sumarizačních metod

The thesis describes the fundamentals of an automatic summarization, then types of summarization distributions, and also it’s pointed on options of automatic and manual methods of summarization and their properties. Next, this work is oriented on valuation of results from the automatic methods of summarization. This topic points on some of methods and their theoretical basics. Main part is focused on a multi-document summarization and methods of this summarization type. That includes implementation of a three automatic multi-document methods of summarization. Then is carried out a valuation of a resulting summaries (from implemented methods), by using an automatic valuation software package (ROUGE). These results (from automatic valuation) are compared (with each other) and data are evaluated. From that is decided, which of methods could be better for practical using.

Key words: automatic summarization, summarization of news articles, multi-document summarization, evaluation of summarization methods, test of automatic summarization methods

Obsah

1 Úvod

Tato diplomová práce se zabývá multi-dokumentovou sumarizací novinových článků. Slovo sumarizace bude v této práci používáno jako pojem procesu vytváření zkrácené verze textu, která informativně odpovídá originálnímu textu. Toto jasné stanovení významu je dáno kvůli dalším možnostem použití. Výsledná zkrácená verze textu se též může nazývat sumarizace, a proto bude zde používán pojem souhrn. Aby bylo možné vysvětlit multi-dokumentovou sumarizaci novinových článků je v prvé řadě důležité vysvětlení právě pojmů sumarizace a souhrn. Dále existuje mnoho možností rozdělení sumarizačních metod. Z toho důvodu byla vybrána jedna možnost a obecně popsána. Sumarizace novinových článků (obecný popis) a metody s tím spojené jsou zde dostatečně rozebrány. Metody mohou být manuální nebo automatické. V druhém případě je důležité si dát určitý pozor na vstup, který musí být v daném tvaru, na což zde nebylo opomenuto. Pro možnost otestování alespoň některých automatických sumarizačních metod na dané novinové články, byly vybrány tři metody. Latentní sémantická analýza je první zástupce, druhý je metoda založená na váze středů shluků a jako poslední metoda založená na nezáporné maticové faktorizaci a K-means shlukování. Všechny tyto metody jsou detailně popsány s použitými algoritmy. Aby bylo možné porovnat použitelnost těchto metod v praxi, je potřeba ohodnotit výsledné souhrny. To nás přivádí k ohodnocování výsledků sumarizačních metod. Celá jedna část této práce se zabývá touto problematikou. Použití softwarového balíčku ROUGE (a jeho popis) umožnilo ohodnotit výsledky metody, z kterých lze učinit určité závěry.

Začátek práce je věnován sumarizaci a souhrnu obecně. Důkazem velkého zájmu o sumarizaci, ať již automatickou nebo manuální, je velké množství definic souhrnů jako takových a velmi pestrý výběr sumarizačních metod. Teoretická část práce obsahuje možné rozdělení sumarizačních metod z různých pohledů. Každé rozdělení obsahuje krátký popis. Někdy jsou uvedeny i příklady sumarizačních metod, které se používají pro automatickou sumarizaci (vytvářenou strojově). Práce se zabývá sumarizací novinových článků. To zahrnuje i možné automatické a manuální metody (teoreticky popsané). Také se pamatuje na teoretický popis ohodnocení výsledných souhrnů. Další část je věnována multi-dokumentové sumarizaci. K výhodám a nevýhodám tohoto tématu je přidán teoretický popis automatických metod. Jak již bylo řečeno, byly vybrány tři automatické sumarizační metody k implementaci. Ta byla provedena v programovacím jazyce Python. Zatímco na ohodnocení výsledných souhrnů byl použit již vytvořený program.

Jedním cílem práce bylo představit sumarizaci z teoretického hlediska, dalším a to hlavním cílem práce bylo zaměřit se na automatické metody vytvářející sumarizaci se vstupem více-dokumentů, vybrat vhodné metody a nakonec provést implementaci vybraných metod. Poté aplikovat implementované metody na originální články, získat výsledky a vyhodnotit úspěšnost těchto výsledků a popsat vhodnost metod pro reálné použití.

2 Teoretické základy sumarizace

2.1 Úvod do sumarizace

Pojem sumarizace bude v této práci chápán jako proces vytváření zkrácené verze originálního textu se stejnou informativností. Zatímco pojem souhrn bude označovat již danou výslednou zkrácenou verzi originálního textu.

Souhrn se kolem nás vyskytuje v mnoha formách. Od úvodní stránky novin, přes nejdůležitější informace z článků časopisu nebo knížek až po výtahy informací na internetu. V případě úvodní stránky v novinách bývá souhrn velmi krátký, zato obsahuje nejdůležitější informace na odkazující článek, jako například téma článku, kdy a kde se událost přihodila, popřípadě ještě další informace.

Hlavním cílem sumarizace je získání obsahově nejdůležitějších informací z článku, tak aby souhrn byl co nejkratší. Za předpokladu, že by všechny věty v článku byly informativně na stejné úrovni, pak by výsledný souhrn nebyl velmi efektivní. Jakékoliv zkrácení článku by vedlo k výraznému snížení informativní úrovně. Naštěstí se informativní věty v článku objevují v dávkové formě, v tom případě lze odstranit nepodstatné věty. A tedy hlavní výzvou vytvoření souhrnu je identifikovat informativně důležité věty v článku od informativně nepodstatných.

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

1) Souhrn může být vytvořen z jednoho nebo více dokumentů (článků).

2) Souhrn musí zachovávat důležité informace originálu.

3) Souhrn by měl být krátký.

Další kapitoly jsou především zaměřeny na automatické sumarizační metody a na ohodnocení kvality výsledných souhrnů. Základ tvoří metoda (postavená např. na algebraickém výpočtu, sémantice, abstrakci nebo extrakci), podle které je napsán algoritmus zpracovávající daný vstup (text ve formě článku nebo více článků) do výsledné sumarizace (zkrácený text ideálně stejně informativní jako celý originál).

2.2 Definice souhrnu a sumarizace

Souhrn lze podle článku [17] 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). “

Sumarizace můžeme být definována podle článku [18]:

„ 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

Sumarizační metody lze dělit různými způsoby, záleží například na tom, jak chceme danou úlohu řešit nebo na vstupních datech. V různých článcích se používá rozdělení sumarizačních metod pro daný úkol a nezahrnují se další možnosti. Lze dělit základně na dvě hlavní skupiny, ale pro lepší přehled a získání obecného pohledu na sumarizační metody zde uvedeme podrobné rozdělení podle článku [7]

2.3.1 Na základě rozsahu

Jedno dokumentová sumarizace:

Sumarizace vychází z jednoho textu (článku nebo dokumentu) o určitém tématu.

Multi-dokumentová sumarizace:

Tato sumarizace vychází z více textů (tedy více článků nebo dokumentů), které jsou o stejném tématu nebo jsou situované v daném časovém období. Snižuje čas vyhledávání, zvláště když cíl uživatele je najít co nejvíce možných informací o daném tématu.

2.3.2 Podle jazyka

Mono-jazykové:

Mono-jazykový typ textové sumarizace spočívá ve vytvoření souhrnu ve stejném jazyce, jako je zdrojový text. Také většina těchto sumarizačních metod dokáže zpracovávat jen jeden jediný jazyk (řeč).

Multi-jazykové:

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

2.3.3Úrovně zpracování sumarizace

Povrchní přístup:

Informace jsou 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 požadavku. Výsledný souhrn je vytvářen extrakcí vět.

Hlubší přístup:

Je využito sémantické zpracování k určení významných částí textu. Zjišťují se textové jednotky a jejich vzájemné vztahy. Informace stavby textu a rétorické struktury (popřípadě hypertextových značek) mohou být využity pro identifikování (určení) vzájemných vztahů. Výsledný souhrn je vytvářen extrakcí vět nebo abstrakcí nejdůležitějších informací z článku.

2.3.4 Účel, pro který je souhrn vytvářen

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 sumarizace zařadit do skupiny automaticky vytvářených.

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 celého dokumentu.

2.3.5 Sumarizace z pohledu uživatelů

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

Tato sumarizace je zaměřena na určité požadavky konkrétního uživatele nebo skupiny uživatelů.

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

Tyto sumarizace se zaměřují na informace v dokumentech, které jsou spojeny s daným tématem dokumentu. Tedy výsledný souhrn (v případě nějaké extraktivní metody) je tvořen větami originálního textu, které se nejvíce blíží k tématu.

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

Do těchto sumarizačních metod lze zařadit například metody, které nejdříve ve vstupním textu identifikují klíčová slova (například ohodnocením tf-idf vahami), dále je proveden dotaz na uživatele, který vybere důležitá slova pro provedení sumarizace a po té je vytvořen souhrn podle sémantické podobnosti vybíraných vět s vybranými slovy.

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.

Edmunsonova 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 nejen o identifikované a zachycené duplikované termy, ale také o související termy. Tento koncept (známý jakou soudržnost), spojuje sémanticky související termy, jež jsou důležitou součástí uceleného textu. Nejjednodušší formou této soudržnosti je slovní soudržnost.

Grafové metody:

Graf je podle článku [15] nejefektivnější reprezentace vztahů mezi dvěma proměnnými. Tedy tyto metody jsou následně navrženy jako prohledávací nástroj grafové struktury.

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 výsledného souhrnu. Tato metoda byla původně vytvořena pro použití na jedno-dokumentové sumarizace a později upravena a přizpůsobena na multi-dokumentovou sumarizaci.

2.4 Charakteristické vlastnosti souhrnu

Pro lepší pochopení a identifikování souhrnů, budou definovány následující aspekty (možnosti), aby podle článku [14] mohl být jakýkoliv souhrn charakterizován třemi hlavními charakteristickými třídami:

2.4.1 Vstup: charakteristika vstupního textu(ů)

Velikost vstupu: jeden dokument nebo více dokumentů

Jedno-dokumentový souhrn je odvozen z jediného vstupního textu.

Multi-dokumentový souhrn je text, který je tvořen (a tudíž obsahuje informace) z více vstupních textů (minimálně 2) a většinou je využíván v případě stejného tématu vstupních textů nebo že vstupní texty jsou z daného časového období.

Specifičnost: specifická oblast nebo obecná

V případě, že se vstupní texty týkají jedné oblasti, která se zaměřuje na specifický obsah. V porovnání s obecným případem má výstup specifické formáty. Souhrn specifické oblasti vychází ze vstupního textu(ů), jehož téma(ta) je vztaženo k jedné jediné oblasti. Tedy lze předpokládat menší nejednoznačnost termů, idiosynkratické slovo, použití gramatiky, specializované formátování, atd. Tyto předpoklady se mohou odrážet ve výsledné sumarizaci.

Souhrn obecné oblasti je tvořen ze vstupního textu (ů) jakékoliv oblasti a nemůže vytvářet žádné výjimky jako souhrn specifické oblasti.

Žánr a rozsah

Vstup (článek) zaměřený na žánr (styl nebo téma) se zaměřuje jen na daný žánr (téma). Například sportovní novinový článek (takovýto článek se zaměřuje jen na sportovní událost popisovanou v něm a na nic jiného). Tedy souhrn bude zaměřený na stejný žánr jako vstup.

Vstup o určitém rozsahu se může lišit od délky knih k délce odstavců.

Na vstupu, který je zaměřen na určitý žánr nebo je o určitém rozsahu, může záležet výběr automatické sumarizační metody. Některé metody se více hodí na zpracování například článku se sportovním žánrem než na článek s žánrem jiným.

2.4.2 Výstup: charakteristika textu souhrnu

Odvození: Extrakt nebo abstrakt

Extrakt vybírá části z vstupního textu(ů) (od jednotlivých vět až po celé odstavce) a vytváří z těchto částí výsledný souhrn.

Abstrakt vytváří souhrn jako nově vygenerovaný text, který je výsledkem reprezentace vnitřního výpočtu analyzovaného vstupu.

Souvislost: plynulost nebo trhanost

Plynulý souhrn je napsán jako celek, věty gramaticky správně, vztahy mezi větami zohledněny a věty jdou po sobě podle pravidel struktury souvislého projevu.

Trhaný souhrn je rozdělen na části skládající se z jednotlivých slov nebo textu, které nejsou složeny do gramaticky správně postavených vět nebo nejsou složené do souvislých odstavců.

Zaujatost: neutrální nebo vyhodnocující

Tato charakteristika platí zejména v případě, kdy je vstupní materiál předmětem stanoviska nebo zaujatosti.

V neutrálním souhrnu (částečném nebo nestranném) je odražen obsah vstupního textu (ů).

Ve vyhodnocujícím souhrnu je zahrnuto zkreslení ať už explicitně (použitím vyjádření názoru) nebo implicitně (prostřednictvím zahrnutím materiálu s jedním vychýlením a vynecháním materiálům s jinými vychýleními).

Konvenčnost: pevná nebo volná

Souhrn, který pevně shrnuje situaci je vytvářen pro specifické použití čtenáře (nebo více čtenářů) a danou situaci. Jako takový může vyhovovat příslušné vnitřní konvenci zvýraznění, formátování a takové.

Souhrn, který volně shrnuje situaci, nepředpokládá pevné konvence, ale je vytvářen a zobrazován rozmanitostmi v nastavení, ke čtenářům a pro účel.

2.4.3 Účel: charakteristiky použití souhrnu

Publikum (uživatelské): obecné nebo dotazově orientované

Obecný souhrn dává pohled autora na vstupní text (y) se stejnou důležitostí pro všechny hlavní témata ve vstupu. Dotazově (uživatelsky) orientovaný souhrn upřednostňuje specifická témata nebo hlediska textu s ohledem na přání uživatele dozvědět se více o vybraných tématech. To může být provedeno explicitně zvýrazněním relativních témat nebo implicitně vynecháním témat, které neodpovídají přání uživatele.

Použití: orientační nebo informativní

Orientační souhrn poskytuje pouze informaci o hlavním předmětu nebo oblasti vstupního textu(ů), aniž by zahrnul obsah.

Z informativního souhrnu lze vysvětlit, o čem vstupní text pojednává, ale ne vždy co obsahuje. Informativní souhrn zahrnuje (pouze některý) obsah a dovoluje (částečně) popsat, co obsahoval vstupní text.

Rozpínavost: souvislost nebo jen daná informace (zpráva)

Souhrn souvislostí, předpokládá, že čtenář nemá velkou předchozí znalost o obecném obsahu vstupního textu (ů), a proto souhrn obsahuje vysvětlující materiál o okolnosti místa, času a informaci vystupujících osob.

Souhrn jen dané informace obsahuje jen určitou informaci nebo základní téma, s tím, že se předpokládá dostatečná čtenářova znalost souvislostí. Ty jsou interpretovány formou kontextu ve výsledném souhrnu.

3 Sumarizace novinových článků

3.1 Úvod do sumarizace novinových článků

Sumarizace novinových (zpravodajských) článků vznikla 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 automatická sumarizace, která nahrazuje celé články kratšími verzemi se stejným přínosem informací. Tudíž čtenáři nemusejí číst celé články, ale pouze nejvíce informativní části z nich (souhrny).

Díky tak velkému množství (zpravodajských) článků a úmyslu uspořit čas začal být velký zájem právě o automatické sumarizační metody.

3.2 Příklad metod pro sumarizaci novinových článků

V literatuře můžeme nalézt dva hlavní přístupy, lingvistický přístup, statistický (automatické sumarizace) a popřípadě jejich kombinace. Výsledné souhrny lze v základně dělit na abstrakční nebo extrakční typ.

3.2.1 Jednoduchá manuální sumarizační metoda

Pro zajímavost je v této práci uvedeno, jak může být sumarizován novinový článek bez použití automatické metody. Lze podotknout, že tento postup může být pro člověka náročný v případě delších, a nebo více informativních (více důležitých informací) článků. Tato jednoduchá metoda vychází z článku [13].

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 a umístění v novinách. Při zpracovávání novinového článku musíme tedy zachytit hlavní zprávu, jež obsahuje. Postup na vytvoření takovéto sumarizace lze 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 souhrnu 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 určité informace a právě tato informace je nazývána hlavní myšlenka. Ta má přímý vztah k otázce „proč“ v článku, jelikož tato otázka ji rozšiřuje. Část souhrnu zachycující hlavní myšlenku, by měla být složena maximálně z tří vět. V některých případech, může mít

článek více hlavních myšlenek, tedy musí být zachovány všechny části souhrnů obsahující jednotlivé 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 souhrnu vložíme zakončující větu. Proces sumarizace nemusí končit v místě, kde končí samostatný článek, ale v místě zakončení příběhu (informativní části článku).

3.2.2 Definice ontologie

Jelikož jsou v této práci uvedené metody, které jednak mají a jednak nemají základ postavený na ontologii, je zde uvedena definice ontologie.

V minulých letech se velmi rozšířil zájem o aplikování technik související s ontologií. Nicméně v literatuře stále chybí unikátní definice. Zde je tedy uvedena jedna možnost a to Gruberova definice ontologie z článku[12]:

> *„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 některých 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 [11], že ontologie rozčleňuje svět do několika objektů, aby je mohla lépe popsat. Definice popsání těchto objektů nebo jejich reprezentace závisí na jednotlivých aplikacích.

Ontologie je analyzování a shromažďování sémantických informačních tříd z článku. Předpokládá se, že každý článek obsahuje několik podtémat. Za použití ontologie lze identifikovat tyto podtémata (získání sémantických informací) článku.

3.2.3 Metoda založená na ontologii

Za použití sémantické informace, obsažené (zakódované) v ontologii, tento systém určí, která témata jsou užitečná pro extrakci odstavců. Navržení a vytvoření ontologie jsou první dva kroky potřebné pro vytvoření sumarizačního systému. V článku [11] je navržen postup, jak vytvořit metodu založenou na ontologii.

3.2.4 Metoda, jež není založená na ontologii

Metoda, jež není založená na ontologii, může být například metoda extrakční, která využívá frekvenci termů, délku vět, bonusová slova (předem určená databází) a klíčová slova. Po získání těchto proměnných můžeme použít následujícího vzorce z článku [11] pro výsledné ohodnocení:

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

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

3.2.5 Multi-dokumentová sumarizace

Pojmy, které lze využít pro multi-dokumentovou sumarizaci:

Extrakce událostí a jejich referencí

V tomto případě lze podle článku [10] říci, že shluk je tvořen články, které obsahují buď dané události, nebo články s referencemi o nich. Z daného pohledu se články ve shluku dělí na hlavní (obsahující určitou událost) a na vedlejší. Vedlejší články buď pouze obsahují referenci na hlavní článek, nebo mohou popřípadě i rozšířit událost o další část (informace). Toho lze využít v multi-dokumentové sumarizaci. Lze identifikovat vedlejší články, které pak nemusí být zařazeny do souhrnu (aby se neopakovala stejná událost nebo informace několikrát).

Metoda lexikální soudržnosti (Lexical cohesion) 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. Událost je definovaná následovně: Událost je informace popisující fakta a podobné informace konkrétního data.

Extrakce informace o událostech

Informace o událostech popisuje nějakou akci (událost, např. průběh voleb) v článku. Může být vztažena k jediné větě nebo až k celému článku. Obsahuje význam o dané akci, a tedy lze pomocí těchto informací porovnávat části (celé) článků. Tím může být zabráněno, aby byly vybrány do výsledného souhrnu významově stejné texty (články). Podle článku [10] je dobré, pro získání informace o událostech, hlouběji analyzovat struktury vět dokumentu (článku). Ovšem provedení takovéto analýzy je velmi obtížné. Místo toho lze použít podobnostní analýzu. Lze také využít přídavné informace o času (v případě, že je dostupná). Tisk může vydat v dvou po sobě jdoucích měsících články o stejné události. Pohled druhého článku na událost může být jiný než v předchozím článku. Tyto články, na základně informace o událostech každého z nich, si budou podobné. V tomto případě lze využít přídavnou informaci o času a použít článek neblíže přítomnosti. K identifikování informace o událostech mohou být použity následující parametry:

- **Kořen**, základní část informace o událostech (většinou sloveso věty reprezentující akci).
- **Modifikátor**, slova rozšiřující kořen.
- **Zápor**, reprezentuje pružnost vyjádření.
- **Hloubka**, je hodnota délky, mezi kořenem informace o událostech a kořenem věty, získaná ze stromu reprezentujícího závislosti vět.
- **Datum**, je časový údaj, kdy se událost stala. Tento parametr není potřebný k identifikování (přídavná informace).
- **Datum vydání článku**.
- **Kusy**, slouží k reprezentaci polohy slov ve větě.

Metoda pro získání informace o událostech z vět může být popsána následně:

1) Aplikuje se Cabocha [20] metoda. Tím se získá strom reprezentující závislosti vět.
2) Výběr slov, která jsou kandidáty na kořen informace o událostech věty (slovesa a podstatná jména).
3) Nastaví se zápor. Podle toho, jestli kořen obsahuje zápor.
4) Modifikátor je získán ze stromu reprezentujícího závislosti vět. Slova v modifikátoru nejsou jen slova přímo rozšiřující kořen, ale také slova, která rozšiřují samotná slova v modifikátoru.
5) V případě, že lze z věty získat datum (kdy se událost stala), tento údaj je zaznamenán jako datum.
6) Z informacích o článku se stanoví datum vydání článku.
7) Hloubka a kusy jsou hodnoty vypočítané z porovnání informace o událostech (doposud získaných) a stromu reprezentujícího závislosti vět.

Reference informace o událostech a jejich použití

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, a to analyzováním odkazů. Základní koncept algoritmu závisí na předávání důležitosti stránek skrz analyzování odkazů. Jinými slovy, stránka obsahující velké množství odkazů získává svojí důležitost z ostatních stránek. Odkaz ze stránky s vyšší důležitostí má vyšší důležitost v porovnání s odkazem ze stránky s menší důležitostí. Další popis lze nalézt v článku [10].

Využití pozice vět a úvodního dotazu (otázky)

Je velmi časté, že důležité věty se vyskytují na začátku článku. Toho lze využít tak, že se zavede hodnota pozice vět. Tato hodnota je pak využita při výpočtu důležitosti jednotlivých vět. Úvodní dotaz je formalizován jako věta, která tvoří dokument o jedné větě a který se vyskytuje ve více dokumentech. Díky tomuto formalizování lze předávat důležitost dotazu a nastavit počáteční hodnoty důležitosti této věty. Tato věta dotazu se nezahrnuje do výsledného procesu extrakce. Úprava PageRank algoritmu tak, aby využíval tyto dva pojmy, je uvedena v článku [10].

Přeuspořádání textu a sloučení založené na podobnosti vět

Jak již v předchozích pojmech bylo uvedeno, obsahově podobné věty mají nejen podobné odkazy, ale také podobné váhy z ohodnocení. Tedy v případě vybírání vět do konečného souhrnu podle nejvyšší váhy ohodnocení může dojít k vybrání nepotřebné věty, která má stejný význam jako věta již byla vybrána. Proto je vhodné použít nějaký mechanismus k identifikování nepotřebných vět a k jejich odstranění. V článku [10] je uvedeno řešení využitím informace o událostech. To lze provést získáním všech informací o událostech vybírané věty a porovnat s informacemi o událostech vět, které již byly zařazeny do souhrnu. Podrobnější algoritmus pro výpočet nepotřebných vět s využitím již uvedených pojmů je k nalezení v článku [10].

Sumarizace na základě analogie vytvářená z příkladných novinových článků

Tato metoda je představená v článku [9]. Jsou uváděny tři výhody při použití této metody na sumarizaci novinových článků a to: vysoká modularita, bez počítání váhy (důležitosti) každého slova a používání kontextu.

Jak už bylo řečeno, metoda využívá následující tři výhody:

1) **Velká modularita**

   Obecně je důležité, aby mohla být metoda jednoduše upravována. Pouhým přidáváním instancí lze jednoduše zlepšovat vytvořený software. Ovšem přidání určitých instancí může způsobit vedlejší efekty.

2) **Využití podobnosti spíše než váhy (důležitost) vět**

Většina minulých metod na automatickou sumarizaci byla založena na extrakci. Byla počítána váha (důležitost) každého slova. To sloužilo k zařazení vět do souhrnu. Samozřejmě je velice obtížné počítat tyto váhy tak, aby odpovídali manuální sumarizaci. Metoda na základě příkladů tedy nepočítá váhu slov, ale počítá místo toho s podobností.

3) **Použití kontextu**

Obecná statická metoda se snaží počítat pravděpodobnost každého slova, jež se objevuje v souhrnu. To může ztížit zachování kontextu, jelikož se statistický přístup zaměřuje na globální pravděpodobnost. Nicméně metoda na základě příkladů se pokouší nalézt podobné instance ze sady instancí, což může být kontextově vhodnější.

Přístup na základě příkladů vytváří jazyk podle napodobování instancí, které vychází z automatického překladu na základě analogie. V automatickém překladu byl tento přístup implementován a doposud dosáhl efektivních výsledků.

Člověk při sumarizaci používá své znalosti a dřívější poznatky. Z toho důvodu se začíná více zaměřovat na sumarizační metody založené na antologii, zde metoda sumarizace na základě příkladů. Metoda sumarizuje vstupní text ve třech krocích:

1) Nahrazení podobných instancí na vstupním textu (článku).

2) Vytváří spojení odpovídajících vět mezi vstupním textem a podobnými instancemi.

3) Zkombinuje odpovídající věty do výsledného souhrnu.

Mnoho sumarizačních metod vytvořilo souhrn z jedné věty, která byla vybrána například z novinového článku (extrakční typ). Metoda na základě příkladů vytváří výsledný souhrn kombinací více vět (abstrakční typ). Tudíž může vytvořit souhrn s vysokou kompresí, jež zahrnuje informaci z více částí zdrojového textu.

4 Ohodnocování výsledků sumarizačních metod

4.1 Ohodnocování výsledků sumarizačních metod

Ohodnocování výsledků sumarizačních metod slouží k posouzení rozdílů kvality mezi jednotlivými sumarizačními metodami. Samozřejmě můžeme mluvit o ohodnocování, které provádí člověk na základě svých vlastních zkušeností. Výsledky takového ohodnocení jsou u každého jedince jiné, a kvůli tomu nejsou dostatečně přesné ke stanovení závěrů o nejlepší metodě. Budeme se tedy raději zabývat automatickým ohodnocováním výsledků sumarizačních metod, které je prováděno danou metodou řízenou daným algoritmem. Při správném zvolení metody již můžeme stanovit závěry o kvalitě sumarizačních metod.

Obecně nelze přesně definovat, jak metody pro ohodnocení výsledných souhrnů pracují. Sice v pár metodách je použit stejný základ, ale celkové výpočty a konstrukce těchto metod se liší. Většina má společné to, že používají k ohodnocení porovnání automaticky vytvořených souhrnů (tedy strojem) se souhrny, které jsou vytvořeny manuálně (tedy člověkem). Díky tomu automatické ohodnocování by se mohlo blížit ohodnocení manuálnímu.

4.2 Možnosti rozdělení metod pro ohodnocování výsledků

Metody používané k ohodnocování souhrnů lze dělit více způsoby např. podle toho, s jakými vstupními texty pracují. Rozdělení, které je zde uvedeno, je podle článku [5] a lze jej interpretovat následovně:

4.2.1 Přímé (intrinsic)

Zaměřené na lingvistickou kvalitu textu:

Toto hodnocení je nejčastěji prováděno manuálně (člověkem). Každému souhrnu je přiřazena hodnota z předem definovaného měřítka (tabulky).

- Gramatika
- Bez přebytečných slov
- 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
  Používáno v případě souhrnu vytvořeného extrahováním vět.
  Vyhledává shodné věty, jež jsou obsaženy jak v ideálním (manuálně) tak automaticky vytvořeném souhrnu. Pracuje pak s počtem shodných vět.
  - Přesnost
  - Úplnost
  - F-skóre
  - Relativní užitečnost
- Podobnostní míry
  Využívají porovnávání aktuálních slov ve větě, než celých vět. Výhodou je, že porovnávají buď extrakty vytvořené člověkem, nebo extrakty vytvořené automatickou sumarizací, se souhrnem napsaným člověkem (abstrakčně).
  - Kosinová podobnost
  - Překrytí obsahu
  - Nejdelší společný podřetězec
  - Společné n-gramy (ROUGE)
  - Ohodnocování vět (pyramidy)
  - Hodnocení na základě LSA

4.2.2 Nepřímé (extrinsic)

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

- Kategorizace dokumentů
- Vyhledávání informací
- Zodpovídání dotazů

4.3 Stručný pohled na některé metody

4.3.1 Zaměření na lingvistickou kvalitu textu

Ohodnocení kvality s tímto zaměřením je prováděno vyhodnocením všech částí dohromady, tedy hodnotí se veškeré aspekty ve výsledném souhrnu. Zároveň tento typ ohodnocení nemůže být prováděn automaticky. Většinou je každému souhrnu podle kvality přiřazeno označení provedené manuální anotací (např. od A-velmi dobré až po E- velmi špatné).

Gramatika:

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

Bez přebytečných slov:

Výsledný text by neměl obsahovat přebytečná 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 souvislý.

4.3.2 Hodnocení obsahu

Ko-selekční přístup:

Hlavní vyhodnocování kvality souhrnů ko-selekčního přístupu jsou přesnost, úplnost a F-skóre.

Přesnost (P):

Je hodnota počtu vět shodných v obou sadách textů (tedy v hodnoceném a ideálním souhrnu) dělená hodnotou celkového počtu vět hodnoceného souhrnu (automaticky vytvořeného).

$$P = \frac{|H \cap I|}{|H|}$$

Kde H je hodnotící souhrn, I je ideální souhrn, vztah $|H \cap I|$ udává počet společných vět v obou souhrnech a $|H|$ udává počet vět v hodnotícím souhrnu

Úplnost (R):

Je hodnota počtu vět shodných v obou sadách textů (tedy v hodnoceném a ideálním souhrnu) dělená hodnotou celkového počtu vět ideálního souhrnu (manuálně vytvořeného).

$$P = \frac{|H \cap I|}{|I|}$$

Kde *H* je hodnotící souhrn, *I* je ideální souhrn, vztah $|H \cap I|$ udává počet společných vět v obou souhrnech a $|I|$ udává počet vět v ideálním souhrnu

F-skóre:

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

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

Lze uvést obecnější 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 β > 1 a při hodnotě β < 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 úsudku člověka přibližně stejné skóre, ale podle této metody je jedna ohodnocena podstatně vyšším skóre než druhá.

Relativní užitečnost:

Tato metoda řeší hlavní problém přesnosti a úplnosti. Proto se zavádí pojem Relativní užitečnost (RU). Sumarizační model této metody reprezentuje všechny věty vstupu společně s jejich hodnotami spolehlivosti. To usnadňuje vybrání určitých 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 podle důležitosti. Tuto hodnotu určuje člověk a 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 může být definován 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.

Podobnostní míry:

Kosinová podobnost:

Základní podobnostní míra založená na obsahu 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 (základních podob slov nebo frází). $\|X\|$ 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.

ROUGE-N: N-gramové statistiky opakovaných výskytů:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), tato metoda je používána jako automatická metoda k ohodnocení. 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. Další metody používané ROUGE jsou uvedeny v kapitole 6.2.

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 jsou získány z dodatečné informace těla souhrnů (omezené určitými podmínkami). S dodatečnou informací se začínají identifikovat podobné věty. Po té začíná přesnější prověřování, které může lépe nalézt související podčásti. SCU_S, jenž se více vyskytuje v manuálně vytvořených souhrnech, dostává vyšší váhu, takže pyramida se vytvoří až po ohodnocení *SCU* dodatečné informace 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 a manuálně získaným souhrnem.

5 Multi-dokumentová sumarizace

5.1 Teorie

Jak může být z názvu patrné, vstupem vybrané multi-dokumentové automatické metody není pouze jeden vstupní text (dokument), ale několik textů (dokumentů) najednou. Vstup tedy může být reprezentován jedním nebo ve většině případů i více soubory obsahující texty (články). V podstatě lze říct, že jedno-dokumentová sumarizace je součást multi-dokumentové sumarizace. Jako velký rozdíl mezi jedno a multi dokumentovou sumarizací může být tedy uveden formát vstupu. Automatická sumarizace s multi-dokumentovým vstupem pracuje na podobném principu jako s jedno-dokumentovým. V multi-dokumentové sumarizaci je velmi těžké a důležité identifikovat podobné věty (z různých dokumentů). To aby nebyly do výsledné sumarizace zařazeny věty se stejnou informací (obsahem).

Většina automatických sumarizačních metod pro zpracovávání multi-dokumentů předpokládá, že články (dokumenty) budou odpovídat jednomu tématu nebo určitému časovému období (tedy budou zaměřeny na určitou událost). V případě, že tomu tak není, je nutné pro většinu automatických sumarizačních metod vstup roztřídit a to např. automatickou metodou na bázi shlukování. V této práci se používají již předem připravené vstupní texty (dokumenty), které mají vždy stejné téma nebo jsou z daného časového období. Z toho důvodu nebylo potřeba další studie metod zaměřených na přerozdělování a shlukování dokumentů.

Automatická sumarizace multi-dokumentů má velký potenciál v případech velkého množství informací. Díky sumarizaci získáme nejdůležitější informace ze všech dokumentů, bez opakování stejných nebo podobných informací zmiňovaných ve více dokumentech.

Mezi požadavky automatické sumarizace multi-dokumentů se dá řadit například určitý formát vstupu, kdy je potřeba nejdříve provést shlukování daných témat dokumentů k sobě. Dále pokud jsou vstupní texty (dokumenty) shlukovány tak, že se jedná o určité časové období, může dojít k tomu, že informace se v průběhu času mění (například si po čase důležité informace mohou odporovat). Automatické sumarizační metody, jejich dělení, stručný popis a použití je rozebíráno viz kapitola 4.

5.2 Vybrané metody

Byly vybrány tři metody k implementaci a otestování vhodnosti použití v reálných úlohách. Každá metoda načítá stejný vstupní text. Tento text je při načítání upravován. Úprava je pouze v odstranění stop slov (např. předložky a velmi často se opakující slova) a zaměnění velkých písmen na malá (pouze u prvních slov každé věty).

První byla k implementaci vybrána automatická sumarizační metoda používající Latentní sémantickou analýzu (LSA – „Latent Semantic Analysis“). Tato metoda je sice původně postavená pro automatické sumarizování jedno-dokumentového vstupu, ale byla upravena tak, aby prováděla automatické sumarizace z multi-dokumentového vstupu.

Upravení bylo provedeno z důvodu otestování, jakých výsledků tato předělaná metoda dosáhne v poměru s metodami určených přímo k automatické sumarizaci multi-dokumentů.

Automatická sumarizační metoda založená na váze středů shluků (MEAD – „Centroid-based method“) byla vybrána jako druhá metoda k implementaci. Tato metoda je velmi rozšířená a podle mnoha článků velmi úspěšná, z těchto důvodů byla vybrána k porovnávání s ostatními metodami.

Jako třetí automatická metoda vybrána k implementaci je metoda postavená na nezáporné maticové faktorizaci (Non-negative matrix factorization – NMF) a K-means shlukování. Byla vybrána z důvodů NMF rozkladu matic (obsahující váhové ohodnocení vět), která pracuje na podobném principu jako LSA metoda, která používá SVD rozklad matic. Dále tato metoda používá rozklad na k shluků (K-means shlukování). Tato metoda je též porovnávána s ostatními implementovanými metodami.

5.2.1 Latentní sémantická analýza (LSA – „Latent Semantic Analysis“)

Tato metoda (zde uvedená) je původně navržena na jedno-dokumentovou sumarizaci, ale pro zajímavost je zde přepracován vstup této metody, tak aby metoda pracovala s multi-dokumenty, s tím, že předpokládá, že dokument je jen jeden, tedy jedno-dokumentová sumarizace. Podklady pro tuto metodu byly čerpány v článku [7] a [8].

Teoretická analýza LSA:

Latentní sémantická analýza je jedna z metod, jež dokazují, že metody nemusí být např. ryze statistické. Tato metoda je statisticko-algebraitská metoda. Výsledný souhrn z této metody je extrakce smyslu slov a podobnosti vět, při čemž používá informace o slovech použitých v kontextu.

Tato metoda používá rozklad matic singulární dekompozicí (SVD). SVD (“Singular value decomposition”) je numerický proces, který je používán při redukci dat. Nejdříve z článku musí být vytvořena matice A, jež je matice termů (slova a znaky ve větě) proti větám. Tedy $A = [A_1, A_2, \dots, A_N]$ kde A_i je sloupcový vektor reprezentující frekvenci termů ve větě *i* daného dokumentu. Matice *A* má rozměry *m x n*, kde *m* je počet termů dokumentu a *n* počet vět. Jelikož se ve většině případů každé slovo v každé větě nevyskytuje, můžeme říci, že matice *A* je řídká.

SVD rozklad je dán vztahem:

$$A = U\Sigma V^T$$

Kde $U = [u_{ij}]$ je *m x n* sloupcově ortonormální matice, jejíž sloupce jsou nazývány levými singulárními vektory, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ je *n x n* diagonální matice, kde diagonální prvky matice jsou nezáporná singulární čísla řazená sestupně a $V = [v_{ij}]$ je *n x n* ortonormální matice, jejíž sloupce jsou nazvány singulárními vektory.

Rozměr matic je následně redukován na *k* dimenzí, kde *k*<*n*, tedy dostáváme matice o rozměrech: *U* s rozměru *m x k*, Σ o rozměru *k x k* a V^T o rozměru *k x n*.

Použití LSA na sumarizaci dokumentu:

1) První postup k použití LSA metody k sumarizaci dokumentu je využití matice V^T získané z rozkladu SVD, která popisuje míru významnosti vět v hlavních tématech dokumentu. Tedy vybereme do souhrnu *P* vět (1,…,j,…,P), které jsou reprezentovány j-tým pravým singulárním vektorem z matice V^T. Každá věta je tedy reprezentována sloupcovým vektorem $[v_{i1}, v_{i2}, .., v_{ik}]$. Do souhrnu je tedy zařazena věta s největší indexovou hodnotou v j-tém pravém singulárním vektoru. Tato metoda má ovšem nevýhodu stejné důležitosti všech *P* v souhrnu obsažených témat.

2) Druhý postup je zdokonalení prvního postupu. Díky nevýhodě prvního postupu, tedy stejná důležitost všech *P* v souhrnu obsažených témat, ale jejichž významnost se může výrazně lišit. To lze sledovat v matici Σ. Tedy v případě násobení Σ^2 zohledňuje statistickou významnost hlavních témat, která je úměrná kvadrátu příslušného singulárního čísla. Může být tedy řečeno, že počítáme v *k* rozměrném latentním prostoru témat délku vektoru s_r pro r-tou větu dle vzorce:

$$s_r = \sqrt{\sum_{i=1}^{k} {v^2}_{ri} * {\sigma^2}_i}$$

   Vybírá se tedy *P* vět s největší délkou s_r.

3) Další postup je zkombinování grafové a LSA sumarizační metody. Zde též využijeme SVD rozkladu matice *A*, její redukci a následnou rekonstrukci matice *A* z redukovaného stavu. Tedy sloupce zrekonstruované matice *A* představují sémanticky reprezentované věty. Na tuto matici aplikujeme ohodnovací algoritmus grafové metody a do výsledného souhrnu zařadíme *P* vět s největší hodnotou.

Implementace LSA metody:

Implementace byla provedena v programovacím jazyce Python. Tato metoda je extraktivní. Můžeme tedy napsat algoritmus následovně:

a) 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).

b) Vytvoříme matici *A*, tedy $A = [A_1, A_2, \dots, A_N]$ kde A_i je sloupcový vektor reprezentující frekvenci termů ve větě i daného dokumentu.

c) Pro přesnější počítání vytvoříme z matice *A* matici TF-IDF vah. Pro každý pozici v matici TF-IDF vah je počítán následovně: $\text{TF} - \text{IDF}_{ij} = tf_{ij} * idf_{ij}$

Kde inverzní frekvenci termů (idf) pro každý term (slovo nebo znak) je počítáno následujícím způsobem:

$$idf_{ij} = \log \frac{N}{n_{ij}}$$

Kde *N* reprezentuje počet slov ve větě *i* a n_{ij} je výskyt termu (slova) *j* ve větě *i*.

A normovanou frekvenci termů (tf_{norm}) pro každý term (slovo nebo znak) je počítána následujícím způsobem:

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

Kde $tf_{i,j}$ je frekvence 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.

d) Následně byly použity tři různé postupy konečného zpracování:

Postup I:

a) Byl proveden SVD rozklad matice TF-IDF vah. SVD rozklad byl použit za pomoci knihovny scipy a funkce linalg.svd(TF-IDF)v programovacím jazyce Python.

b) Zredukujeme rozměr matic SVD a dále budeme používat tyto matice zredukované.

c) Z tohoto rozkladu je využita matice V^T, která popisuje míru významnosti vět v hlavních tématech dokumentu. Tedy vybereme do souhrnu *P* vět (1,…,j,…,P), které jsou reprezentovány j-tým pravým singulárním vektorem z matice V^T. Každá věta je tedy reprezentována sloupcovým vektorem $[v_{i1}, v_{i2}, .., v_{ik}]$.

d) Do výsledného souhrnu zařazujeme tedy větu s největší indexovou hodnotou v j-tém pravém singulárním vektoru.

Tento postup byl implementován, ale neuvádíme zde výsledky, jelikož tato metoda má nevýhodu, která je řešená v postupu II.

Postup II (délka vet):

a) Opět byl proveden rozklad SVD matice TF-IDF vah.

b) Zredukujeme rozměr matic SVD a dále budeme používat tyto matice zredukované.
c) Dále vypočítáme délku věty podle vorzce:

$$s_r = \sqrt{\sum_{i=1}^{k} v^2{}_{ri} * \sigma^2{}_i}$$

Kde v_{ri} je hodnota termu *i* ve větě *r* a σ_i je hodnota σ věty *i*.

d) Do výsledného souhrnu zařazujeme věty s největší hodnotou délky *s*.

Postup III (kombinace LSA a LexRank metody):

a) Opět provedeme rozklad SVD matice TF-IDF vah.

b) Zredukujeme rozměr matic SVD a dále budeme používat tyto matice zredukované.

c) Zrekonstruujeme matici TF-IDF z zredukovaných matic SVD.

d) První krok k vytvoření kosinové matice.
- Pokud na pozici [i,j] je hodnota v zrekonstruované matici TF-IDF větší než prahová hodnota, přiřadíme do matice na pozici [i,j] hodnotu 1 a do pole stupňů na pozici *i* přičteme 1.
- Do kosinové matice přiřadíme na pozici [i,j] hodnotu 0. Tento bod je proveden pouze v případě, že není splněna předchozí podmínka.

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

e) Druhý a konečný krok pro získání kosinové matice:

V tomto bodě v podstatě doopravíme kosinovou matici z minulého bodu následujícím způsobem:

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.

f) Vytvoříme konečný vektor *L*, jež obsahuje konečné LexRank váhy.

Tento vektor vytvoříme pomocí „Silové metody“ (Power Method) [23]. 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.

g) Vybereme *p* (reprezentuje číslo počtu vět, jež uživatel požaduje extrahovat) vět původního článku (který je počítán z podle počtu vět 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.

5.2.2 Metoda založená na váze středů shluků (MEAD – „Centroid-Based summarizer“)

Teorie metody založené na váze středů shluků:

Vstupem pro vytváření souhrnu musí být již předem připravený shluk dokumentů, v kterém jsou dokumenty se stejným tématem nebo dokumenty psané v daném časovém období.

Střed je tvořen kombinací sady slov, které jsou statisticky důležité vzhledem ke shluku dokumentů. Takovýto střed lze použít jednak na klasifikaci dokumentů, tak na identifikování charakteristických vět ve shluku.

MEAD je používána ke klasifikaci charakteristických vět, extrakci těchto vět a vytvoření výsledného souhrnu. Ohodnocování vět v shluku dokumentů je prováděno pomocí několika parametrů. Vstupem MEAD je předem připravený shluk dokumentů a výstupem je

shluk vytvořený z požadovaného počtu vět, které jsou extrahovány ze vstupu a seřazeny podle původního výskytu ve vstupním shluku dokumentů.

Pro výpočet hodnot vět, z kterých lze pak určit charakteristické věty shluku dokumentů, se používají tři funkce: Středová hodnota („Centroid value“), Poziční hodnota („Positional value“) a Překrytí první věty („First-Sentece overlap“).

Implementace MEAD:

Implementace této metody je podle článku [1].

D reprezentuje shluk dokumentů a $|D|$ reprezentuje počet dokumentů v *D*. Pro výpočet tří funkcí je nejdříve nutné rozdělit shluk dokumentů *D* na jednotlivé dokumenty D_k, kde *k* reprezentuje *k*-tý dokument. Dále $S_{i,k}$ je *i*-tá věta dokumentu *k*, která obsahuje slova w_j (j reprezentuje j-té slovo v dané větě $S_{i,k}$). Je nutné spočítat každému slovu ze shluku dokumentů váhy TF(w_i)*IDF(w_i), kde *i* reprezentuje *i*-té slovo ze všech slov ve shluku dokumentů.

Tedy: TF(w_i) = $f(w_i, D)$, což je tzv. frekvence termu, tedy hodnota počtu výskytů slova w_i v celém shluku dokumentů.

IDF(w_i) = log($\frac{počet\ dokumentů}{počet\ dokumentů\ obsahující\ slovo\ w_i}$), což je tzv. inverzní frekvence dokumentu a je též počítána přes všechny dokumenty ve shluku dokumentů.

Výpočet tří funkcí:

Středová hodnota:

Středová hodnota věty $S_{i,k}$ je definována jako normalizovaná suma jednotlivých centroidů slov ve větě. Tedy:

$$C'_{i,k} = \sum_{w \in S_{i,k}} \frac{TF(w) * IDF(w)}{|D|} * f(w, S_{i,k})$$

Kde *w* jsou všechny slova věty $S_{i,k}$(tedy i-tá věta z k-tého dokumentu) a $f(w, S_{i,k})$ je četnost slova *w* ve větě $S_{i,k}$.

Po spočítání všech středových hodnot $C'_{i,k}$ k-tého dokumentu se tyto váhy následně normují:

$$C_{i,k} = \frac{C'_{i,k}}{\max_{\hat{i},\hat{k}} C'_{\hat{i},\hat{k}}}$$

Poziční hodnota:

Pro každou větu $S_{i,k}$ z dokumentu D_k se poziční hodnota počítá následně:

$$P_{i,k} = \frac{(n - i + 1)}{n}$$

Kde *n* je počet vět dokumentu D_k

Překrytí první věty:

Tato hodnota je počítána jako hodnota společného výskytu slov ve větách $S_{i,k}$ a $S_{1,k}$, tedy:

$$F_{i,k} = \frac{\langle \overrightarrow{S_{i,k}}, \overrightarrow{S_{1,k}} \rangle}{\langle \overrightarrow{S_{1,k}}, \overrightarrow{S_{1,k}} \rangle}$$

Kde: $\langle \overrightarrow{S_{i,k}}, \overrightarrow{S_{J,k}} \rangle = \sum_{w \epsilon S_{i,k} \cap S_{j,k}} f(w, S_{i,k}) * f(w, S_{j,k})$

Po výpočtu všech třech funkcí, lze vypočítat konečné váhy vět:

$$Váha(S_{i,k}) = w_c C_{i,k} + w_p P_{i,k} + w_f F_{i,k}$$

Kde w_c, w_p a w_f jsou váhy jednotlivých funkcí. Podle článku [3] jsou nastaveny následovně:

$w_c = 3$, v případě celkového počtu vět shluku dokumentů je menší než 200.

$w_c = 4$, v případě celkového počtu vět shluku dokumentů je větší než 200.

$w_p = 2$ a $w_f = 1$

Výstupem algoritmu je požadovaný počet vět vzhledem k počtu vět celého shluku dokumentů, které mají nejvyšší konečné váhy a jsou řazeny podle původního výskytu ve shluku dokumentů.

5.2.3 Metoda založená na NMF a K-means shlukování

Podklady k této metodě jsou čerpány z článku [4].

Vstupem této metody musí být sada dokumentů, které mají stejné téma nebo jsou z daného časového období. Vstup může být uložen v jednom souboru nebo ve více souborech.

Teorie nezáporné maticové faktorizace:

Použitím nezáporné maticové faktorizace (Non-negative matrix factorization – NMF) lze identifikovat části reprezentující data, protože nezáporné omezení metody NMF jsou

kompatibilní s intuitivními pojmy (jejichž kombinace tvoří celek). Také lze efektivně reprezentovat velké množství informací, díky řídké distribuci reprezentované v NMF.

Implementace NMF+K-means:

Implementace byla provedena v programovacím jazyce Python.

Nejdříve je provedeno načtení vstupních dokumentů. Každá věta z jednotlivých dokumentů je načtena zvlášť. V případě citace je předpokládáno, že celá citace je důležitá a proto je načtená celá část s citací jako samostatná věta.

Dále je potřeba vytvořit váhovou matici *A* tvořenou frekvencemi termů. Matice *A* je dána velikostí *m x n*, kde *m* je počet termů vyskytujících se ve všech dokumentech a *n* je počet vět ze všech dokumentů. Tedy element A_{ji} je váhová termová-frekvence termu *j* ve větě *i*.

Váhová termová frekvence je počítáná následovně:

$$A_{ji} = L_{ji} G(j)$$

Kde L_{ji} je lokální váha (tedy počet výskytů) termu *j* ve větě *i*. Zatímco $G(j)$ je globální váha (inverzní frekvence dokumentu – IDF) termu *j* ve všech dokumentů a lze počítat následovně:

$$G(j) = \log(\frac{N}{N(j)})$$

Kde *N* je celkový počet vět ve všech dokumentů a $N(j)$ je počet vět, v kterých se term *j* vyskytuje.

Ve chvíli, kdy je vytvořená váhová matice *A* využijeme metodu K-Means shlukování. V algoritmu byla použita již naprogramovaná metoda K-means za použití knihovny scipy pro programovací jazyk Python.

Teorie k-means shlukování:

K-means shlukování je algoritmus, který rozděluje *n* objektů do *f* shluků. Při shlukování je využívána funkce Euklidovské vzdálenosti s respektováním matice *A*. Tedy lze počítat Euklidovskou vzdálenost následovně:

$$d(A_{*a}, A_{*b}) = \sqrt{\sum_{k=1}^{m} |A_{ka} - A_{kb}|^2}$$

Kde A_{*a} a A_{*b} je *a*-tý sloupcový vektor a *b*-tý sloupcový vektor matice *A*. Tedy dostáváme váhy *a*-té a *b*-té věty.

Dále vytváříme *f* shluků. Počet shluků je v algoritmu zvolen jako desetina celkového počtu vět. Tedy lze po rozdělení matice *A* do *f* shluků reprezentovat tuto matici jako:

$$\{A_{*j}|j = 1, \dots, n\} = \bigcup_{i=1}^{f} \{C^i|l = 1, \dots, s_i\}, C^i \cap C^j \neq \phi, \quad i \neq j$$

Kde C^i je váhová matice *i*-tého shluku tvořeného větami, s_i je počet sloupců shluku C^i, dále *n* je počet vět a *f* je počet shluků. Shluky, které mají méně než 5 vět se přiřadí do $f_{šum}$, což je počet šumových shluků.

V algoritmu přichází na řadu NMF rozklad matic. Nejdříve se odstraní shluky, které jsou považovány za šum. Po té se provede NMF rozklad jednotlivých matic shluků na matice *W* a *H*, tedy:

$$C^i = W^i H^i, \qquad i = 1,2, \dots, f, \qquad f = f - f_{šum}$$

Teorie NMF rozkladu:

Nezáporná maticová faktorizace rozkládá vstupní matici (v tomto případě matic reprezentující shluk C^i) na nezáporný sémantický příznak matice *W* a na nezápornou sémantickou proměnou matice *H* následovně: $C^i \approx W^i H^i$, *i* reprezentuje daný shluk. Jestliže vstupní matice C^i má rozměry m *x* n, pak matice W^i má rozměry *m x r* a matice H^i rozměry *r x n*. Většinou se *r* volí menší než *m* nebo *n*, tedy pak velikost matic *W* a *H* je pak menší než velikost původní matice.

Pro rozklad se používá objektivní funkce, která minimalizuje Euklidovskou vzdálenost mezi jednotlivými sloupci vstupní matice, a tedy dochází k aproximaci: $\tilde{C}^i = W^i H^i$. Jako objektivní funkce je použita Frobeniova norma:

$$\Theta_E(W, H) \equiv \|C - WH\|_F^2 \equiv \sum_{j=1}^{m} \sum_{i=1}^{n} \left(X_{ji} - \sum_{l=1}^{r} W_{jl} H_{li} \right)^2$$

Matice *W* a *H* jsou stále aktualizované až do té doby, kdy $\Theta_E(W, H)$ dokonverguje pod stanovený práh nebo překročí maximální počet opakování. Pravidla pro aktualizování matic *W* a *H*:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T C)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}}$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(C H^T)_{i\alpha}}{(W H H^T)_{i\alpha}}$$

Po té co je každý jednotlivý shluk rozdělen na odpovídající matice *W* a *H* lze sloupcový vektor C_{*j}^i *j*-té věty matice C^i reprezentovat jako lineární kombinací sémantických

rysů vektorů (W^i_{*l}) a sémantické proměnné (H^i_{lj}). Tedy váha *l*-tého sémantického rysu vektoru W^i_{*l} ve větě C^i_{*j} je H^i_{lj}.

$$C^i_{*j} = \sum_{l=1}^{r} H^i_{lj} W^i_{*l}$$

Tedy lze říct, že výhody dvou nezáporných matic W^i *a* H^i jsou:

Všechny sémantické proměnné H^i_{lj} jsou použity pro popis, jak je *j*-tá věta strukturovaná za pomoci sémantických rysů. W^i *a* H^i jsou reprezentovány odděleně. Lze říci, že dává mnohem větší smysl, když je přiřazení malé podmnožiny z velké řady témat W^i_{*l}, než jen jedno téma nebo všechna témata. V každém sémantickém rysu W^i_{*l} jsou sémanticky spřízněné termy shlukovány dohromady právě za pomoci NMF. Navíc k shlukování sémanticky spřízněných termů dohromady do sémantických rysů používá NMF kontext tak, aby byly rozlišeny více-významové vztahy stejných termů.

Dále se určí, kolik vět se bude vybírat z jednotlivých shluků a to následujícím způsobem:

$$S^i_{výběr} = \left[k \times \frac{s_i}{N'}\right]$$

Kde *k* je celkový počet vět, který má být zahrnut v souhrnu, s_i je počet vět v daném shluku a N' je celkový počet vět bez šumových vět.

V algoritmu bude zapotřebí funkce kosinové podobnosti mezi sémantickými rysy vektoru W^i_{*j} a tématu T a tato funkce je vyjádřena následovně:

$$sim\left(W^i_{*j}, T\right) = \frac{W^i_{*j} T}{\left|W^i_{*j}\right| \times |T|} = \frac{\sum_{k=1}^{m} W^i_{kj} \times T_k}{\sqrt{\sum_{k=1}^{m} {W^i_{kj}}^2} \times \sqrt{\sum_{k=1}^{m} {T_k}^2}} \quad , T = (T_1, \ldots, T_n)$$

Kde *T* je vektor tématu, který je tvořen TF-IDF vahami jednotlivých termů vyskytujících se ve všech dokumentech.

Průběh algoritmu v několika krocích:

1) Veškeré načtené dokumenty se rozloží do jednotlivých vět a určí se *k,* které reprezentuje počet vět, který bude vybrán do výsledného souhrnu.
2) Vytvoří se váhová matice *A* z načtených vět podle výše uvedeného postupu.
3) Za použití k-means shlukování se vytvoří *f* shluků (*i*-tý shluk reprezentován maticí C^i) z matice *A* a odstraní se shluky, které jsou považovány za šum (shluky, které obsahují méně než 5 vět). Postup uveden výše.
4) Ke každému shluku reprezentovaného maticí C^i se vytvoří za použití NMF metody matice W^i a H^i (*i*-tého shluku). Postup uveden výše.
5) Zvolí se počet vět, které budou vybrány z daných shluků $S^i_{výběr}$ (*i*-tý shluk).

6) Opakují se následující kroky pro každý shluk C^i:
    a. Výběr $p = \arg\max_{1 \le j \le r}\{sim(W^i_{*j}, T)\}$, kde vztah pro funkci na výpočet kosinové podobnosti $sim(W^i_{*j}, T)$ je uveden výše.
    b. Výběr $q = \arg\max_{1 \le j \le s_i}\{H_{pj}\}$
    c. Věta, která odpovídá C^i_{*q} je vložena do výsledného souhrnu.
    d. Kroky od a. do c. jsou opakovány tak dlouho, dokud není vybráno $S^i_{výběr}$ vět (i-tého shluku). Věty, které již byly vybrány se vynechávají.

6 Ohodnocení výsledků použitých sumarizačních metod

6.1 Provedení ohodnocení sumarizačních metod se softwarovým balíčkem ROUGE

Pro ohodnocení souhrnů získaných z automatických sumarizačních metod je použita automatická metoda ROUGE („Recall-Oriented Understudy for Gisting Evaluation“). Softwarový balíček ROUGE [21] ohodnocuje kvalitu souhrnů (vytvořených automatickými metodami) tím, že jsou korelovány s (ideálními) souhrny, které vytvářel člověk. Tedy pro získání výsledků je zapotřebí mít ideální souhrny (manuálně vytvořené).

6.2 Možné metody ohodnocení softwarového balíčku ROUGE

Popis metod ohodnocení softwarového balíčku ROUGE vychází z článku [6].

Rouge-N: N-gramové statistiky opakovaných výskytů:

Obecně je Rouge-N *n*-gramové odvolávání mezi výslednými souhrny a referenčními (ideálními) souhrny. Rouge-N se počítá následovně:

$$\frac{\sum_{S\in\{referenční\ souhrny\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{referenční\ souhrny\}}\sum_{gram_n\in S} Count(gram_n)}$$

Kde *n* představuje délku *n*-gramů, $gram_n$, a $Count_{match}(gram_n)$ je maximální hodnota opakovaného výskytu *n*-gramů ve výsledném souhrnu a v sadě referenčních souhrnů.

Při vzrůstajícím počtu referenčních souhrnů se zvyšuje i hodnota počtu *n*-gramů (ve jmenovateli Rouge-N vzorce). Tedy lze jednoduše do této výsledné hodnoty integrovat více referenčních souhrnů. Čitatel sumy (přes všechny referenční souhrny) dává efektivně více váhy na souhlasné n-gramy výskytů ve více referenčních souhrnech. Tedy výsledný souhrn, který obsahuje slova shodná se slovy ve více referenčních souhrnech, je zvýhodněný Rouge-N mírou.

Rouge-L: nejdelší společná pod-sekvence („Longest Common Subsequence“-LCS):

Dány dvě sekvence *X* a *Y*, nejdelší společná pod-sekvence (LCS) *X* a *Y* je společná pod-sekvence o nejdelší délce. Pro aplikování LCS v poli ohodnocování souhrnů je věta ze souhrnu vnímána jako sekvence slov a na základě LCS míry pak Rouge-L počítá poměr mezi délkou LCS dvou souhrnů (výsledný a referenční souhrn) a délkou referenčního souhrnu. Jedna z výhod používání LCS je ta, že nevyžaduje po sobě jdoucí shody, ale sekvenční shody, které odráží úroveň slovosledu věty (jako u N-gramů). Další výhodou je, že automaticky zahrnuje *n*-gramy nejdelší v sekvenci a proto není zapotřebí žádná předdefinovaná délka n-gramů.

Rouge-W: vážená nejdelší společná pod-sekvence („Weighted Longest Common Subsequence“):

Základní LCS také obsahuje problém s rozlišením více LCS různých prostorových vztahů v rámci jejich propojených sekvencí. Vážená nejdelší společná pod-sekvence vylepšuje základní LCS s podporou po sobě jdoucími shodami. Rouge-W lze efektivně počítat za použití dynamického programování.

Rouge-S: statistika opakovaně přeskočených bigramů („Skip-Bigram Co-Occurrence Statistics“):

Přeskočením bigramu je myšlen libovolný pár jakýchkoliv slov v jejich pořadí ve větě, za možnosti jakékoliv délky mezery. Míra Rouge-S je překrytí poměru přeskočených bigramů mezi výsledným souhrnem a sadou referenčních souhrnů. Při srovnání přeskočených bigramů s LCS lze poukázat, že přeskočené bigramy počítají všechny odpovídající dvojice slov v pořadí, zatímco LCS počítá pouze nejdelší společnou pod-sekvenci. Aby se snížila možnost výskytů falešných shod, lze nastavit limit na maximální délku skoku (mezery), tzv. d_{skip}, mezi dvěma slovy v pořadí, které mohou tvořit přeskočený bigram. Rouge-S s maximální délkou skoku (mezery) *N* se nazývá Rouge-SN.

Rouge-SU: rozšíření Rouge-S:

Potencionální problém Rouge-S je takový, že nedokáže ohodnotit větu z výsledného souhrnu, pokud věta nemá žádné páry opakovaných výskytů slov s referenčními větami. Pro přizpůsobení tomuto problému se Rouge-S rozšíří o přidaný unigram jako čítací jednotku. Rozšířená verze se nazývá Rouge-SU.

6.3 Vyhodnocení výsledků softwarového balíčku ROUGE

Všechny výsledky implementovaných automatických sumarizačních metod pro sumarizaci multi-dokumentů (vstupy se stejným tématem) byly ohodnocené softwarovým balíčkem ROUGE [21]. Pro ohodnocení výsledků byly použity metody ohodnocení: ROUGE-2, ROUGE-SU a ROUGE-W (popsané v části 6.2).

Jelikož výsledky získané z metod softwarového balíčku ROUGE nejsou v požadovaném tvaru, musí být provedeno ještě vyhodnocení do konečného tvaru. Tento poslední krok lze provést za pomoci skriptu rouge2csv získaného z [22]. Díky tomuto skriptu jsou konečné výsledky reprezentovány hodnotami přesnosti, úplnosti a f-skóre daných metod. Tyto hodnoty popsány v kapitole 4.3.2.

7 Výsledky

7.1 Aplikování automatických sumarizačních metod a jejich ohodnocení

Vybrané sumarizační metody (viz kapitola 5.2) byly použity k sumarizaci novinových článků. Všechny vybrané metody předpokládají, že všechny články na vstupu dané metody mají stejné téma nebo jsou z daného časového období.

Vstupní články pro první část jsou data získané z ACL (Asociace počítačové lingvistiky - „Association for Computational Linguistics“) [16]. ACL je mezinárodní vědecké a profesní sdružení lidí zabývajících se problémy, které se týkají přirozeného jazyka a počítačů. Pořádá každoročně letní konference v místech významných výzkumných pracovišť počítačové lingvistiky. Je vydáván odborný časopis (Computational Linguistics), který slouží k prezentování výsledků výzkumu v oblastech počítačové lingvistiky a počítačového zpracování přirozeného jazyka.

První část (viz kapitola 7.2) výsledků tvořená sumarizací článků (se stejnými tématy) je získána aplikováním všech uvedených automatických sumarizačních metod a následného ohodnocení za využití softwarového balíčku ROUGE. Z těchto výsledků lze pak učinit závěr (viz kapitola 7.4).

Druhá část (viz kapitola 7.3) výsledků je věnována sumarizaci článků (z daných časových období) metodou NMF a K-means shlukování. V tomto případě nemohlo být provedeno ohodnocení. Jelikož softwarový balíček ROUGE využívá k ohodnocení souhrny (které jsou vytvářeny manuálně) a proto jej nelze použít.

7.2 Ohodnocené výsledky (shrnuté) automatických sumarizačních metod tvořené z článků (se stejnými tématy)

Výsledné souhrny byly získány aplikováním vybraných automatických sumarizačních metod (LSA, MEAD a NMF+K-means). Patnáct souhrnů (na různá témata) je vytvořených ze vstupů po deseti článcích (dokumentech). Ukázka těchto článků viz kapitola Příloha A. Jeden výsledný souhrn je pro přehled uveden viz kapitola Příloha B, ostatní lze nalézt na přiloženém CD. Dále uváděné výsledky jsou ohodnocení výsledných souhrnů. Vstupní články lze nalézt na přiloženém CD.

Uvedené výsledky jsou zprůměrované. Byly rozděleny do tabulek podle způsobu ohodnocení ROUGE. V každé tabulce jsou k danému typu ohodnocení zprůměrovány f-skóre hodnoty všech metod vždy pro každé téma zvlášť. Průměrované hodnoty vychází z výsledků: Tabulka 1 až Tabulka 15 (viz kapitola Příloha C).

Tabulka s průměry hodnot f-skóre v případě použití ROUGE-2 ohodnocení:

	Metody			
	LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
téma 1	0.1263	0.1483	0.0974	**0.1492**
téma 2	0.2070	**0.2164**	0.0768	0.1340
téma 3	0.1409	0.1603	0.0794	**0.2044**
téma 4	0.1475	0.1523	0.0527	**0.1672**
téma 5	0.1348	0.1170	0.1519	**0.1648**
téma 6	0.0540	0.0947	0.0841	**0.1725**
téma 7	0.0834	0.0819	0.0464	**0.0970**
téma 8	0.0970	0.1622	0.0991	**0.2635**
téma 9	0.0500	0.1570	0.0639	**0.1823**
téma 10	0.0872	0.1265	0.0838	**0.1873**
téma 11	0.0618	**0.2031**	0.0965	0.2014
téma 12	0.1227	0.1694	0.1715	**0.2425**
téma 13	0.1060	0.1543	0.0314	**0.1785**
téma 14	0.0936	0.1280	0.1646	**0.1898**
téma 15	0.1473	0.2328	0.0479	**0.2642**
Celkový průměr	0.1106	0.1536	0.0898	**0.1866**

Tabulka 1: Průměry hodnot f-skóre (k určitým tématům) v případě použití ROUGE-2 ohodnocení

Tabulka s průměry hodnot f-skóre v případě použití ROUGE-SU ohodnocení:

	Metody			
	LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
téma 1	0.1068	**0.1721**	0.1075	0.1719
téma 2	0.1753	**0.2121**	0.1399	0.1331
téma 3	**0.1725**	0.1602	0.1475	0.1182
téma 4	**0.1757**	0.1690	0.1071	0.1643
téma 5	0.1473	0.1114	**0.1858**	0.1018
téma 6	0.0892	0.1303	0.1017	**0.1717**
téma 7	**0.1278**	0.0843	0.1246	0.0683
téma 8	0.1531	**0.2053**	0.1169	0.1993
téma 9	0.1048	0.1819	0.1030	**0.2070**
téma 10	0.1583	**0.1777**	0.1405	0.1759
téma 11	0.1181	**0.1934**	0.1276	0.1360
téma 12	0.1555	**0.2144**	0.1570	0.1796
téma 13	0.1407	0.1226	0.1059	**0.1409**
téma 14	0.1486	0.0830	**0.1498**	0.1095
téma 15	0.1685	**0.2070**	0.1158	0.1630
Celkový průměr	0.1428	**0.1617**	0.1287	0.1494

Tabulka 2: Průměry hodnot f-skóre (k určitým tématům) v případě použití ROUGE-SU ohodnocení

Tabulka s průměry hodnot f-skóre v případě použití ROUGE-W ohodnocení:

	Metody			
	LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
téma 1	0.0978	0.1131	0.0867	**0.1175**
téma 2	0.1421	0.1403	0.0947	**0.1183**
téma 3	0.1023	0.1085	0.0789	**0.1104**
téma 4	0.1108	0.1284	0.0857	**0.1366**
téma 5	0.1240	0.1166	0.1220	**0.1300**
téma 6	0.0789	0.1023	0.0742	**0.1195**
téma 7	**0.1052**	0.0971	0.0924	0.0996
téma 8	0.0879	0.1037	0.0721	**0.1375**
téma 9	0.0712	0.1115	0.0766	**0.1221**
téma 10	0.0972	0.1124	0.0877	**0.1352**
téma 11	0.0991	**0.1574**	0.1077	0.1544
téma 12	0.1185	0.1411	0.1074	**0.1680**
téma 13	0.1258	0.1397	0.0968	**0.1466**
téma 14	0.1053	0.0933	0.1262	**0.1407**
téma 15	0.1361	0.1635	0.0927	**0.1848**
Celkový průměr	0.1068	0.1219	0.0934	**0.1347**

Tabulka 3: Průměry hodnot f-skóre (k určitým tématům) v případě použití ROUGE-W ohodnocení

7.3 Výsledky NMF+Kmeans metody vytvářené z článků (z daného časového období)

Tyto výsledky nelze ohodnotit, a proto byla vybrána pouze jedna potencionálně nejlepší automatická sumarizační metoda, která je zde aplikována. Vstupem bylo několik článků k daným časovým obdobím (lze je nalézt na přiloženém CD), celkem je získáno 6 souhrnů z časových období. Dále je uveden jeden souhrn, další jsou k nalezení (kvůli obsáhlosti) na přiloženém CD. Tyto výsledky nebylo možné ohodnotit, kvůli ideálním (manuálním) souhrnům, které nejsou k těmto článkům k dispozici.

Výsledný souhrn z časového období Listopad 2013 z novinových článkům (s tématem Ukrajina):

> EU si pro ukrajinské přidružení stanovila řadu podmínek, z nichž většinu Kyjev splnil.
> Chystanou smlouvu Ashtonová označila za nejambicióznější dohodu, kterou kdy unie partnerské zemi nabídla.
> Moskva slova evropských politiků o svém nátlaku na Kyjev odmítá a hovoří naopak o vydírání ze strany Evropské unie.
> "Silnější vztahy s EU nejsou na úkor vztahů mezi našimi východními partnery a jejich dalšími sousedy, jako je Rusko," "zdůraznili v prohlášení lídři unie.

EU to odmítá právě s poukazem na bilaterální charakter smlouvy ; konzultace se třetí zemí by byly možné zřejmě o konkrétních tématech, na nichž se Ukrajina a unie dohodnou.
Evropští politici neskrývali zklamání a lídři unie José Barroso a Herman Van Rompuy v pondělí jednoznačně odsoudili "vnější tlak" kterým na svého západního souseda působí Rusko.
Podle diplomata EU je unie připravena dál s Kyjevem vývoj konzultovat, text smlouvu už měnit nepůjde.
Unie by však případně mohla finančně pomoci Ukrajině s krátkodobými náklady modernizace, které by z podpisu smlouvy plynuly.
Unie ale nehovoří o tom, že by mohla Kyjevu nahradit ztráty, které utrpí jeho obchod s Ruskou federací.
Lídři unie José Barroso a Herman Van Rompuy v pondělí jednoznačně odsoudili "vnější tlak" kterým údajně na svého západního souseda působí Rusko.
Brusel přitom podle podmínek asociační smlouvy nabízel Kyjevu technickou pomoc ve výši 610 milionů eur.
"Evropská unie zformulovala ultimativně požadavek, který bylo možné uspokojit až jako důsledek našeho sblížení," "řekl Juščenko.
Chce, aby tyto problémy řešila Evropská unie společně s Ruskem, "řekla Grybauskaitéová novinářům.
Od nápadu s třístrannou komisí na čtvrteční večeři podle Grybauskaitéové Janukovyče odrazovali zejména noví členové EU ze střední a východní Evropy, kteří se členy unie stali v roce 2004.

7.4 Vyhodnocení výsledků kapitoly 7.2

Veškeré výsledky jsou porovnány a zhodnoceny vzhledem ke kvalitě automatické sumarizační metody. Zde je uvedeno veškeré možné porovnání výsledků získané z tabulek (Tabulka 1 až Tabulka 18). Tabulky s výsledky jsou získány ze softwarového balíčku ROUGE, který porovnává výsledné sumarizace (z automatických sumarizačních metod) s výslednými „ideálními“ sumarizacemi (vytvořenými manuálně).

Uváděné výsledky jsou pro ROUGE-2 (nejpoužívanější), ROUGE-SU a ROUGE-W (získání těchto hodnot je teoreticky popsáno viz kapitola 6.2), kde pro každou možnost jsou hodnoty přesnosti, úplnosti a f-skóre (teoretický výpočet těchto hodnot je popsán viz kapitola 4.3.2). Z teoretického výpočtu hodnot přesnosti, úplnosti a f-skóre je patrné, že pro ohodnocení metody stačí použít f-skóre, které je kombinací přesnosti a úplnosti.

7.4.1 Porovnávání metod vzhledem k ohodnocení podle ROUGE-2

Z výsledků uváděných v Tabulka 1 získané z Tabulka 4 až Tabulka 18 (viz kapitola Příloha C) lze určit nejlepší metodu. Podle ROUGE-2 porovnání dosahuje nejlepších výsledků metoda založená na NMF a K-means shlukování.

7.4.2 Porovnávání metod vzhledem k ohodnocení podle ROUGE-SU

Z výsledků uváděných v Tabulka 2 získané z Tabulka 4 až Tabulka 18 (viz kapitola Příloha C) lze určit nejlepší metodu. Podle ROUGE-SU porovnání dosahuje nejlepších výsledků metoda založená na LSA (délka vět).

7.4.3 Porovnávání metod vzhledem k ohodnocení podle ROUGE-W

Z výsledků uváděných v Tabulka 3 získané z Tabulka 4 až Tabulka 18 (viz kapitola Příloha C) lze určit nejlepší metodu. Podle ROUGE-W porovnání dosahuje nejlepších výsledků metoda založená na NMF a K-means shlukování.

7.5 Závěr k výsledkům

Z výsledků lze říci, že nejlepší metoda pro použití v praxi (v případě vstupů se stejnými tématy) je metoda založená na NMF a K-means shlukování. Na druhém místě je metoda založená na Latentní sémantické analýze (LSA+LexRank a LSA-délka vět). Nejhorších výsledků dosáhla metoda založená na váze středů shluků (MEAD). Je zajímavé, že metoda založená na Latentní sémantické analýzy, která byla předělána z jedno-dokumentové metody na multi-dokumentovou metodu, dosáhla lepších výsledků než multi-dokumentová metoda založená na váze středů shluků. Metoda založená na NMF a K-means shlukování a metoda založená na Latentní sémantické analýze dosahuje lepších výsledků pravděpodobně díky využití sémantiky vět. Metoda založená na LSA dosáhla v případě ohodnocení pomocí ROUGE-SU dokonce lepších výsledků než metoda založená na NMF a K-means shlukování. To může být zapříčiněno díky ohodnocení pomocí ROUGE-SU, které používá přeskočené bigramy a přidaný unigram (v procesu porovnávání souhrnů), na rozdíl od ohodnocení pomocí ROUGE-2 a ROUGE-W.

Bohužel výsledné souhrny z článků z daného časového období nelze ohodnotit. Tedy na tyto články byla aplikována pouze jediná automatická sumarizační metoda a to NMF+K-means. To z důvodů dosažení nejlepších výsledků ze všech testovaných metod.

8 Závěr

V této práci jsou v prvé řadě popsány teoretické základy sumarizace dokumentů. Jsou zahrnuty jak manuální, tak automatické sumarizační metody a samozřejmě jejich vlastnosti a částečný popis. Také byly probrány a popsány možnosti ohodnocení úspěšnosti (kvality) výsledných souhrnů. Metody, použitelné k sumarizaci novinových článků jsou zde rozebrány teoreticky, to samozřejmě zahrnuje i popis těchto metod. Největší část práce je zaměřena na metody použitelné pro multi-dokumentovou sumarizaci. Jsou uvedeny výhody a nevýhody těchto metod a je zde detailně seznámeno s určitými automatickými metodami, které jsou nejvíce vhodné pro sumarizování novinových článků. Vybrané automatické metody jsou tři a to Latentní sémantická analýza, metoda založená na váze středů shluků (MEAD) a metoda založená na NMF a K-means shlukování. Tyto metody byly detailně popsány a naprogramovány v programovacím jazyce Python. Programy zpracovávají (sumarizují) shluky dokumentů, které jsou již předem přerozděleny do skupin s daným tématem nebo z daného časového období. Ohodnocení výsledných souhrnů získaných z naprogramovaných metod je důležitá část práce. Toto téma je rozebráno teoreticky a jsou nastíněny možnosti ohodnocování. Aby bylo možné zjistit nejlepší metodu (která byla naprogramována) jsou výsledné souhrny ohodnoceny. To bylo provedeno za použití softwarového balíčku ROUGE, jehož funkce je nastíněna a popsána. K provedení ohodnocení výsledků je důležité mít ideální souhrny (vytvořené člověkem). Ty jsou dostupné pouze pro vstupy (články) s danými tématy. Proto bylo ohodnocení souhrnů provedeno pouze pro případ vstupů se stejnými tématy. Všechny metody, které byly ozkoušeny lze použít jak na vstupy (články) z daného časového období tak na vstupy se stejným tématem. Ze zpracovaných výsledků lze říci, že pro vstupy se stejným tématem je pro použití v praxi nejlepší metoda založená na NMF a K-means shlukování. U vstupů z daného časového období nelze určit, jaká testovaná metoda dosahuje lepších výsledků (díky nedostatečným datům). Také z těchto důvodů byla na tyto vstupy aplikována pouze metoda založená na NMF a K-means shlukování.

Velký zájem o sumarizaci vede stále k dalšímu rozvoji sumarizačních metod ať již automatických, tak i manuálních. V posledních letech byl proveden pokrok nejen ve vytváření nových metod, ale i v zlepšení nebo zkombinování již používaných metod. Podle mého názoru mají budoucnost kombinované metody. Například jedna z metod může být ryze matematická a druhá sémantická. Touto kombinací lze dosáhnout kvalitního souhrnu, který obsahuje související věty. Toto kombinování nemusí být nutně použitelné jen na extraktivní sumarizační metody, ale i na abstraktivní.

Jsem velmi rád o možnost rozšíření znalostí o další sumarizační metody a o obecný pohled na sumarizaci. Práce je zaměřena především na multi-dokumentové extrakční automatické sumarizační metody. Tyto metody lze nadále zdokonalovat. Podle mého názoru se tyto metody budou v nejbližší době rozvíjet více než metody abstraktivní. Ty by měly vytvářet kvalitnější souhrny, tedy mají velký potenciál, ovšem musí obsahovat data o jazyku, v kterém se provádí sumarizace. Popřípadě by bylo zajímavé se těmito metodami zabývat.

Literatura

[1] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś. *Centroid-based summarization of multiple documents***.** Information Processing & Management. 2004, číslo 40, Issue 6, strany 919-938, ISSN 0306-4573, http://dx.doi.org/10.1016/j.ipm.2003.10.006.

[2] Yuan Ding. *A Survey on Multi-Document Summarization***.** University of Pennsylvania: Department of Computer and Information Science. October 2004 Dostupné online z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.2153&rep=rep1&type=pdf.

[3] Dragomir R. Radev, Sasha BlairGoldensohn and Zhu Zhang. *Experiments in Single and MultiDocument Summarization Using MEAD***.** Ann Arbor: University of Michigan. Dostupné online z: http://www.summarization.com/~radev/papers/045.pdf

[4] Sun Park, Ju-Hong Lee, Deok-Hwan Kim and Chan-Min Ahn. *Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization.* Jan van Leeuwen et al. (Eds.): SOFSEM 2007. Dostupné z: http://link.springer.com/chapter/10.1007%2F978-3-540-69507-3_66

[5] Josef Steinberger, Karel Ježek. *Evaluation Measures For Text Summarization***.** 28 vyd. Computing and Informatics. 2009. Plzeň: University of West Bohemia in Pilsen. Dostupné online z: http://www.cai.sk/ojs/index.php/cai/article/download/37/24

[6] Chin-Yew Lin. *Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples Are Enough?*. Proceeding of NTCIR-4, April 2003–June 2004. Dostupné online z: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/OPEN/NTCIR4-OPEN-LinCY.pdf

[7] Karel Ježek , Josef Steinberger. *Sumarizace textů*. Mikulov: DATAKON 2010, 16-19. 10. 2010. Dostupné z: http://textmining.zcu.cz/publications/SumarizDATAKON.pdf

[8] Günes Erkan , Dragomir R. Radev. *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization***.** AI Access Foundation, 2004. Dostupné online z: http://www.aaai.org/Papers/JAIR/Vol22/JAIR-2214.pdf

[9] Megumi Makino and Kazuhide Yamamoto. *Summarization by Analogy: An Example-based Approach for News Articles***.** Nagaoka University of Technology. Dostupné online z: http://aclweb.org/anthology//I/I08/I08-2102.pdf

[10] Masaharu Yoshioka, Makoto Haraguchi. *Multiple News Articles Summarization Based on Event Reference Information*. Proceeding of NTCIR-4, April 2003 – June 2004. Dostupné online z: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-YoshiokaM.pdf

[11] Chia-Wei Wu and Chao-Lin Liu. *Ontology-based Text Summarization for Business News Articles*. National Chengchi University. Dostupné online z: http://www.cs.nccu.edu.tw/~chaolin/papers/wu03.pdf

[12] T. Gruber. *Ontology Definition.* Dostupné online z: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[13] Shannon Johnson. *How to Summarize a Newspaper Article***.** eHow Contributor. 2011. Dostupné z: http://www.ehow.com/how_7706132_summarize-newspaper-article.html

[14] Eduard Hovy and Chin-Yew Lin. *Automated Text Summarization And The Summarist Systém.* University of Southern California. Dostupné z: http://acl.ldc.upenn.edu/X/X98/X98-1026.pdf

[15] *Graphical Methods-Summary***.** Modeling Workshop Project 2002. Dostupné online z: http://modeling.asu.edu/Modeling-pub/Mechanics_curriculum/1-Sci%20Thinking/Resources/U1-GMsummary.pdf

[16] *Association for Computational Linguistics*. Dostupné online z: https://www.aclweb.org/

[17] Eduard Hovy. *Text Summarization.* Dostupné online z: http://www.isi.edu/natural-language/people/hovy/papers/05Handbook-Summ-hovy.pdf

[18] I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, B. Sundheim. *The TIPSTER SUMMAC Text Summarization Evaluation*. In Proceedings of EACL, 1999. Dostupné online z: http://acl.ldc.upenn.edu/E/E99/E99-1011.pdf

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

[20] T. Kudo, Y. Matsumoto. *Japanese dependency analysis using cascaded chunking*. Nara Institute of Science and Technology. Dostupné online z: http://acl.ldc.upenn.edu/W/W02/W02-2016.pdf

[21] *ROUGE: Recall-Oriented Understudy of Gisting Evaluation*. Dostupné online z: http://www.berouge.com/

[22] Kavita Ganesan. *rouge2csv - Script to Interpret ROUGE Scores*. © Kavita Ganesan 2013. Dostupné online z: http://kavita-ganesan.com/software-rouge2csv

[23] Günes Erkan , Dragomir R. Radev. *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.* Ann Arbor: University of Michigan, 2004. Dostupné online z: http://www.aaai.org/Papers/JAIR/Vol22/JAIR-2214.pdf

Dodatky

- Obsah přiloženého CD

- Elektronická podoba diplomové práce
- Zdrojové kódy všech třech implementovaných automatických sumarizačních metod v jazyce Python a soubory, které tyto metody využívají
- Vstupní texty (články o daných tématech) s jejich modely (ideálními souhrny vytvořené manuálně).
- Vstupní texty (články z daných časových období)
- Veškeré výsledné souhrny vytvářené automatickými sumarizačními metodami

Příloha A Originální článek (ukázka)

Zde pro přehled je uvedena jedna sada vstupních článků týkajících se jednoho tématu a výsledky z implementovaných sumarizačních metod k tomuto tématu. Ostatní vstupní sady článků a výsledné automatické sumarizace implementovaných metod jsou k dispozici na přiloženém CD.

Téma č. 9:

Toto téma obsahuje celkem 10 článků jako všechny ostatní témata.

Článek č.1:

Letadlo Air France s 228 na palubě zmizel
Pondělí, 1. června 2009
Let Air France 447 převážející 228 lidí z Ria de Janeira, Brazílie, do Paříže, letiště Charlese de Gaulla, je nezvěstný od brazilského pobřeží. Letadlo mělo přiletět do Paříže v 11:10 SEČ dne 1. června 2009.
Airbus A330-200 byl naposledy slyšet přes rádio v 22:30 místního času (01:30 GMT). Brazilské letectvo potvrzuje zmizení letadla, které zmizelo z radaru přibližně 190 mil (306 km) od brazilského pobřeží. Reuters hlásí, že elektrický zkrat během turbulencí mohl způsobit zřícení letadla do Atlantského oceánu. Letadlo s registrací F-GZCP má celkem 18.870 letových hodin a šlo do služby dne 18. dubna 2005. Jeho poslední kontrola údržby byla 16. dubna 2009.
Plánovaná trať letu 447.
"Neobdrželi jsme žádné zprávy z letu AF447", řeklo Air France v prohlášení. Letadlo vezlo 12 členů posádky a 216 cestujících, z toho 126 mužů, 82 žen, 7 dětí a jednoho novorozence.
V tiskové zprávě Air France uvedlo, že AF447 se dostal do oblasti těžké turbulence v 02:00 GMT, a že letadlo odeslalo automatickou zprávu s oznámením elektrické poruchy v 02:14 UTC.
Pátrací a záchranné operace brazilského letectva se zpočátku soustředily okolo ostrova Fernando de Noronha. Letadlo je pohřešováno tak dlouho, že už by mu nezbývalo žádné palivo, kdyby bylo ve vzduchu.
Francouzský prezident Sarkozy řekl, že vyhlídky nalezení přeživších jsou "velmi malé". Požádal také Ameriku, aby použila jejich sledovací satelity při hledání letadla.
Air France nabízí tři telefonní čísla pro rodiny a přátele cestujících na palubě AF447:
* 0800 800 812 pro volání v rámci Francie
* 0800 881 20 20 pro volání z Brazílie
* +33 1 57 02 10 55, pro volání ze zemí mimo Francii nebo Brazílii.

"Speciálně vyhrazené oblasti" 2. terminálu letiště Charlese de Gaulla jsou používány pro péči o rodiny pasažérů na palubě AF447.

Článek č.2:

Airbus nabízí financování hledání černé skříňky z katastrofy Air France
Čtvrtek, 30. červenec 2009
Airbus oznámil, že je ochoten přispět mezi 12 a 20 miliony eur (asi 16 až 28 milionů dolarů) na financování rozšířeného hledání černé skříňky z letu Air France 447. Tryskáč Airbus A330 spadl do Atlantského oceánu v červnu, zabito bylo všech 228 lidí na palubě.
Generální ředitel Thomas Enders uvedl v prohlášení: "Chceme vědět, co se stalo, protože zlepšit bezpečnost leteckého provozu je naší nejvyšší prioritou. Jsme plně odhodláni podporovat prodloužení hledání významným příspěvkem." I když výrobci letadel normálně poskytují technickou pomoc vyšetřování, požadovaná nestrannost dělá financování vzácné, mluvčí Airbusu Stefan Schaffrath řekl, že iniciativa firmy byla bezprecedentní.
"Je to výjimečná nehoda a výjimečná situace", vysvětlil Schaffrath. Francouzská vyšetřovací agentura BEA požádala o finanční pomoc pro vyhledávání, jak od Airbusu, tak i od Air France. Air France projednává tuto možnost s BEA.

Vyšetřovatelé se již vzdali hledání hlasového záznamníku z pilotní kabiny a černé skřínky s využitím konvenční metody hledání 'audiobzučáků', jejichž baterie by se vybily po 30-40 dnech. Pokračují snahy pomocí citlivého vybavení v rámci úsilí z francouzské námořní lodi, ale pokud se to ukáže jako marné, BEA pak bude shánět finance na další tříměsíční vyhledávání.
Komunikační a hlásící systém letadla (ACARS) byl schopen předat informace o problémech na palubě před havárii. Z údajů ACARS vyplývá, že letadlo nemělo k dispozici důležité hodnoty, včetně rychlosti, což vede k podezření na pilotní statický systém, který dodává různá měření.
Airbus již doporučil, aby jedna složka tohoto systému, pilotní trubice, byla nahrazena v letadlech typu A330. Air France to neudělal v případě havarovaného letounu, i když celá flotila už upravený design má. Tři další případy s podobnými okolnostmi byly zjištěny. Národní přepravní bezpečnostní komise Spojených Států zkoumá dva na americké půdě, zatímco na začátku tohoto měsíce další Air France A330 vybaven novým trubkami se potýkal s podobnou řadou problémů na cestě z Itálie do Francie.

Článek č.3:

Trosky zmizelého letadla letu Air France nalezeny v Atlantiku
Úterý, 2. června 2009
Brazilské vojenské letadlo našlo podle brazilského letectva trosky letounu v Atlantském oceánu, asi 650 kilometrů od severního pobřeží Brazílie.
Má se za to, že pocházejí z letu Air France 447, tryskáče, který zmizel nad Atlantským oceánem 1. června.
Trosky obsahují sedadla z letounu a kovové předměty. Piloti brazilské letecké společnosti TAM dříve řekli, že viděli požár v Atlantiku, ale francouzští představitelé říkají, že žádné známky vraku nebyly nalezeny.
Brazilská letadla se senzory v současné době skenují oblast oceánu, kde byly nalezeny trosky.
"Letadlo brazilských aerolinií TAM údajně vidělo něco hořet v Atlantském oceánu. To byl letoun, který přistál dnes," řekl brazilský viceprezident Jose Alencar.
Let Air France AF447 byl Airbus A330 převážející 216 cestujících a 12 členů posádky, který zmizel poté, co vlétl do hrozné bouřky. Úřady si zatím nejsou jisté, co způsobilo incident, ale jsou pesimistické, co se týče nalezení přeživších.

Článek č.4:

Bylo potvrzeno, že trosky letadla nalezené v Atlantiku jsou ze zmizelého letu Air France.
Úterý, 2. června 2009
Brazilská vláda potvrdila, že trosky letadla nalezené v Atlantickém oceánu 650 kilometrů od pobřeží Brazílie jsou z letu Air France 447.
Brazilský ministr obrany Nelson Jobim řekl, že tam nebyli žádní přeživší.
V úterý bylo nalezeno v Atlantiku jedno sedadlo a záchranná vesta. Francouzská loď dorazila do této oblasti ve stejný den a potvrdila nalezení trosek. Podle CNN by mělo ve středu dorazit na scénu plavidlo brazilského námořnictva.
Hlavním cílem hledající posádky bude najít hlas z kokpitu a černou skříňku, která by pomohla zjistit, co způsobilo pád.
Nicméně brazilský ministr obrany řekl, že to může být těžké najít kvůli velké hloubce oceánu v oblasti, řekl, že "by to mohlo být v hloubce 2000 nebo 3000 m [od 6500 do 9800 stop] v této oblasti z oceánu.
"Letem Air France 447 byl Airbus A330 na cestě z Rio de Janeira do Paříže, Francie, s 228 lidmi na palubě, když zmizel z radarových obrazovek po vstupu do oblasti silné turbulence.

Článek č.5:

Ocas z tryskáče Air France byl nalezen v Atlantickém oceánu
Pondělí, 8. června 2009

Vyhledávací tým z Brazílie nalezl kousek zadní části tryskáče Air France letu 447, který se zřítil do Atlantického oceánu 1. června. Brazilské ozbrojené síly zveřejnili fotografie potápěčů u nalezeného ocasu, který je pomalován barvami Air France - modrou, červenou a bílou.
Mezitím Francie řekla, že odeslala jadernou ponorku, aby hledala černou skříňku letu tryskáče, která by mohla odhalit informace o tom, jak a kdy se letadlo zřítilo.
Šestnáct těl bylo nalezeno ve vodách oceánu minulý víkend.
Brazilský prezident Luz Inácio Lula da Silva řekl, že "děláme všechno pro to [...], abychom nalezli, pokud možno, všechna těla, protože víme, jak moc pro rodinu znamená dostat své ztracené milované."

Článek č.6:

V Atlantiku pokračuje pátrání po zmizelém letadle společnosti Air France
Čtvrtek, 2. června 2009
Letadla a lodě francouzské, americké a brazilské armády v pondělí pátraly v Atlantiku poté, co nad oceánem zmizelo letadlo společnosti Air France s 228 lidmi na palubě.
Pátrací týmy z Brazílie se zaměřily hlavně na oblast severně od ostrova Fernando de Noronha, který leží přibližně 200 mil od brazilského pobřeží. Francouzská armáda byla o několik set mil dále a prohledávala oceán v blízkosti Kapverdských ostrovů.
Na pomoc při pátrání byly vyslány také letouny ze Senegalu a ze Španělska. Francouzské úřady rovněž požádaly Spojené státy o data ze satelitů, která by jim při hledání pomohla.
"Chceme najít místo, kde bylo letadlo naposled ve spojení, což je asi 1200 km severovýchodně od Natalu [ležícím v Brazílii]," řekl plukovník Jorge Amaral, mluvčí brazilského letectva.
Francouzský prezident Nicolas Sarkozy prohlásil, že je pesimistický ohledně nálezu přeživších osob.
"Toto je katastrofa, jakou Air France nikdy předtím nezažila. Řekl jsem jim pravdu: vyhlídky na to, že se najde někdo, kdo přežil, jsou velmi špatné," řekl prezident novinářům na pařížském letišti Charlese de Gaulla ve Francii. Také Pentagon povolal letectvo, aby pomohlo při pátrání po letadle.
Podle mluvčí společnosti Air France odletěl Airbus A330 z Ria de Janeira v neděli večer se dvanácti členy posádky a 216 pasažéry na palubě. Když letadlo asi čtyři hodiny po odletu zmizelo, letělo normálně rychlostí 840 kilometrů za hodinu ve výšce asi 10700 metrů nad hladinou moře. Když bylo naposledy v radarovém spojení, nebyly zřejmé žádné problémy.
Společnost Air France uvedla, že letadlo "prolétalo bouřkovou oblastí se silnými turbulencemi" v asi 23:00 místního času (02.14 GMT). Přibližně čtrnáct minut později byla vyslána automatická zpráva, že v kabině poklesl tlak a že selhal elektrický systém letadla.
Jestliže se nenajdou žádní přeživší, pak to bude nejhorší letecká katastrofa od havárie letadla na letu 587 společnosti American Airlines v listopadu 2001.

Článek č.7:

Těla pasažérů a vrak letadla z letu 447 společnosti Air France zřejmě nalezeny
Sobota, 6. června 2009
Brazilské letectvo lokalizovalo těla pasažérů a vrak letadla z letu 447 společnosti Air France v Atlantickém oceánu. Nejméně dvě těla byla dosud vyzvednuta.
Plukovník brazilského letectva Jorge Amaral potvrdil, že pátrací týmy vyzvedly dvě mužská těla a vrak včetně sedadla, kufříku, dýchací masky, přenosného počítače a část letenky patřící jednomu z cestujících. Dodal, že na místo nálezu už cestují odborníci na lidské ostatky, aby je prozkoumali. Všech 228 lidí na palubě letadla se pokládá za mrtvé.
Včera se ukázalo, že dříve nalezené trosky z míst pravděpodobného zřícení letadla jsou odpadky z projíždějících lodí. Našla se dřevěná nákladní paleta a dvě bóje, ale zřejmě pocházejí z lodi.
Letadlo na letu 447 společnosti Air France letící z Ria de Janeira v Brazílii na letiště Charlese de Gaulla v Paříži zmizelo nad Atlantikem 1. června. Rádiové spojení s Airbusem A330-220 bylo naposledy navázáno 1. června ve 22:30 místního času (01:30 UTC) a letadlo zmizelo z radarů asi 306 km od brazilského pobřeží.

Příčina havárie není zatím známa. Dnes francouzské úřady oznámily, že letadlo z letu 447 krátce před zmizením vyslalo 24 chybových zpráv. Také informovaly, že autopilot byl v letadle vypnutý, ale že z chybových zpráv není možné zjistit, proč tomu tak bylo. Havárie letadla se považuje za nejhorší letecké neštěstí od roku 2001.

Článek č.8:

Podle leteckých odborníků se letadlo společnosti Air France na letu 447 rozlomilo ve vzduchu
Čtvrtek, 18. června 2009
Pitvy mrtvých těl z letu Air France 447, jehož letadlo se začátkem měsíce zřítilo do vod Atlantického oceánu, odhalily zlomeniny nohou, což by mohlo znamenat, že se letadlo rozlomilo na kusy během letu.
Mluvčí vyšetřujících brazilských lékařů ve středu uvedl, že pitvy určité části mrtvých těl, které se dosud z vraku letadla našly, odhalily fraktury.
"Můžeme říci, že máme trochu méně nejistoty, takže vládne trochu více optimismu," řekl Paul-Louis Arslanian, ředitel Francouzské agentury pro vyšetřování leteckých neštěstí BEA. Nicméně dodal, že "je předčasné teď říkat, co se stalo."
Frank Ciacco, bývalý forenzní expert Americké národní rady pro dopravní bezpečnost, prohlásil, že "pokud vidíte neporušená těla s četnými zlomeninami - rukou, nohou, kyčlů, tak je to typicky dobrý indikátor, že se letadlo rozpadlo za letu. Obzvláště když také vidíte velké kusy letadla."
Let společnosti Air France 447 byl Airbus A330, který letěl z Ria de Janeira v Brazílii do Paříže ve Francii a 1. června zmizel nad Atlantickým oceánem. Při vyšetřování se zatím našlo několik částí letadla a nějaká mrtvá těla cestujících, ale stále se hledá záznamník letových dat a záznamník komunikace v kabině pilotů, které by mohly obsahovat důležité informace o tom, jak se nehoda přesně udála.

Článek č.9:

Falešná stopa z letu společnosti Air France - trosky nejsou z letadla, pátrání pokračuje
Pátek, 5. června 2009
Pátrací týmy z Francie a Brazílie potvrdily, že nenalezly žádné trosky z letu 447 společnosti Air France. Airbus A330 s 228 lidmi na palubě zmizel z radarů 1. června nad Atlantickým oceánem.
Brazilské letectvo původně 2. června informovalo, že nalezlo trosky letadla. Vyzvedlo dřevenou nákladní paletu a dvě bóje, ale nyní se předpokládá, že pocházejí z lodi. Letadlo žádné dřevěné palety nevezlo. 2. června byla vydána zpráva, kterou brazilské letectvo potvrdilo, že se ve vodě našlo jedno sedadlo a jedna záchranná vesta.
Mastná skvrna na hladině, o které se myslelo, že je z letadla, se nakonec ukázala pocházet z jiného zdroje, nejspíše z lodi. Přítomnost pohonných hmot se pokládala za důkaz, že na palubě letadla nedošlo k požáru nebo výbuchu. Španělský list El Mundo cituje pilota společnosti Air Comet, který letěl nad Atlantikem v době zřícení letadla, že zažil "silné vyšlehnutí bílého světla", které padalo z nebe a rozpadlo se na kusy. Hrozba bombovým útokem, která se později ukázala planou, byla také hlášena na jiném letu společnosti Air France z Buenos Aires do Paříže 27. května, jen několik dní před ztracením letu 447.
Francouzský úřad pro vyšetřování a analýzu bezpečnosti civilního letectví (BEA) vede vyšetřování a zatím zjistil, že v oblasti byly bouřky a že série automatických signálů z letadla oznamovala nestálou rychlost a řetězec selhání systémů.
Ve 23:10 se vypnul autopilot a nejméně jeden počítač se přepnul do režimu napájení ze záložního zdroje. Úřad má signály toho, že se poškodil kontrolní letový systém a po něm měřiče rychlosti, směru a výšky a jiné kontrolní systémy. Kontrolní systém hlavního počítačového systému a spoilery na křídlech se také rozbily. Čtyři minuty po začátku problémů oznamovala poslední zpráva úplný výpadek elektrického proudu a ztrátu tlaku v kabině pilotů.
Mluvčí francouzských ozbrojených sil Christophe Prazuck uvedl, že "každý teď pochybuje o všem a zatím nemáme nejmenší náznak odpovědí." Francouzský deník Le Monde napsal, že rychlost letadla byla příliš nízká, na což BEA zareagoval varováním proti "překotným interpretacím a spekulacím"

ohledně příčiny havárie. Airbus připomněl všem pilotům letadel A330 správné postupy za krajně nepříznivého počasí.
Francie vyslala na místo výzkumnou loď "Pourquoi Pas?" a ta je teď na cestě se svými miniponorkami, aby pomohla při hledání a vyzvedávání trosek. Stále probíhá pátrání po zařízeních pro záznam komunikace v kabině pilotů a letových údajů, tzv. černých skříněk. Brazilské úřady už dříve uvedly, že by skříňky mohly být na dně této části oceánu v hloubce 2000 až 3000 metrů.

Článek č.10:

Černé skříňky z letu 447 společnosti Air France byly lokalizovány
Čtvrtek, 6. května 2010
Letadlo A330-200 F-GZCP společnosti Air France přistálo na letišti Charlese de Gaulla v Paříži 28. března 2007. A toto letadlo bylo zničeno během letu 447 společnosti Air France, když se zřítilo do Atlantického oceánu a zahynulo tak všech 228 lidí na palubě.
Zařízení pro nahrávání letových dat a zařízení pro nahrávání hlasů z kokpitu neboli "černé skříňky" z letadla společnosti Air France, které se 1. června loňského roku zřítilo do Atlantického oceánu, byla podle čtvrtečního vyjádření francouzského představitele lokalizována v okruhu asi pěti čtverečních kilometrů.
Francouzská vláda a vojenští představitelé nabádají k opatrnosti, neboť není zaručeno, že se černé skříňky najdou. Mluvčí francouzského námořnictva Hugues du Plessis d'Argentre řekl agentuře AFP: "Je to jako hledat krabici od bot v prostoru velkém jako plocha Paříže, v hloubce 3000 m a v terénu rozeklaném jako Alpy."
Airbus A330-200, na jehož palubě bylo celkem 228 lidí - 216 pasažérů a 12 členů posádky, se za špatného počasí zřítil do Atlantického oceánu. K havárii pravděpodobně přispěly Pitotovy trubice měřící rychlost letadla. Ale vlastní příčina neštěstí není dosud zjištěna.
Hledání je nyní ve své třetí fázi, která začala 30. března až 1. dubna 2010 a měla původně trvat 30 dní. Avšak 4. května bylo hledání prodlouženo do 25. května.

Příloha B Souhrn získaný z implementovaných automatických sumarizačních metod (ukázka)

Zde pro přehled je uvedena jedna sada výstupů z implementovaných automatických sumarizačních metod. Výsledné sumarizace jsou vytvořeny ze 17 vět vybraných z originální sady článků. Výsledky odpovídají přiložené sadě vstupů, tedy téma č.9. Podle výsledků ze softwarového balíčku ROUGE vychází nejlépe automatická sumarizační metoda NMF+K-means.

Téma č. 9:

LSA+LexRank:

Francouzský prezident Sarkozy řekl, že vyhlídky nalezení přeživších jsou "velmi malé" .
"Je to výjimečná nehoda a výjimečná situace" vysvětlil Schaffrath.
Pokračují snahy pomocí citlivého vybavení v rámci úsilí z francouzské námořní lodi, ale pokud se to ukáže jako marné, BEA pak bude shánět finance na další tříměsíční vyhledávání.
Tři další případy s podobnými okolnostmi byly zjištěny.
Trosky obsahují sedadla z letounu a kovové předměty.
Úřady si zatím nejsou jisté, co způsobilo incident, ale jsou pesimistické, co se týče nalezení přeživších.
Brazilský ministr obrany Nelson Jobim řekl, že tam nebyli žádní přeživší.
V úterý bylo nalezeno v Atlantiku jedno sedadlo a záchranná vesta.
Brazilský prezident Luz Inácio Lula da Silva řekl, že "děláme všechno pro to [...], abychom nalezli, pokud možno, všechna těla, protože víme, jak moc pro rodinu znamená dostat své ztracené milované."
Na pomoc při pátrání byly vyslány také letouny ze Senegalu a ze Španělska.
Francouzské úřady rovněž požádaly Spojené státy o data ze satelitů, která by jim při hledání pomohla.
Francouzský prezident Nicolas Sarkozy prohlásil, že je pesimistický ohledně nálezu přeživších osob.
Také Pentagon povolal letectvo, aby pomohlo při pátrání po letadle.
GMT).
Nejméně dvě těla byla dosud vyzvednuta.
Dodal, že na místo nálezu už cestují odborníci na lidské ostatky, aby je prozkoumali.
Našla se dřevěná nákladní paleta a dvě bóje, ale zřejmě pocházejí z lodi.

LSA-délka vět:

Letadlo mělo přiletět do Paříže v 11: 0 SEČ dne 1. června 2009.
Požádal také Ameriku, aby použila jejich sledovací satelity při hledání letadla.
Pokračují snahy pomocí citlivého vybavení v rámci úsilí z francouzské námořní lodi, ale pokud se to ukáže jako marné, BEA pak bude shánět finance na další tříměsíční vyhledávání.
Národní přepravní bezpečnostní komise Spojených Států zkoumá dva na americké půdě, zatímco na začátku tohoto měsíce další Air France A330 vybaven novým trubkami se potýkal s podobnou řadou problémů na cestě z Itálie do Francie.
Trosky obsahují sedadla z letounu a kovové předměty.
Hlavním cílem hledající posádky bude najít hlas z kokpitu a černou skříňku, která by pomohla zjistit, co způsobilo pád.
Vyhledávací tým z Brazílie nalezl kousek zadní části tryskáče Air France letu 447, který se zřítil do Atlantického oceánu 1. června.
Pátrací týmy z Brazílie se zaměřily hlavně na oblast severně od ostrova Fernando de Noronha, který leží přibližně 200 mil od brazilského pobřeží.
Francouzská armáda byla o několik set mil dále a prohledávala oceán v blízkosti Kapverdských ostrovů.
Také Pentagon povolal letectvo, aby pomohlo při pátrání po letadle.
Podle mluvčí společnosti Air France odletěl Airbus A330 z Ria de Janeira v neděli večer se dvanácti členy posádky a 216 pasažéry na palubě.

Nicméně dodal, že "je předčasné teď říkat, co se stalo."
Pátrací týmy z Francie a Brazílie potvrdily, že nenalezly žádné trosky z letu 447 společnosti Air France.
Brazilské letectvo původně 2. června informovalo, že nalezlo trosky letadla.
2. června byla vydána zpráva, kterou brazilské letectvo potvrdilo, že se ve vodě našlo jedno sedadlo a jedna záchranná vesta.
Mastná skvrna na hladině, o které se myslelo, že je z letadla, se nakonec ukázala pocházet z jiného zdroje, nejspíše z lodi.
Airbus A330 - 00, na jehož palubě bylo celkem 228 lidí - 216 pasažérů a 12 členů posádky, se za špatného počasí zřítil do Atlantického oceánu.

MEAD:

Plánovaná trať letu 447.
Letadlo je pohřešováno tak dlouho, že už by mu nezbývalo žádné palivo, kdyby bylo ve vzduchu.
Požádal také Ameriku, aby použila jejich sledovací satelity při hledání letadla.
+33 1 57 02 10 55, pro volání ze zemí mimo Francii nebo Brazílii.
Air France to neudělal v případě havarovaného letounu, i když celá flotila už upravený design má.
Tři další případy s podobnými okolnostmi byly zjištěny.
Úřady si zatím nejsou jisté, co způsobilo incident, ale jsou pesimistické, co se týče nalezení přeživších.
Když bylo naposledy v radarovém spojení, nebyly zřejmé žádné problémy.
GMT).
Příčina havárie není zatím známa.
Také informovaly, že autopilot byl v letadle vypnutý, ale že z chybových zpráv není možné zjistit, proč tomu tak bylo.
Havárie letadla se považuje za nejhorší letecké neštěstí od roku 2001.
Kontrolní systém hlavního počítačového systému a spoilery na křídlech se také rozbily.
Francie vyslala na místo výzkumnou loď "Pourquoi Pas?"
K havárii pravděpodobně přispěly Pitotovy trubice měřící rychlost letadla.
Ale vlastní příčina neštěstí není dosud zjištěna.
Avšak 4. května bylo hledání prodlouženo do 25. května.

NMF+K-means:

Let Air France 447 převážející 228 lidí z Ria de Janeira, Brazílie, do Paříže, letiště Charlese de Gaulla, je nezvěstný od brazilského pobřeží.
Airbus A330 - 00 byl naposledy slyšet přes rádio v 22: 0 místního času (01: 0 GMT).
Francouzský prezident Sarkozy řekl, že vyhlídky nalezení přeživších jsou "velmi malé" .
Francouzská vyšetřovací agentura BEA požádala o finanční pomoc pro vyhledávání, jak od Airbusu, tak i od Air France.
Vyšetřovatelé se již vzdali hledání hlasového záznamníku z pilotní kabiny a černé skřínky s využitím konvenční metody hledání 'audiobzučáků', jejichž baterie by se vybily po 30 - 0 dnech.
Komunikační a hlásící systém letadla (ACARS) byl schopen předat informace o problémech na palubě před havárii.
Brazilská vláda potvrdila, že trosky letadla nalezené v Atlantickém oceánu 650 kilometrů od pobřeží Brazílie jsou z letu Air France 447.
Nicméně brazilský ministr obrany řekl, že to může být těžké najít kvůli velké hloubce oceánu v oblasti, řekl, že "by to mohlo být v hloubce 2000 nebo 3000 m [od 6500 do 9800 stop] v této oblasti z oceánu."
Letadla a lodě francouzské, americké a brazilské armády v pondělí pátraly v Atlantiku poté, co nad oceánem zmizelo letadlo společnosti Air France s 228 lidmi na palubě.
Francouzský prezident Nicolas Sarkozy prohlásil, že je pesimistický ohledně nálezu přeživších osob.
Brazilské letectvo lokalizovalo těla pasažérů a vrak letadla z letu 447 společnosti Air France v Atlantickém oceánu.

Letadlo na letu 447 společnosti Air France letící z Ria de Janeira v Brazílii na letiště Charlese de Gaulla v Paříži zmizelo nad Atlantikem 1. června.
Pitvy mrtvých těl z letu Air France 447, jehož letadlo se začátkem měsíce zřítilo do vod Atlantického oceánu, odhalily zlomeniny nohou, což by mohlo znamenat, že se letadlo rozlomilo na kusy během letu.
Při vyšetřování se zatím našlo několik částí letadla a nějaká mrtvá těla cestujících, ale stále se hledá záznamník letových dat a záznamník komunikace v kabině pilotů, které by mohly obsahovat důležité informace o tom, jak se nehoda přesně udála.
Pátrací týmy z Francie a Brazílie potvrdily, že nenalezly žádné trosky z letu 447 společnosti Air France.
Airbus A330 s 228 lidmi na palubě zmizel z radarů 1. června nad Atlantickým oceánem.
2. června byla vydána zpráva, kterou brazilské letectvo potvrdilo, že se ve vodě našlo jedno sedadlo a jedna záchranná vesta.

Příloha C Výsledky automatických sumarizačních metod (pro sady článků se stejnými tématy)

Zde jsou uvedeny celkové výsledky získané ze softwarového balíčku ROUGE pro všech 15 sad článků (Tabulka 4 až Tabulka 15). Zobecněné (zprůměrované) výsledky viz kapitola 7.2.

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.1:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.11505	0.15244	0.0815	**0.186**
		Přesnos	0.13986	**0.14441**	0.12057	0.1246
		F-skóre	0.12625	0.14832	0.09726	**0.14923**
	ROUGE-SU	Úplnost	0.08962	0.18189	0.07828	**0.277**
		Přesnos	0.1322	0.16331	**0.17063**	0.12466
		F-skóre	0.10682	**0.1721**	0.10732	0.17194
	ROUGE-W	Úplnost	0.0642	0.07948	0.05475	**0.09253**
		Přesnos	0.19962	0.19273	**0.20702**	0.15878
		F-skóre	0.09715	0.11255	0.0866	**0.11692**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.16812	0.13913	0.16522	**0.18841**
		Přesnos	0.2028	0.13079	**0.24255**	0.12524
		F-skóre	0.18384	0.13483	**0.19655**	0.15046
	ROUGE-SU	Úplnost	0.09828	0.17281	0.10541	**0.27163**
		Přesnos	0.14276	0.15279	**0.22628**	0.12038
		F-skóre	0.11642	0.16218	0.14382	**0.16683**
	ROUGE-W	Úplnost	0.07689	0.07747	0.07565	**0.09037**
		Přesnos	0.23254	0.18273	**0.27826**	0.15086
		F-skóre	0.11557	0.10881	**0.11896**	0.11303
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.04598	**0.11782**	0.04885	**0.11782**
		Přesnos	0.05594	**0.11172**	0.07234	0.079
		F-skóre	0.05047	**0.11469**	0.05832	0.09458
	ROUGE-SU	Úplnost	0.07537	0.17019	0.06996	**0.24202**
		Přesnos	0.11138	**0.15309**	0.1528	0.10912
		F-skóre	0.0899	**0.16119**	0.09598	0.15042
	ROUGE-W	Úplnost	0.04154	0.06313	0.04026	**0.0674**
		Přesnos	0.14192	**0.16819**	0.16727	0.12707
		F-skóre	0.06427	**0.0918**	0.0649	0.08808
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.13143	0.2	0.03143	**0.25143**
		Přesnos	0.16084	**0.19074**	0.04681	0.16956
		F-skóre	0.14466	0.19526	0.03761	**0.20253**
	ROUGE-SU	Úplnost	0.0953	0.20228	0.06012	**0.31681**
		Přesnos	0.14245	**0.18405**	0.13281	0.14448
		F-skóre	0.1142	0.19273	0.08277	**0.19846**
	ROUGE-W	Úplnost	0.07696	0.10068	0.04902	**0.12373**
		Přesnos	0.22182	**0.22633**	0.17184	0.19685
		F-skóre	0.11427	0.13937	0.07628	**0.15195**

Tabulka 4: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.1

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.2:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.24234	**0.25905**	0.07149	0.21634
		Přesnos	0.18125	**0.18675**	0.0828	0.09757
		F-skóre	0.20739	**0.21704**	0.07673	0.13449
	ROUGE-SU	Úplnost	0.24372	0.31001	0.12197	**0.39192**
		Přesnos	0.13687	0.16178	**0.16367**	0.08023
		F-skóre	0.1753	**0.21261**	0.13978	0.13319
	ROUGE-W	Úplnost	0.10802	0.10804	0.06327	**0.11036**
		Přesnos	**0.20761**	0.20015	0.18806	0.12801
		F-skóre	**0.1421**	0.14033	0.09468	0.11853
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.28382	0.27851	0.06101	**0.33952**
		Přesnos	**0.22292**	0.21084	0.07419	0.1608
		F-skóre	**0.24971**	0.24	0.06696	0.21824
	ROUGE-SU	Úplnost	0.23182	0.30571	0.11721	**0.4367**
		Přesnos	0.14325	**0.17554**	0.17306	0.09837
		F-skóre	0.17708	**0.22302**	0.13976	0.16057
	ROUGE-W	Úplnost	0.1256	0.11779	0.06529	**0.1401**
		Přesnos	**0.23976**	0.21673	0.19274	0.16139
		F-skóre	**0.16484**	0.15263	0.09754	0.14999
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	**0.20649**	0.15929	0.07375	0.17109
		Přesnos	**0.14583**	0.10843	0.08065	0.07286
		F-skóre	**0.17094**	0.12903	0.07705	0.1022
	ROUGE-SU	Úplnost	0.25962	0.27066	0.13419	**0.43411**
		Přesnos	**0.12983**	0.12577	0.16034	0.07913
		F-skóre	**0.1731**	0.17174	0.1461	0.13386
	ROUGE-W	Úplnost	0.08757	0.08041	0.05969	**0.09646**
		Přesnos	0.17294	0.15306	**0.18232**	0.11496
		F-skóre	**0.11627**	0.10543	0.08994	0.1049
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.23269	**0.33241**	0.08033	0.13019
		Přesnos	0.175	**0.24096**	0.09355	0.05905
		F-skóre	0.19976	**0.27939**	0.08644	0.08125
	ROUGE-SU	Úplnost	0.24267	**0.34941**	0.11637	0.30588
		Přesnos	0.13754	**0.18403**	0.1576	0.0632
		F-skóre	0.17557	**0.24108**	0.13388	0.10476
	ROUGE-W	Úplnost	0.11109	**0.1262**	0.06496	0.09363
		Přesnos	0.20906	**0.22893**	0.18908	0.10634
		F-skóre	0.14508	**0.16271**	0.0967	0.09958

Tabulka 5: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.2

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.3:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.17343	0.23096	0.0753	**0.38917**
		Přesnos	0.11884	0.12281	0.08404	**0.13851**
		F-skóre	0.14104	0.16035	0.07943	**0.20431**
	ROUGE-SU	Úplnost	0.26968	0.36264	0.13312	**0.52336**
		Přesnos	0.12695	0.10291	**0.16572**	0.06664
		F-skóre	**0.17263**	0.16032	0.14764	0.11823
	ROUGE-W	Úplnost	0.07593	0.08826	0.05117	**0.1073**
		Přesnos	**0.15384**	0.13881	0.1687	0.11301
		F-skóre	0.10168	0.10791	0.07852	**0.11008**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.15897	0.24872	0.06667	**0.4359**
		Přesnos	0.10783	0.1309	0.07365	**0.15357**
		F-skóre	0.1285	0.17153	0.06999	**0.22712**
	ROUGE-SU	Úplnost	0.23146	0.34489	0.1234	**0.53087**
		Přesnos	0.10674	0.09589	**0.15051**	0.06622
		F-skóre	0.1461	**0.15006**	0.13561	0.11775
	ROUGE-W	Úplnost	0.07799	0.09562	0.04984	**0.12174**
		Přesnos	0.15236	0.14502	**0.15843**	0.12363
		F-skóre	0.10317	0.11525	0.07583	**0.12268**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.21144	0.28358	0.07463	**0.33831**
		Přesnos	0.14783	**0.15385**	0.08499	0.12285
		F-skóre	0.174	**0.19948**	0.07947	0.18025
	ROUGE-SU	Úplnost	0.30529	0.42286	0.14504	**0.49079**
		Přesnos	0.14955	0.12488	**0.18791**	0.06503
		F-skóre	**0.20076**	0.19282	0.16372	0.11484
	ROUGE-W	Úplnost	0.09808	0.11613	0.06247	**0.11773**
		Přesnos	0.18455	0.16962	**0.19125**	0.11516
		F-skóre	0.12809	**0.13787**	0.09418	0.11643
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.14872	0.15897	0.08462	**0.39487**
		Přesnos	0.10087	0.08367	0.09348	**0.13911**
		F-skóre	0.12021	0.10964	0.08883	**0.20574**
	ROUGE-SU	Úplnost	0.27006	0.31642	0.13016	**0.55045**
		Přesnos	0.12454	0.08797	**0.15875**	0.06866
		F-skóre	**0.17047**	0.13767	0.14304	0.12209
	ROUGE-W	Úplnost	0.0551	0.05749	0.0428	**0.08588**
		Přesnos	0.1234	0.09995	**0.15598**	0.09999
		F-skóre	0.07618	0.07299	0.06717	**0.0924**

Tabulka 6: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.3

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.4:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.14663	0.16325	0.0479	**0.21408**
		Přesnos	**0.14925**	0.1431	0.05875	0.13748
		F-skóre	0.14793	0.15251	0.05277	**0.16743**
	ROUGE-SU	Úplnost	0.17314	0.19451	0.08926	**0.28092**
		Přesnos	**0.17946**	0.14971	0.13411	0.11628
		F-skóre	**0.17624**	0.16919	0.10718	0.16448
	ROUGE-W	Úplnost	0.07822	0.09544	0.05719	**0.11251**
		Přesnos	0.19132	**0.20112**	0.16846	0.1738
		F-skóre	0.11104	0.12945	0.08539	**0.13659**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.21203	0.10315	0.0659	**0.24355**
		Přesnos	**0.2209**	0.09254	0.08273	0.16008
		F-skóre	**0.21637**	0.09756	0.07336	0.19318
	ROUGE-SU	Úplnost	0.2177	0.16756	0.09526	**0.31729**
		Přesnos	**0.23619**	0.13499	0.1498	0.13746
		F-skóre	**0.22657**	0.14952	0.11646	0.19182
	ROUGE-W	Úplnost	0.10295	0.07843	0.0606	**0.12114**
		Přesnos	**0.25863**	0.16975	0.18333	0.1922
		F-skóre	0.14728	0.10729	0.09109	**0.14861**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.14244	**0.28488**	0.0407	0.22674
		Přesnos	0.14627	**0.25193**	0.05036	0.14689
		F-skóre	0.14433	**0.26739**	0.04502	0.17828
	ROUGE-SU	Úplnost	0.15842	0.25317	0.08436	**0.25638**
		Přesnos	0.16701	**0.19818**	0.12891	0.10793
		F-skóre	0.1626	**0.22233**	0.10198	0.15191
	ROUGE-W	Úplnost	0.05378	**0.11477**	0.04852	0.10207
		Přesnos	0.14347	**0.26379**	0.15589	0.17198
		F-skóre	0.07823	**0.15995**	0.07401	0.12811
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.08182	0.1	0.03636	**0.1697**
		Přesnos	0.0806	0.08483	0.04317	**0.10546**
		F-skóre	0.08121	0.09179	0.03947	**0.13008**
	ROUGE-SU	Úplnost	0.1393	0.16092	0.08789	**0.2669**
		Přesnos	**0.13519**	0.11597	0.12363	0.10344
		F-skóre	0.13721	0.1348	0.10274	**0.1491**
	ROUGE-W	Úplnost	0.07811	0.09001	0.06403	**0.11541**
		Přesnos	**0.16809**	0.16688	0.16594	0.15685
		F-skóre	0.10666	0.11694	0.0924	**0.13298**

Tabulka 7: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.4

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.5:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.17778	0.1942	0.16232	**0.3314**
		Přesnos	0.10855	0.08375	**0.14249**	0.10962
		F-skóre	0.1348	0.11703	0.15176	**0.16475**
	ROUGE-SU	Úplnost	0.27054	0.35391	0.21335	**0.51337**
		Přesnos	0.10123	0.06615	**0.16461**	0.0565
		F-skóre	0.14733	0.11147	**0.18584**	0.1018
	ROUGE-W	Úplnost	0.10205	0.11158	0.08827	**0.14237**
		Přesnos	0.15877	0.12267	**0.19729**	0.12009
		F-skóre	0.12424	0.11686	0.12197	**0.13028**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.23209	0.2808	0.14327	**0.40688**
		Přesnos	**0.14336**	0.1225	0.12723	0.13615
		F-skóre	0.17724	0.17058	0.13477	**0.20403**
	ROUGE-SU	Úplnost	0.32675	0.45857	0.22589	**0.58633**
		Přesnos	0.12508	0.08769	**0.17831**	0.06602
		F-skóre	0.18091	0.14723	**0.1993**	0.11868
	ROUGE-W	Úplnost	0.11217	0.12689	0.08452	**0.15107**
		Přesnos	0.17973	0.14366	**0.19454**	0.13123
		F-skóre	0.13813	0.13476	0.11784	**0.14045**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.17699	0.18289	0.21239	**0.41888**
		Přesnos	0.10619	0.0775	**0.18321**	0.13615
		F-skóre	0.13274	0.10887	0.19672	**0.2055**
	ROUGE-SU	Úplnost	0.26285	0.33759	0.21532	**0.53641**
		Přesnos	0.09496	0.06093	**0.16041**	0.057
		F-skóre	0.13952	0.10323	**0.18385**	0.10305
	ROUGE-W	Úplnost	0.10096	0.10449	0.09519	**0.15402**
		Přesnos	0.1565	0.11446	**0.21199**	0.12944
		F-skóre	0.12274	0.10925	0.13138	**0.14066**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.12392	0.11816	0.13256	**0.17003**
		Přesnos	0.07611	0.05125	**0.11705**	0.05657
		F-skóre	0.0943	0.07149	**0.12432**	0.08489
	ROUGE-SU	Úplnost	0.22103	0.26363	0.19878	**0.41759**
		Přesnos	0.08365	0.04984	**0.15513**	0.04649
		F-skóre	0.12137	0.08383	**0.17426**	0.08367
	ROUGE-W	Úplnost	0.09214	0.10216	0.0851	**0.12059**
		Přesnos	0.13956	0.10934	**0.18516**	0.09902
		F-skóre	0.111	0.10563	**0.11661**	0.10875

Tabulka 8: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.5

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.6:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.07811	0.13351	0.08174	**0.24251**
		Přesnos	0.04125	0.07357	0.08646	**0.13363**
		F-skóre	0.05399	0.09487	0.08403	**0.17231**
	ROUGE-SU	Úplnost	0.20385	0.27897	0.10342	**0.36726**
		Přesnos	0.05714	0.08513	**0.11578**	0.11208
		F-skóre	0.08926	0.13045	0.10925	**0.17175**
	ROUGE-W	Úplnost	0.06574	0.08392	0.0488	**0.09757**
		Přesnos	0.09527	0.12691	0.14144	**0.14754**
		F-skóre	0.0778	0.10103	0.07256	**0.11746**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.08023	0.11748	0.09456	**0.30086**
		Přesnos	0.04029	0.06156	0.0951	**0.15766**
		F-skóre	0.05364	0.08079	0.09483	**0.2069**
	ROUGE-SU	Úplnost	0.20603	0.26229	0.1151	**0.39372**
		Přesnos	0.05217	0.07232	**0.11643**	0.10856
		F-skóre	0.08326	0.11338	0.11576	**0.17019**
	ROUGE-W	Úplnost	0.09111	0.10757	0.06684	**0.13663**
		Přesnos	0.10338	0.12736	0.15168	**0.16177**
		F-skóre	0.09686	0.11663	0.09279	**0.14814**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.09164	0.09973	0.06739	**0.18868**
		Přesnos	0.04892	0.05556	0.07205	**0.10511**
		F-skóre	0.06379	0.07136	0.06964	**0.13501**
	ROUGE-SU	Úplnost	0.22451	0.2892	0.06739	**0.37182**
		Přesnos	0.06422	0.09006	0.07205	**0.11579**
		F-skóre	0.09987	0.13735	0.06964	**0.17659**
	ROUGE-W	Úplnost	0.05008	0.05818	0.03402	**0.06592**
		Přesnos	0.08744	0.106	0.11881	**0.12009**
		F-skóre	0.06369	0.07513	0.05289	**0.08512**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.06299	0.1811	0.08399	**0.24147**
		Přesnos	0.03453	0.1036	0.09222	**0.13814**
		F-skóre	0.04461	0.1318	0.08791	**0.17574**
	ROUGE-SU	Úplnost	0.18243	0.28326	0.10251	**0.34072**
		Přesnos	0.05502	0.09301	**0.12349**	0.11188
		F-skóre	0.08454	0.14004	0.11203	**0.16845**
	ROUGE-W	Úplnost	0.06525	0.0967	0.05271	**0.10546**
		Přesnos	0.09487	0.14671	0.15327	**0.16**
		F-skóre	0.07732	0.11657	0.07844	**0.12713**

Tabulka 9: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.6

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.7:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.10644	0.14846	0.05415	**0.21755**
		Přesnos	**0.06872**	0.05662	0.04053	0.06248
		F-skóre	0.08352	0.08198	0.04636	**0.09708**
	ROUGE-SU	Úplnost	0.21682	0.33049	0.17322	**0.44538**
		Přesnos	0.09065	0.04834	**0.09725**	0.03697
		F-skóre	**0.12785**	0.08434	0.12457	0.06827
	ROUGE-W	Úplnost	0.08386	0.09724	0.06952	**0.11654**
		Přesnos	**0.14012**	0.09607	0.13464	0.08672
		F-skóre	**0.10492**	0.09665	0.09169	0.09944
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.07932	0.14731	0.07082	**0.20963**
		Přesnos	0.05063	0.05556	0.05241	**0.05953**
		F-skóre	0.06181	0.08069	0.06024	**0.09273**
	ROUGE-SU	Úplnost	0.22158	0.34995	0.18936	**0.45068**
		Přesnos	0.09057	0.05004	**0.10393**	0.03657
		F-skóre	0.12858	0.08756	**0.1342**	0.06765
	ROUGE-W	Úplnost	0.09966	0.12326	0.08941	**0.14031**
		Přesnos	0.14242	0.10415	**0.14809**	0.08929
		F-skóre	**0.11726**	0.1129	0.1115	0.10913
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.06215	0.11582	0.05085	**0.14407**
		Přesnos	0.03978	**0.0438**	0.03774	0.04103
		F-skóre	0.04851	0.06356	0.04332	**0.06387**
	ROUGE-SU	Úplnost	0.18304	0.31719	0.16845	**0.39847**
		Přesnos	0.07523	0.04561	**0.09298**	0.03251
		F-skóre	0.10663	0.07975	**0.11982**	0.06012
	ROUGE-W	Úplnost	0.06148	0.08149	0.06044	**0.09241**
		Přesnos	0.11115	0.08711	**0.12665**	0.0744
		F-skóre	0.07917	**0.08421**	0.08183	0.08243
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.17582	0.18132	0.04121	**0.2967**
		Přesnos	**0.11573**	0.07051	0.03145	0.08689
		F-skóre	**0.13958**	0.10154	0.03567	0.13442
	ROUGE-SU	Úplnost	0.2443	0.32476	0.16254	**0.48476**
		Přesnos	**0.10614**	0.04936	0.09484	0.04181
		F-skóre	**0.14799**	0.0857	0.11979	0.07698
	ROUGE-W	Úplnost	0.09359	0.09245	0.06285	**0.12197**
		Přesnos	**0.16571**	0.09679	0.12897	0.09618
		F-skóre	**0.11962**	0.09457	0.08451	0.10755

Tabulka 10: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.7

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.8:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.10636	0.18182	0.08091	**0.39091**
		Přesnos	0.08904	0.1462	0.12787	**0.19852**
		F-skóre	0.09693	0.16208	0.09911	**0.26332**
	ROUGE-SU	Úplnost	0.18553	0.2609	0.08188	**0.48422**
		Přesnos	0.13024	0.16902	**0.20364**	0.12543
		F-skóre	0.15304	**0.20514**	0.1168	0.19925
	ROUGE-W	Úplnost	0.06152	0.07339	0.04349	**0.11504**
		Přesnos	0.14646	0.16782	**0.19508**	0.16629
		F-skóre	0.08665	0.10212	0.07112	**0.136**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.12603	0.1726	0.09041	**0.37808**
		Přesnos	0.10502	0.13816	0.14224	**0.19114**
		F-skóre	0.11457	0.15347	0.11055	**0.25391**
	ROUGE-SU	Úplnost	0.18912	0.2573	0.08736	**0.46276**
		Přesnos	0.13151	0.16512	**0.21522**	0.11875
		F-skóre	0.15514	**0.20115**	0.12428	0.189
	ROUGE-W	Úplnost	0.08095	0.09064	0.05304	**0.13482**
		Přesnos	0.16996	0.1828	**0.20982**	0.17187
		F-skóre	0.10967	0.12119	0.08468	**0.15111**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.11142	0.21727	0.07242	**0.47632**
		Přesnos	0.09132	0.17105	0.11207	**0.23684**
		F-skóre	0.10037	0.19141	0.08798	**0.31637**
	ROUGE-SU	Úplnost	0.19022	0.28931	0.08387	**0.53**
		Přesnos	0.12798	0.17963	**0.19993**	0.13158
		F-skóre	0.15301	**0.22164**	0.11817	0.21082
	ROUGE-W	Úplnost	0.06747	0.09008	0.04719	**0.14507**
		Přesnos	0.14958	0.19186	**0.19715**	0.1953
		F-skóre	0.09299	0.1226	0.07615	**0.16648**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.08245	0.15691	0.07979	**0.32181**
		Přesnos	0.07078	0.12939	0.12931	**0.16759**
		F-skóre	0.07617	0.14183	0.09869	**0.2204**
	ROUGE-SU	Úplnost	0.17788	0.23838	0.0749	**0.46271**
		Přesnos	0.13123	0.1623	**0.19578**	0.12597
		F-skóre	0.15103	0.19312	0.10835	**0.19803**
	ROUGE-W	Úplnost	0.04228	0.04711	0.03359	**0.07633**
		Přesnos	0.11891	0.12729	**0.178**	0.13036
		F-skóre	0.06238	0.06877	0.05652	**0.09628**

Tabulka 11: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.8

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.9:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.04667	0.17378	0.05462	**0.22542**
		Přesnos	0.05365	0.14262	0.07671	**0.15255**
		F-skóre	0.04992	0.15667	0.06381	**0.18196**
	ROUGE-SU	Úplnost	0.09218	0.22556	0.07773	**0.32868**
		Přesnos	0.12175	0.15229	**0.1529**	0.15109
		F-skóre	0.10492	0.18182	0.10306	**0.20702**
	ROUGE-W	Úplnost	0.04595	0.07869	0.04741	**0.09182**
		Přesnos	0.14845	0.18168	**0.187**	0.17489
		F-skóre	0.07018	0.10982	0.07564	**0.12042**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.04893	0.25382	0.08869	**0.29358**
		Přesnos	0.05479	**0.20293**	0.12134	0.19355
		F-skóre	0.05169	0.22554	0.10248	**0.23329**
	ROUGE-SU	Úplnost	0.0935	0.25468	0.08952	**0.35822**
		Přesnos	0.11713	0.16309	**0.16702**	0.15618
		F-skóre	0.10399	0.19885	0.11656	**0.21752**
	ROUGE-W	Úplnost	0.05164	0.10878	0.06272	**0.11964**
		Přesnos	0.15009	**0.22592**	0.22255	0.20499
		F-skóre	0.07684	0.14685	0.09786	**0.1511**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.04011	0.12034	0.05158	**0.16619**
		Přesnos	0.04795	0.10269	0.07531	**0.11694**
		F-skóre	0.04368	0.11082	0.06123	**0.13728**
	ROUGE-SU	Úplnost	0.09363	0.20811	0.08033	**0.31144**
		Přesnos	0.13353	0.15172	**0.17061**	0.15458
		F-skóre	0.11008	0.1755	0.10923	**0.20661**
	ROUGE-W	Úplnost	0.03436	0.05097	0.03689	**0.06223**
		Přesnos	0.13244	0.1404	**0.17362**	0.14142
		F-skóre	0.05456	0.07479	0.06085	**0.08643**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.05136	0.15106	0.02417	**0.22054**
		Přesnos	0.05822	0.12225	0.03347	**0.14718**
		F-skóre	0.05458	0.13514	0.02807	**0.17654**
	ROUGE-SU	Úplnost	0.08928	0.21653	0.06334	**0.31901**
		Přesnos	0.11458	0.14206	0.12106	**0.14249**
		F-skóre	0.10036	0.17156	0.08317	**0.19699**
	ROUGE-W	Úplnost	0.05578	0.08486	0.04605	**0.1031**
		Přesnos	0.16251	0.17668	0.1638	**0.17709**
		F-skóre	0.08305	0.11465	0.07189	**0.13033**

Tabulka 12: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.9

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.10:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.09212	0.15195	0.07882	**0.26971**
		Přesnos	0.08291	0.1084	0.08925	**0.14343**
		F-skóre	0.08727	0.12653	0.08371	**0.18727**
	ROUGE-SU	Úplnost	0.17677	0.26299	0.12505	**0.39772**
		Přesnos	0.14332	0.1342	**0.16016**	0.11295
		F-skóre	0.1583	**0.17771**	0.14044	0.17594
	ROUGE-W	Úplnost	0.06727	0.08401	0.05688	**0.11275**
		Přesnos	0.16418	0.16262	**0.17454**	0.16278
		F-skóre	0.09544	0.11079	0.0858	**0.13322**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.08406	0.15652	0.09275	**0.29275**
		Přesnos	0.07436	0.10976	0.10323	**0.15303**
		F-skóre	0.07891	0.12903	0.09771	**0.20099**
	ROUGE-SU	Úplnost	0.17138	0.25419	0.13228	**0.36179**
		Přesnos	0.13425	0.12531	**0.16368**	0.09926
		F-skóre	0.15056	**0.16787**	0.14631	0.15578
	ROUGE-W	Úplnost	0.08418	0.11453	0.0773	**0.14525**
		Přesnos	0.16864	0.18196	**0.19469**	0.17212
		F-skóre	0.1123	0.14058	0.11066	**0.15755**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.03955	**0.19209**	0.06497	0.12712
		Přesnos	0.0359	**0.13821**	0.07419	0.06818
		F-skóre	0.03764	**0.16076**	0.06927	0.08876
	ROUGE-SU	Úplnost	0.15147	0.27593	0.12349	**0.31295**
		Přesnos	0.12489	0.14319	**0.16084**	0.09038
		F-skóre	0.1369	**0.18854**	0.13971	0.14025
	ROUGE-W	Úplnost	0.04077	**0.06432**	0.0388	0.06276
		Přesnos	0.11978	**0.14989**	0.14331	0.10908
		F-skóre	0.06083	**0.09001**	0.06107	0.07968
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.15254	0.10734	0.0791	**0.38983**
		Přesnos	0.13846	0.07724	0.09032	**0.20909**
		F-skóre	0.14516	0.08984	0.08434	**0.27219**
	ROUGE-SU	Úplnost	0.20719	0.25842	0.11974	**0.51662**
		Přesnos	**0.17084**	0.1341	0.15595	0.14921
		F-skóre	0.18727	0.17657	0.13547	**0.23155**
	ROUGE-W	Úplnost	0.08551	0.08309	0.06221	**0.14609**
		Přesnos	0.20199	0.15566	0.18474	**0.20413**
		F-skóre	0.12015	0.10835	0.09308	**0.1703**

Tabulka 13: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.10

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.11:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.06573	0.27887	0.08826	**0.34085**
		Přesnos	0.05877	**0.15968**	0.10731	0.14286
		F-skóre	0.06206	**0.20308**	0.09686	0.20133
	ROUGE-SU	Úplnost	0.13304	0.3904	0.10706	**0.45282**
		Přesnos	0.10654	0.12854	**0.15806**	0.07999
		F-skóre	0.11832	**0.1934**	0.12765	0.13596
	ROUGE-W	Úplnost	0.07317	0.13695	0.07282	**0.15531**
		Přesnos	0.15401	0.18473	**0.20818**	0.15342
		F-skóre	0.09921	**0.15729**	0.1079	0.15436
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.03488	0.2936	0.06105	**0.34012**
		Přesnos	0.03023	**0.1629**	0.07192	0.13813
		F-skóre	0.03239	**0.20954**	0.06604	0.19647
	ROUGE-SU	Úplnost	0.12447	0.42779	0.11241	**0.45857**
		Přesnos	0.09356	0.1322	**0.15577**	0.07603
		F-skóre	0.10682	**0.20198**	0.13058	0.13043
	ROUGE-W	Úplnost	0.06694	0.14295	0.07047	**0.15901**
		Přesnos	0.1366	0.18696	**0.19534**	0.1523
		F-skóre	0.08985	**0.16202**	0.10357	0.15558
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.05932	0.27684	0.07345	**0.37006**
		Přesnos	0.0529	**0.15806**	0.08904	0.15466
		F-skóre	0.05593	0.20123	0.0805	**0.21815**
	ROUGE-SU	Úplnost	0.12239	0.38127	0.09745	**0.47883**
		Přesnos	0.09741	0.12474	**0.14298**	0.08405
		F-skóre	0.10848	**0.18798**	0.1159	0.143
	ROUGE-W	Úplnost	0.07528	0.13903	0.07019	**0.16366**
		Přesnos	0.1544	0.18276	**0.19556**	0.15754
		F-skóre	0.10121	0.15792	0.1033	**0.16054**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.10082	0.26703	0.12807	**0.31335**
		Přesnos	0.0932	0.15806	**0.16096**	0.13577
		F-skóre	0.09686	**0.19858**	0.14264	0.18945
	ROUGE-SU	Úplnost	0.15047	0.36602	0.11129	**0.42357**
		Přesnos	0.12866	0.12867	**0.17544**	0.07989
		F-skóre	0.13871	**0.19041**	0.13619	0.13443
	ROUGE-W	Úplnost	0.07682	0.12958	0.07729	**0.14427**
		Přesnos	0.17065	0.18448	**0.23321**	0.15041
		F-skóre	0.10595	**0.15223**	0.1161	0.14728

Tabulka 14: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.11

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.12:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.12837	0.19721	0.1507	**0.37488**
		Přesnos	0.11735	0.14846	**0.19926**	0.17911
		F-skóre	0.12261	0.1694	0.17161	**0.2424**
	ROUGE-SU	Úplnost	0.17075	0.29592	0.12356	**0.48129**
		Přesnos	0.14279	0.16805	**0.21546**	0.11035
		F-skóre	0.15552	**0.21436**	0.15705	0.17954
	ROUGE-W	Úplnost	0.08628	0.10955	0.07061	**0.15758**
		Přesnos	0.18922	0.19796	**0.22373**	0.18085
		F-skóre	0.11852	0.14105	0.10734	**0.16841**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.10526	0.20222	0.17729	**0.27147**
		Přesnos	0.09694	0.15336	**0.23616**	0.13067
		F-skóre	0.10093	0.17443	**0.20253**	0.17642
	ROUGE-SU	Úplnost	0.16293	0.29562	0.12602	**0.42436**
		Přesnos	0.13827	0.17037	**0.22302**	0.09874
		F-skóre	0.14959	**0.21616**	0.16104	0.1602
	ROUGE-W	Úplnost	0.08308	0.11105	0.07521	**0.13014**
		Přesnos	0.18004	0.19826	**0.23547**	0.14757
		F-skóre	0.1137	**0.14236**	0.11401	0.13831
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.1573	0.1882	0.10112	**0.46348**
		Přesnos	0.14286	0.14076	0.13284	**0.22**
		F-skóre	0.14973	0.16106	0.11483	**0.29837**
	ROUGE-SU	Úplnost	0.18483	0.29373	0.10048	**0.53435**
		Přesnos	0.15256	0.16465	**0.17295**	0.12092
		F-skóre	0.16715	**0.21102**	0.12711	0.19721
	ROUGE-W	Úplnost	0.08893	0.10381	0.06475	**0.17599**
		Přesnos	0.20124	0.19354	0.2117	**0.20839**
		F-skóre	0.12335	0.13514	0.09917	**0.19082**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.12291	0.20112	0.17318	**0.39106**
		Přesnos	0.11224	0.15126	**0.22878**	0.18667
		F-skóre	0.11733	0.17266	0.19713	**0.25271**
	ROUGE-SU	Úplnost	0.16478	0.29838	0.14387	**0.48671**
		Přesnos	0.13754	0.16913	**0.25041**	0.11138
		F-skóre	0.14993	**0.21589**	0.18275	0.18128
	ROUGE-W	Úplnost	0.08667	0.1141	0.0721	**0.16495**
		Přesnos	0.18628	0.20205	**0.22391**	0.18553
		F-skóre	0.1183	0.14584	0.10908	**0.17464**

Tabulka 15: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.12

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.13:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.12311	0.25025	**0.03229**	0.27346
		Přesnos	0.09306	0.11156	0.03039	**0.13265**
		F-skóre	0.106	0.15432	0.03131	**0.17864**
	ROUGE-SU	Úplnost	0.19311	0.36804	0.11255	**0.3685**
		Přesnos	**0.11061**	0.07352	0.09976	0.08713
		F-skóre	0.14066	0.12256	0.10577	**0.14094**
	ROUGE-W	Úplnost	0.09996	0.13955	0.07125	**0.14096**
		Přesnos	**0.16983**	0.13996	0.15063	0.15381
		F-skóre	0.12585	0.13975	0.09674	**0.1471**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.10272	0.19637	**0.02719**	0.25378
		Přesnos	0.0778	0.08772	**0.02564**	0.12335
		F-skóre	0.08854	0.12127	**0.02639**	0.16601
	ROUGE-SU	Úplnost	0.15639	0.29996	0.09608	**0.34613**
		Přesnos	**0.08992**	0.06015	0.08549	0.08215
		F-skóre	0.11419	0.10021	0.09048	**0.13278**
	ROUGE-W	Úplnost	0.08922	**0.11648**	0.06638	0.11462
		Přesnos	**0.15234**	0.1174	0.14104	0.12568
		F-skóre	0.11253	0.11694	0.09027	**0.1199**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.14583	0.28571	0.02083	**0.35714**
		Přesnos	0.11213	0.12955	0.01994	**0.17621**
		F-skóre	0.12678	0.17827	0.02038	**0.23599**
	ROUGE-SU	Úplnost	0.19957	0.37718	0.09959	**0.4167**
		Přesnos	**0.11822**	0.07793	0.0913	0.1019
		F-skóre	0.14848	0.12917	0.09526	**0.16376**
	ROUGE-W	Úplnost	0.10277	0.13348	0.06795	**0.16756**
		Přesnos	0.18065	0.13851	0.14862	**0.18916**
		F-skóre	0.13101	0.13595	0.09326	**0.17771**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.12037	**0.26852**	0.04938	0.20679
		Přesnos	0.08924	**0.11741**	0.04558	0.09838
		F-skóre	0.10249	**0.16338**	0.0474	0.13333
	ROUGE-SU	Úplnost	0.22449	**0.42925**	0.14367	0.34003
		Přesnos	**0.12369**	0.08249	0.12251	0.07734
		F-skóre	0.1595	**0.13839**	0.13225	0.12602
	ROUGE-W	Úplnost	0.10803	**0.1695**	0.07984	0.13862
		Přesnos	0.17623	**0.16323**	0.16208	0.14524
		F-skóre	0.13395	**0.16631**	0.10698	0.14185

Tabulka 16: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.13

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.14:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.09249	0.19872	0.16209	**0.34615**
		Přesnos	0.09457	0.09443	**0.16809**	0.13057
		F-skóre	0.09352	0.12802	0.16504	**0.18962**
	ROUGE-SU	Úplnost	0.1454	0.22456	0.14492	**0.43721**
		Přesnos	0.15203	0.05094	**0.15586**	0.06255
		F-skóre	0.14864	0.08304	**0.15019**	0.10944
	ROUGE-W	Úplnost	0.07444	0.08802	0.08929	**0.14873**
		Přesnos	0.18032	0.09924	**0.21935**	0.13314
		F-skóre	0.10538	0.09329	0.12692	**0.1405**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.07775	0.20375	0.23592	**0.24933**
		Přesnos	0.08146	0.09922	**0.25071**	0.09637
		F-skóre	0.07956	0.13345	**0.24309**	0.13901
	ROUGE-SU	Úplnost	0.15042	0.22513	0.17646	**0.41692**
		Přesnos	0.16507	0.0536	**0.19917**	0.0626
		F-skóre	0.1574	0.08659	**0.18713**	0.10886
	ROUGE-W	Úplnost	0.07456	0.07809	0.10184	**0.11591**
		Přesnos	0.19335	0.09425	**0.26785**	0.11108
		F-skóre	0.10762	0.08541	**0.14757**	0.11344
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.09804	0.15406	0.09244	**0.43697**
		Přesnos	0.09831	0.0718	0.09402	**0.16166**
		F-skóre	0.09817	0.09795	0.09322	**0.23601**
	ROUGE-SU	Úplnost	0.14544	0.21936	0.1172	**0.48302**
		Přesnos	**0.14626**	0.04786	0.12122	0.06646
		F-skóre	**0.14585**	0.07858	0.11918	0.11684
	ROUGE-W	Úplnost	0.07232	0.08159	0.07368	**0.1804**
		Přesnos	0.1681	0.08828	**0.1737**	0.15497
		F-skóre	0.10113	0.0848	0.10347	**0.16672**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.10221	0.23757	0.1547	**0.35635**
		Přesnos	0.10393	0.11227	**0.15954**	0.13368
		F-skóre	0.10306	0.15248	0.15708	**0.19442**
	ROUGE-SU	Úplnost	0.14003	0.22902	0.13841	**0.4142**
		Přesnos	0.14477	0.05137	**0.14718**	0.05859
		F-skóre	0.14236	0.08392	**0.14266**	0.10266
	ROUGE-W	Úplnost	0.0764	0.10507	0.09009	**0.15292**
		Přesnos	0.17934	0.11479	**0.21447**	0.13266
		F-skóre	0.10715	0.10972	0.12688	**0.14207**

Tabulka 17: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.14

Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.15:

			Metoda			
			LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
Porovnání se všemi třemi ideálními souhrny	ROUGE-2	Úplnost	0.16066	0.31451	0.04576	**0.42454**
		Přesnos	0.13614	0.18468	0.05054	**0.19173**
		F-skóre	0.14739	0.23271	0.04803	**0.26416**
	ROUGE-SU	Úplnost	0.20165	0.40249	0.10535	**0.47908**
		Přesnos	**0.14504**	0.13933	0.12841	0.09822
		F-skóre	**0.16872**	0.207	0.11574	0.16302
	ROUGE-W	Úplnost	0.10243	0.14109	0.06426	**0.17947**
		Přesnos	**0.20375**	0.19463	0.16647	0.19049
		F-skóre	0.13633	0.16359	0.09273	**0.18482**
Porovnání s 1. ideálním souhrnem	ROUGE-2	Úplnost	0.13814	0.33634	0.03604	**0.44144**
		Přesnos	0.11386	0.19211	0.03871	**0.19393**
		F-skóre	0.12483	0.24454	0.03733	**0.26948**
	ROUGE-SU	Úplnost	0.18359	0.41609	0.11628	**0.4847**
		Přesnos	0.12493	**0.13627**	0.13408	0.09402
		F-skóre	0.14868	**0.2053**	0.12455	0.15749
	ROUGE-W	Úplnost	0.09448	0.13479	0.06496	**0.17792**
		Přesnos	0.18419	0.18224	0.16493	**0.18509**
		F-skóre	0.1249	0.15496	0.09321	**0.18143**
Porovnání s 2. ideálním souhrnem	ROUGE-2	Úplnost	0.18625	0.21777	0.0745	**0.40688**
		Přesnos	0.16089	0.13036	0.08387	**0.18734**
		F-skóre	0.17264	0.16309	0.07891	**0.25655**
	ROUGE-SU	Úplnost	0.22643	0.32294	0.11601	**0.45551**
		Přesnos	**0.16917**	0.11612	0.14688	0.09701
		F-skóre	**0.19366**	0.17082	0.12963	0.15995
	ROUGE-W	Úplnost	0.11319	0.12114	0.06701	**0.17798**
		Přesnos	**0.23161**	0.17191	0.17856	0.19434
		F-skóre	0.15206	0.14213	0.09745	**0.1858**
Porovnání s 3. ideálním souhrnem	ROUGE-2	Úplnost	0.15652	0.3913	0.02609	**0.42609**
		Přesnos	0.13366	**0.23156**	0.02903	0.19393
		F-skóre	0.14419	**0.29095**	0.02748	0.26655
	ROUGE-SU	Úplnost	0.19314	0.47121	0.08426	**0.49797**
		Přesnos	0.14102	**0.1656**	0.10426	0.10364
		F-skóre	0.16302	**0.24507**	0.0932	0.17157
	ROUGE-W	Úplnost	0.09884	0.16738	0.06067	**0.18256**
		Přesnos	0.19484	**0.22881**	0.15575	0.19202
		F-skóre	0.13115	**0.19333**	0.08732	0.18717

Tabulka 18: Porovnání výsledných hodnot (úplnost, přesnost a f-skóre) všech implementovaných metod pro sadu článků s č.15