T A

Č R

Software pro generování textu ze zadaných informací

Technická dokumentace

RNDr. Pavel Ševeček

Lingea s.r.o.

Mgr. Michal Kašpar

Lingea s.r.o.

Mgr. Michal Hala

Lingea s.r.o.

Mgr. Michal Auersperger

ÚFAL MFF UK

Mgr. Michal Novák, Ph.D.

ÚFAL MFF UK

doc. RNDr. Pavel Pecina, Ph.D.

ÚFAL MFF UK

Mgr. Dušan Variš, Ph.D.

ÚFAL MFF UK

Brno, leden 2023

Číslo projektu: FW03010656-V2
Projekt Technologické agentury České republiky
Program: TREND 3

LINGEA

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

T A

Č R

Obsah

Úvod

Zpracovat větší množství textu a extrahovat z něj maximum informace v minimálním výsledném textu je úkol, jehož vyřešení má potenciál ušetřit velké množství lidského času.

Tato technická dokumentace popisuje systém, který se snaží zpracovat vstupní texty na základě sémantické podobnosti vět a blízkosti informací v nich obsažených. Výsledkem zpracování jsou informace seskupené podle významové podobnosti, uspořádané podle důležitosti, a s odfiltrovanými sémantickými duplicitami. Vzniklý nástroj umožňuje uživateli zpracování parametrizovat podle jeho potřeb a dosahovat tak relevantních výsledků za různých podmínek (zejména při různých požadavcích). Povedlo se dosáhnout dobře škálovatelné implementace, schopné zpracovávat množství textů, které by člověk četl dlouho, nebo vůbec nebyl schopen v rozumném čase přečíst. Ačkoliv už teď jsou výsledky zpracování tímto nástrojem velmi zajímavé, předpokládáme, že se nástroj v budoucnosti stane podkladem pro další vývoj a ještě detailnější parametrizaci/těsnější adaptaci na potřeby uživatele.

T A

Č R

Analýza funkčních požadavků

V této části je uveden stručný souhrn vytyčených požadavků, které vyplynuly z konzultací mezi hlavním řešitelem projektu – firmou Lingea a kolegy z ÚFAL MFF UK.

Hlavním cílem systému je **spojovat informace z více zdrojů do jednoho celku**. Důležitým požadavkem na takto vzniklý celek je, aby obsahoval pokud možno **všechnu relevantní informaci** z podkladové kolekce a nedocházelo tedy k vypouštění důležitých částí.

Z tohoto pohledu se nabízí užití metod extraktivní sumarizace. Ty prezentují zdrojové části textu přímo, bez parafrází, které v sobě nebezpečí ztráty nebo změny významu obsahují.

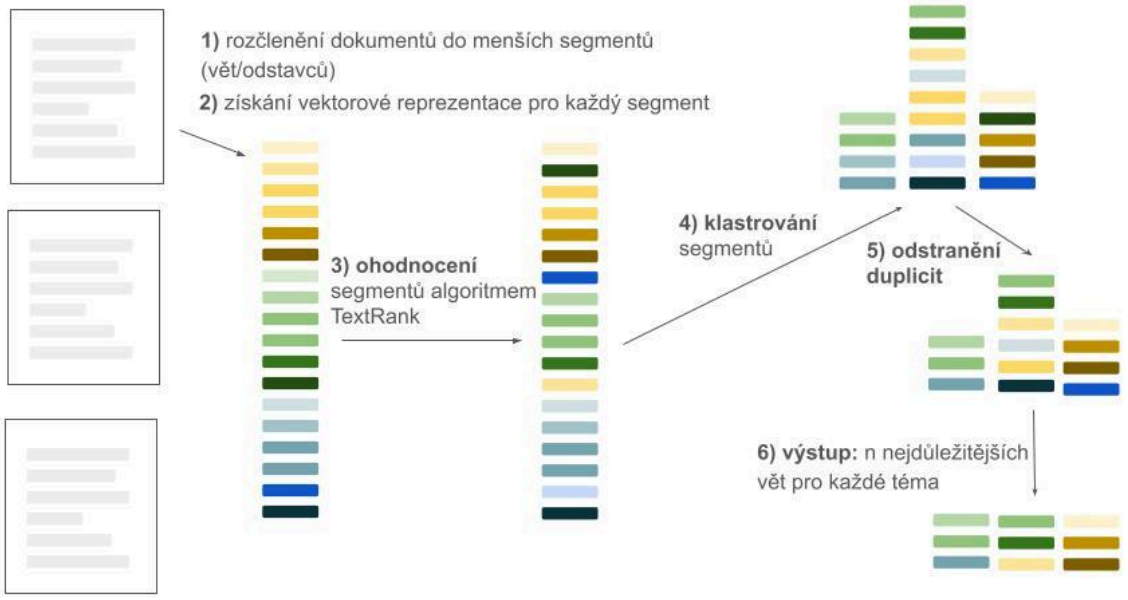
Předpokládá se značná **doménová různorodost zpracovávaných kolekcí** a **různorodost informačních potřeb** uživatelů, kterou má předkládaný systém pokrýt. Je tedy vhodné, aby měl uživatel možnost postupně ovlivňovat složení výstupu podle své aktuální situace. Tento požadavek je v současné době mimo možnosti klasického end-2-end neuronového systému a vede k nutnosti hledat alternativu.

Požadavkem společným pro celý projekt je pak schopnost pracovat s **vícejazyčnými textovými kolekcemi**.

Systém má být jako **samostatná komponenta** začlenitelná do MASAPI Asistenta i do dalších produktů třetích stran.

Technická dokumentace

Navržený systém předkládá uživateli kolekci potenciálně vícejazyčných dokumentů jako sadu témat, kde každé téma je reprezentováno vybranými pasážemi ze vstupní kolekce. Počet témat i počet citovaných segmentů v tématech může uživatel měnit na základě své specifické informační potřeby.



Obrázek 1 - schematický popis systému

Fungování systému je schematicky znázorněno na Obrázku 1. Níže jsou podrobněji popsány jednotlivé kroky.

1) Kolekce dokumentů je segmentována do úseku podle zvolené granularity (věty nebo odstavce, dále jen věty)

2) Věty jsou neuronovou sítí převedeny do vektorové reprezentace. Neuronová síť (enkodér) je (před)trénována tak, aby uměla pracovat s více jazyky a zároveň aby vracela reprezentace víceslovných textových pasáží přímo. Takové reprezentace jsou vhodnější pro určování významové podobnosti mezi větami než například reprezentace získané průměrováním embeddingů slov/tokenů při užití klasického jazykového modelu (Reimers & Gurevych, 2019).

3) Dále probíhá i jejich hierarchické shlukování do skupin podle podobnosti významu (t.j. na základě podobnosti reprezentací z enkodéru). Výstupem hierarchického shlukování je dendrogram,

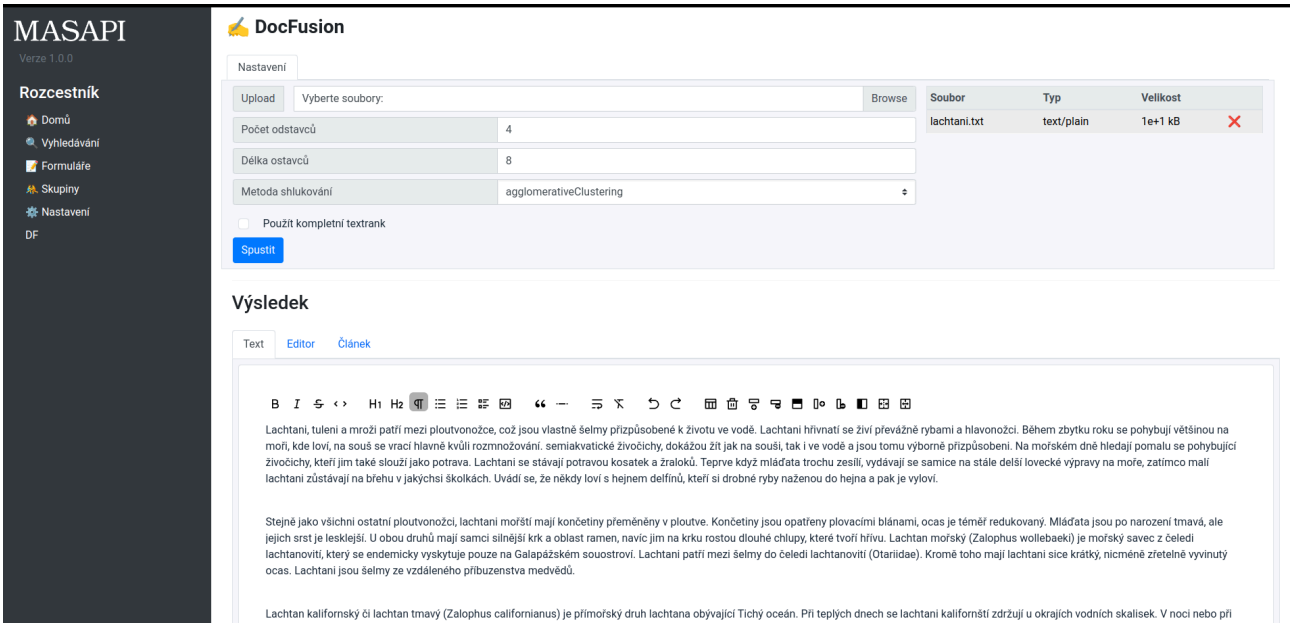
který definuje skladbu témat v závislosti na jejich počtu. Alternativně lze použít shlukovací algoritmus kmeans spolu s dodanými iniciálními centroidy.

4) V dalším kroku jsou segmenty ohodnoceny algoritmem TextRank (Mihalcea & Tarau 2004). Vstupem algoritmu je textová kolekce vyjádřená jako graf: vrcholům grafu odpovídají textové pasáže (věty) a hrany mezi vrcholy jsou ohodnoceny podobností mezi větami. Podobnost je vyjádřena jako kosinová vzdálenost ve vektorovém prostoru větných reprezentací. Výstupem je ohodnocení vět vzhledem k jejich reprezentativnosti pro danou kolekci dokumentů. TextRank lze počítat v globálním kontextu (vznikne graf všech vět), nebo jen v rámci jednotlivých shluků (což umožňuje paměťové škálování).

Při prezentaci výsledků se nejprve řadí věty v jednotlivých tématech a poté i jednotlivá témata podle ohodnocení získaného dříve (bod 3). Na závěr dochází k odstranění duplicit, t.j. takových vět, jejichž (kosinová) podobnost k nějaké větě s vyšším hodnocením přesáhne zvolený práh (5).

Náhled na kolekci dokumentů pak uživatel dotváří postupnou volbou počtu témat, počtu zobrazených vět z každého tématu a volbou míry pro odstranění duplicit. Výpočty provedené v krocích 1 - 4 se už ale při této práci nemění.

K nástroji byl následně implementován HTTP server umožňující integraci s grafickým rozhraním MASAPI, kde byl DocFusion integrován s editorem textu. Viz. obrázek níže.



T A

Č R

Programátorská dokumentace

Systém je implementován primárně jako jednoduše importovatelná Pythonová knihovna, ale obsahuje i jednoduchý HTTP server.

Použité technologie

- Python 3.10.8
- sentence-transformers 2.2.2
- networkx 2.5.1
- rouge-metric 1.0.1

Použití jako knihovna

Pro použití jako knihovna v jiném pythonovém projektu importujte soubor *fusion.py*, s funkcí *summarize*, která má nastavitelné následující parametry:

- *plaintext (string)* - konkatenovaný text dokumentů pro sumarizaci.
- *encoder (SentenceTransformer)* - jazykový model pro výpočet větných embeddingů, doporučujeme *sentence-transformers/distiluse-base-multilingual-cased-v2*.
- *full_textrank (bool, default=False)* - pokud je parametr nastaven na *True*, počítá Textrank napříč všemi páry vět v dokumentu (použitelné jen pro dokumenty do 400-500 kB).
- *n_cluster (int, default=0)* - parametr udávající počet výsledných odstavců. Má smysl uvádět jen při použití aglomerativního shlukování.
- *method (str, default="agglomerativeClustering")* - použitá metoda shlukování. Knihovna podporuje následující methody:
    - *agglomerativeClustering*
    - *kmeans*
- *seeds (List[str], default=[])* - seznam textů použitých jako iniciální centroidy při použití metody *kmeans*.

Použití jako server

Pokud DocFusion běží jako server, komunikuje se přes protokol HTTP s využitím následujícího endpointu:

- cesta: /
- ContentType: *application/x-www-form-urlencoded*
- Method: *POST*
- Body: formulář

- *data (string)* - textový dokument
- *n_clusters (int)* - počet odstavců
- *n_tokens (int)* - počet vět v jednom odstavci
- *method (string)* - použitá shlukovací metoda
- *seeds (string[])* - texty použité jako iniciální centroidy

Popis ověření funkčnosti softwaru

Úloha, kterou daný systém řeší, je nová a nejsou tedy k dispozici alternativy pro automatickou evaluaci. Přesto, v režimu kdy se výstup omezí jen na několik málo témat a několik málo vět, má smysl výstup porovnat s výstupy pro extraktivní sumarizaci. Toto porovnání proběhlo pro češtinu na datasetu SumeCzech (Straka et al. 2018), viz Tabulka 1, a pro angličtinu na datasetu Multi-News (Fabbri et al. 2019) viz Tabulka 2.

CZ	unigram	bigram	lcs
First	0.14	0.02	0.10
Random	0.13	0.01	0.08
Textrank	0.14	0.02	0.09
Tensor2tensor	0.11	0.01	0.09
MASAPI	0.14	0.02	0.09

Tabulka 1 - Rouge metrika (F1) pro extraktivní sumarizaci českého textu na datasetu SumeCzech. Výsledky referenčních modelů převzaty ze Straka et al. (2018). Výstup MASAPI systému je pro konfiguraci: 2 témata, 2 věty na téma.

EN	unigram	bigram	su
First-3	0.39	0.12	0.15
LexRank	0.38	0.13	0.13
TextRank	0.38	0.13	0.14
MASAPI	0.41	0.13	0.17

Tabulka 2 - Rouge metrika (F1) pro extraktivní sumarizaci anglického textu na datasetu Multi-News. Výsledky referenčních modelů převzaty z Fabbri et al. (2019). Výstup MASAPI systému je pro konfiguraci: 3 témata, 5 vět na téma. V souladu s referenčními modely byla každá hypotéza (text sumarizace) omezena na prvních 300 slov.

Uživatelská příručka

Instalace

Před instalací doporučujeme zprovoznit *minicondu* pro jednodušší správu pythonových virtuálních prostředí.

1. Vytvořte a aktivujte virtuální prostředí s pythonem ve verzi 3.10.8
2. Nainstalujte knihovny pomocí
   `pip3 install -r requirements.txt`

Spuštění HTTP serveru z příkazové řádky

1. Aktivujte virtuální prostředí
2. Spusťte server pomocí
   `python3 fusion_server.py`

Poznámka: server se defaultně pouští na portu 8003.

Příklad JS dotazu na server

```
await fetch('http://localhost:8003', {
    method: 'POST',
    headers: {
        'Content-Type': 'application/x-www-form-urlencoded',
    },
    body: new URLSearchParams({
        data: document.text,
        n_clusters: 4,
        n_tokens: 8,
        method: "agglomerativeClustering",
        full_textrank: true,
    }).toString()
})
```

T A

Č R

Reference

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

Straka, M. et al. SumeCzech (2018): Large Czech News-Based Summarization Dataset. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Fabbri, A. et al. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*