Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

DETEKCE OSOBNÍCH ÚDAJŮ POMOCÍ METOD HLUBOKÉHO UČENÍ V NESTRUKTUROVANÉM TEXTU

AUTOREFERÁT DIPLOMOVÉ PRÁCE

Studijní program: Informační systémy a technologie

Specializace: Vývoj informačních systémů

Autor: Bc. David Ondrášek

Vedoucí diplomové práce: Ing. Josef Doležal

Praha, červen 2022

Cíl práce

Hlavním cílem práce je navrhnout a implementovat prototyp nástroje, který pomocí algoritmů hlubokého učení a případně i dalších relevantních metod dokáže v nestrukturovaném textu klasifikovat osobní údaje a umožní tak rychlejší a přesnější zpracování dokumentů v případě potřeby anonymizace těchto údajů.

Klíčovým požadavkem pro tento prototyp je jeho modularita. To znamená, že vyvinutý nástroj musí být dostatečně modulární do té míry, aby se dal jednoduše programaticky přizpůsobit klasifikaci osobních údajů v různých doménách. Různými doménami se pak v tomto případě myslí konkrétní aplikace klasifikačního nástroje na konkrétních typech textových dokumentů s doménově specifickými typy osobních údajů – jmenných entit.

V tomto případě je konkrétní implementace nástroje přizpůsobena úloze povinné anonymizace osobních údajů při nahrávání dokumentů do veřejného registru smluv. Může tedy sloužit jako rozšíření existujícího nástroje "Nástroj pro anonymizaci dokumentů", dostupném na Portálu veřejné správy[1] a pomáhat může všem subjektům povinným v tomto registru smlouvy zveřejňovat. Nástroj je následně evaluován na vhodně vybraném evaluačním datasetu, obsahujícím vhodný typ doménově specifického dokumentu.

Dílčím cílem práce je analyzovat zákony a nařízení, které je potřeba respektovat při anonymizaci osobních údajů při nahrávání dokumentů do Veřejného registru smluv. Na základě této analýzy je vypracován model, který vyjadřuje míru potřeby anonymizace konkrétního identifikátoru v nestrukturovaném textu. Tento model slouží jako podklad pro vývoj konkrétních klasifikátorů v modulární struktuře vyvíjeného prototypu.

Dalším dílčím cílem je provést rešerši existujících metod hlubokého učení, využitelných pro klasifikaci osobních údajů v nestrukturovaném textu, zejména pak metod, které umožňují tuto klasifikaci v textech v českém jazyce. Dochází také k analýze možností kombinace těchto metod za účelem dosažení co nejlepšího výsledku detekce a klasifikace údajů.

Problémy, které jsou v práci řešené jsou tedy následující:

- Absence jednoznačné definice osobního údaje v různých kontextech doménově specifických aplikací.
- Nedostatečně modulární implementace existujících anonymizačních nástrojů.
- Limitující závislost na jazyce dostupných velkých trénovacích datasetů.

Použité metody

Byla provedena systematická rešerše, týkající se problematiky samotné definice toho, co v různých kontextech znamená osobní údaj. Následně došlo k analýze osobních údajů v kontextu Veřejného registru smluv. Byl prozkoumán existující Nástroj pro anonymizaci

[1] https://anonymizace.gov.cz/crossroad

dokumentů, dostupný na Portálu veřejné správy a typy dokumentů, které se běžně do Veřejného registru smluv nahrávají. Na základě toho byla vytvořena komplexní doménově specifická mapa kategorií osobních údajů, která byla následně logicky redukována. Tato mapa potom sloužila jako seznam doménově specifických osobních údajů, které musí být schopen vyvíjený prototyp anonymizovat.

Další část systematické rešerše se zabývala technikami hlubokého učení, vhodnými pro implementaci anonymizačního nástroje. Nejprve došlo k hrubému představení technik z oblasti hlubokého učení (Deep Learning), následované konkrétněji zaměřenými metodami z oblasti zpracování přirozeného jazyka (Natural Language Processing) a rozpoznávání jmenných entit (Named Entity Recognition). Byla představena typická architektura NER modelů a problematika rozpoznávání jmenných entit byla následně přenesena do českého jazyka. V rámci toho byly popsány i dodatečné morfologické znaky slov, které je vhodné využít při trénování kvalitního NER modelu. Také došlo k představení možnosti využití state-of-the-art multilingvních modelů, včetně popisu architektury několika konkrétních modelů.

Systematická část rešerše vhodné literatury proběhla s využitím databáze ACM, Google Scholar a vyhledávačem Univerzity Karlovy UKAŽ. Na ACM a Google Scholar byly vyhledávány práce s technickým zaměřením, týkající se nejnovějších poznatků NLP ve spojení s klasifikací a detekcí identifikátorů v textu nebo využití multilingvních modelů v praktických aplikacích. V databázi UKAŽ pak byly vyhledávány práce zabývající se ochranou osobních údajů a jejich identifikací, kterých nebylo na předchozích databázích nalezeno dostatečné množství.

Výběr vhodných technologií byl proveden definováním seznamu důležitých kritérií a následným porovnáním zkoumaných frameworků a nástrojů vůči tomuto seznamu. Vzhledem k tomu, že dále použitá metodika CRISP-DM nedefinuje výběr programového vybavení a nástrojů, je zde tato část uvedená samostatně.

Pro vývoj prototypu a jeho evaluaci bylo využito upravené metodiky CRISP-DM (Cross-industry standard process for data mining) (Chapman et al., 2000). Ta je vzhledem k charakteru vývoje nástrojů založených na strojovém učení často využívána a je vhodnější alternativou než některé v jiných oblastech hojněji využívané agilní metodiky (Pinhasi, 2021).

Tato metodika je také dostatečně flexibilní, aby se dala jednoduše přenést na řešený problém. Zároveň v kontextu této práce došlo k upravení metodiky tak, aby odpovídala vývoji prototypu nástroje, a ne jeho komplexní komerční verze.

Evaluace prototypu anonymizačního nástroje byla provedena na autorem vytvořeném evaluačním datasetu. Tento dataset obsahuje příklad reálné smlouvy, tedy dokumentu, který je běžně nahráván do Veřejného registru smluv.

Dosažené výsledky

Výsledkem této práce je funkční prototyp modulárního anonymizačního nástroje. Tento nástroj je přitom možné jednoduše programaticky upravovat pro klasifikaci rozdílných

doménově závislých osobních údajů, čímž je vyřešen problém s nejednoznačnou definicí toho, co vlastně osobní údaj je, v závislosti na kontextu analyzovaného textu.

Anonymizační nástroj dosáhl na evaluačním datasetu výsledného F1 score 81,67 %, při využití NER modelu GPU_bert_cased a 66,02 % při použití modelu CPU_fine_nomorph. F1 score 81,67 % je přitom dostatečně vysoké, aby se dal výsledek považovat za uspokojivý. Při analýze výsledků bylo nicméně zjištěno, že má nástroj problémy s klasifikací jmenných entit typu „instituce“ či „doména“.

V případných budoucích iteracích vývoje anonymizačního nástroje stále existuje řada možných zlepšení. Bylo by například vhodné se více zaměřit na doimplementování dodatečných recognizerů či na dotrénování dalších NER modelů. Zároveň by bylo vhodné doplnit mechanismus, který by dokázal na základě primárního či sekundárního typu osobního údaje vytvářet další závislosti, které by šlo dále znovu využít pro kvalitnější celkovou klasifikaci jmenných entit.

I přes výše zmíněné nedostatky se dá výsledný prototyp anonymizačního nástroje označit jako dostatečně kvalitní, aby naplňoval všechny cíle této práce. Nástroj je dostatečně modulární a jednoduše přizpůsobitelný, aby se dal dobře použít pro anonymizaci i dalších doménově závislých osobních údajů a tím vyřešil absenci jednoznačné definice osobního údaje jako takového. Využitím multilingvního modelu byl také částečně vyřešen problém s limitující závislostí na jazyce trénovacích datasetů. Ač je zřejmé, že pro komerční nasazení tohoto anonymizačního nástroje do produkčního prostředí by ještě musel proběhnout další rozsáhlejší vývoj, dá se říct, že konceptuálně tato myšlenka o přizpůsobitelném modulárním anonymizačním nástroji dává smysl a je tak vhodná k dalšímu zkoumání.

Vlastní přínos autora

Cílem práce bylo představit myšlenku modulárního anonymizačního nástroje, který by umožnil snadnější a univerzálnější klasifikaci osobních údajů v nestrukturovaných textech. To se v této práci vytvořením funkčního prototypu podařilo a dá se tedy usoudit, že tato myšlenka dává smysl.

Navržený a implementovaný prototyp nástroje používá state-of-the-art aplikovaných principů NLP v praxi, a to s využitím zajímavých moderních technologií. Těmi jsou v tomto případě hlavně framework spaCy a SDK Presidio (doplněné grafickým frameworkem Streamlit).

Přínosem je také fakt, že celý nástroj dokáže pracovat s nestrukturovanými texty, psanými v českém jazyce. To s sebou nese mimo jisté unikátnosti i další přínos ve smyslu ukázky nutných úprav spaCy i Presidio, které mohou posloužit jako návod dalším výzkumníkům, kteří by chtěli stejné technologie v českém jazyce využít.

Došlo také k vytrénování řady NER modelů (5 variant utilizujících GPU a 3 varianty utilizující CPU) s různými state-of-the-art architekturami, u kterých byla následně evaluována jejich kvalita použitím standartních metrik jako je precision, recall nebo F1 score. Všechny z těchto modelů je možné díky zmiňované modularitě využít

v anonymizačním nástroji a dále i upravit pro detekci vlastních definovaných doménově specifických druhů osobních údajů.

Dále je možné anonymizační nástroj pro doménově specifické použití upravit přiřazováním předpřipravených nebo vytvářením nových recognizerů, sloužících pro další rozšíření seznamu možných druhů klasifikovaných osobních údajů.

V neposlední řadě byl potom autorem vytvořen vlastní evaluační dataset, který byl následně použitý pro evaluaci anonymizačního nástroje upraveného pro zvolené vzorové doménově specifické použití (anonymizace osobních údajů v dokumentech nahrávaných do Veřejného registru smluv). Pro tuto vzorovou aplikaci nástroje byla také předem vytvořena analýza kategorizace osobních údajů v daném kontextu.

Použitá literatura

CHAPMAN, P., J. CLINTON, R. KERBER, T. KHABAZA, T. REINARTZ, C. SHEARER a R. WIRTH, 2000. CRISP-DM 1.0: Step-by-step data mining guide. *undefined* [online]. [vid. 2022-06-16]. Dostupné z: https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

PINHASI, Assaf, 2021. Towards a Development Methodology for Machine Learning — part I. *Medium* [online]. [vid. 2022-06-16]. Dostupné z: https://assaf-pinhasi.medium.com/towards-a-development-methodology-for-machine-learning-part-i-f1050a0bc607