Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

****

Detekce osobních údajů pomocí metod hlubokého učení v nestrukturovaném textu

diplomová PRÁCE

Studijní program: Informační systémy a technologie

Specializace: Vývoj informačních systémů

Autor: Bc. David Ondrášek

Vedoucí diplomové práce: Ing. Josef Doležal

Praha, [měsíc rok]

Poděkování

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean id metus id velit ullamcorper pulvinar. In convallis. Sed convallis magna eu sem. Maecenas lorem. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Ut tempus purus at lorem. Phasellus et lorem id felis nonummy placerat. Mauris elementum mauris vitae tortor. Aliquam id dolor. Vivamus porttitor turpis ac leo.

Abstrakt

Klíčová slova

amet, consectetuer, dolor, Lorem ipsum, sit.

JEL klasifikace

amet, consectetuer, dolor, Lorem ipsum, sit.

Abstract

Keywords

amet, consectetuer, dolor, Lorem ipsum, sit.

JEL Classification

amet, consectetuer, dolor, Lorem ipsum, sit.

Obsah

[Úvod 13](#_Toc105343881)

[1.1 Vymezení problému 14](#_Toc105343882)

[1.1.1 Absence jednoznačné definice osobního údaje 15](#_Toc105343883)

[1.1.2 Limitující závislost na jazyce trénovacích datasetů 15](#_Toc105343884)

[1.1.3 Nemodulární implementace existujících nástrojů 15](#_Toc105343885)

[1.2 Účel a cíle práce 16](#_Toc105343886)

[1.3 Omezení projektu 16](#_Toc105343887)

[1.4 Význam a přínos práce 17](#_Toc105343888)

[2 Rešerše 18](#_Toc105343889)

[2.1 Rešeršní strategie 18](#_Toc105343890)

[2.2 Absence jednoznačné definice osobního údaje 19](#_Toc105343891)

[2.3 Limitující závislost na jazyce trénovacích datasetů 20](#_Toc105343892)

[2.4 Nemodulární implementace existujících nástrojů 20](#_Toc105343893)

[2.5 Shrnutí 21](#_Toc105343894)

[3 Metodika práce 22](#_Toc105343895)

[4 Kategorizace osobních údajů 23](#_Toc105343896)

[4.1 Osobní údaje v kontextu Veřejného registru smluv 24](#_Toc105343897)

[4.1.1 Zákon o registru smluv 24](#_Toc105343898)

[4.1.2 Registr smluv 24](#_Toc105343899)

[4.1.3 Nástroj pro anonymizaci dokumentů 25](#_Toc105343900)

[4.2 Typy vstupních dokumentů 26](#_Toc105343901)

[4.3 Kategorizace osobních údajů 27](#_Toc105343902)

[4.3.1 Kompletní seznam 29](#_Toc105343903)

[4.4 Mapa kategorií 30](#_Toc105343904)

[4.4.1 Jmenné identifikátory 32](#_Toc105343905)

[4.4.2 Číselné identifikátory 32](#_Toc105343906)

[4.4.3 Znakové identifikátory 32](#_Toc105343907)

[4.4.4 Lokalizační identifikátory 32](#_Toc105343908)

[4.4.5 Ostatní identifikátory 32](#_Toc105343909)

[4.4.6 Shrnutí 33](#_Toc105343910)

[5 Použité techniky hlubokého učení 34](#_Toc105343911)

[5.1 Hluboké učení 34](#_Toc105343912)

[5.2 Natural Language Processing 35](#_Toc105343913)

[5.3 Named Entity Recognition 37](#_Toc105343914)

[5.4 NER v českém jazyce 39](#_Toc105343915)

[5.4.1 Stematizace a lematizace 40](#_Toc105343916)

[5.4.2 POS 40](#_Toc105343917)

[5.5 Multilingvní modely 41](#_Toc105343918)

[5.5.1 mBERT 42](#_Toc105343919)

[5.5.2 XLM 42](#_Toc105343920)

[5.5.3 XML-RoBERTa 43](#_Toc105343921)

[5.6 Context-based modely 44](#_Toc105343922)

[5.7 Transfer learning 44](#_Toc105343923)

[5.7.1 Fine-tuning 44](#_Toc105343924)

[5.8 Sekvenční modely 44](#_Toc105343925)

[5.9 Meta klasifikátory 44](#_Toc105343926)

[6 Výběr technologií 45](#_Toc105343927)

[6.1 NLP frameworky 45](#_Toc105343928)

[6.1.1 Kritéria pro výběr NLP frameworku 45](#_Toc105343929)

[6.1.2 Stanford CoreNLP (Stanford NER) 46](#_Toc105343930)

[6.1.3 NLTK 46](#_Toc105343931)

[6.1.4 Flair 46](#_Toc105343932)

[6.1.5 SpaCy 47](#_Toc105343933)

[6.1.6 Porovnání 47](#_Toc105343934)

[6.2 Vývojové prostředí 48](#_Toc105343935)

[7 Dataset 49](#_Toc105343936)

[7.1 Anotace NER datasetů 49](#_Toc105343937)

[7.1.1 Granularita NER datasetů 49](#_Toc105343938)

[7.1.2 Způsoby anotace jmenných entit 50](#_Toc105343939)

[7.2 Dostupné datasety 50](#_Toc105343940)

[7.2.1 Czech Named Entity Corpus (CNEC) 50](#_Toc105343941)

[7.2.2 BSNLP 50](#_Toc105343942)

[7.2.3 SumeCzech-NER 51](#_Toc105343943)

[7.2.4 Shrnutí 51](#_Toc105343944)

[7.3 Zpracování datasetu 51](#_Toc105343945)

[7.3.1 Načtení CNEC\_extended 52](#_Toc105343946)

[7.3.2 Načtení CNEC 2.0 52](#_Toc105343947)

[7.4 Analýza dat 54](#_Toc105343948)

[7.4.1 Kategorie klasifikovaných jmenných entit 54](#_Toc105343949)

[7.4.2 Úprava datasetu 55](#_Toc105343950)

[Vývoj prototypu 60](#_Toc105343951)

[7.5 Trénování NER modelu 60](#_Toc105343952)

[7.6 Návrh anonymizačního nástroje 66](#_Toc105343953)

[8 Evaluace modelu 72](#_Toc105343954)

[8.1 Hodnocení naměřených kvalitativních indikátorů 72](#_Toc105343955)

[8.2 Automatická evaluace 72](#_Toc105343956)

[8.3 Evaluace člověkem 72](#_Toc105343957)

[8.4 Zhodnocení výsledků a porovnání 72](#_Toc105343958)

[Závěr 74](#_Toc105343959)

[Použitá literatura 75](#_Toc105343960)

[Přílohy I](#_Toc105343961)

[Příloha A: Název první přílohy I](#_Toc105343962)

[Příloha B: Název druhé přílohy I](#_Toc105343963)

Seznam obrázků

Seznam tabulek

[Tabulka 1: Mapa kategorií osobních údajů 31](#_Toc102753828)

Seznam výpisů programového kódu

Seznam zkratek

NER Named Entity Recognition

NE Named Entity

NLP Natural Language Processing

OCR

GRU

LSTM

RNN

BRNN

TODO rozšířit dle již hotového textu

Úvod

GLOBAL TODO

* Bohužel v češtině kvůli nedostatku anotovaných datasetů se asi k kontext-citlivému rozpoznávání stejně moc nedostanu
* Model bude muset být modulární – meta-classifier, přičemž základními modely budou ty co umějí základní NER – jména, bydliště, atd.
* Czech Named Entity Corpus
* Kontext-citlivé rozpoznávání
  + Anonymizovat ne jen jedno slovo, ale celou větu kde se více identifikačních údajů nachází
  + Určovat kontext na základě více vyhledaných sekundárních identifikátorů
* Spojit každou kapitolu/krok s nějakou metodikou
* Odkazy na GitHub, atd. udělat podle Ondry (indexace na konci strany)
* Přidat labely ke kusům kódu
* Přidat skript do mého repa
* Doplnit do praktické části více podnadpisů?

Nezapomenout přiložit ke zdrojákům i všechny config fily a hlavně streamlit appku

Není to tak dávno, kdy si pod pojmem automatizace šlo představit pouze zavádění robotů a automatických procedur do továren a tento termín byl sám o sobě spojován pouze s průmyslovým sektorem. Dnes je s příchodem digitálních technologií vše jinak a stejný proces automatizace můžeme pozorovat i v oblasti služeb. Tam dochází k automatizaci rutinních kancelářských činností a celkově vzato i většiny nekreativních prací.

Jednou z těchto prací je i detekce a klasifikace osobních údajů v různých textech. Tato práce je ve velké části případů vykonávána kvůli povinnosti anonymizace těchto údajů na základě různých zákonných opatření, týkajících se ochrany osobních údajů, v čele se známým evropským nařízením GDPR.

Problematika detekce identifikátorů v textu je součástí vědecké domény s názvem Named Entity Recognition (NER), která se soustavně vyvíjí již mnoho let. Samotný termín Named Entity byl definován R. Grishmanem již v roce 1996 (Grishman, 1996). Od té doby došlo k objevu řady významných metod detekce identifikátorů, jako jsou některé slovníkové metody, lookup tabulky nebo regulární výrazy (Nadeau, 2007). Poslední dobou se však jako nejvhodnější způsob pro detekci identifikátorů v textu jeví využití hlubokého učení, zejména pak technik Natural Language Processing (NLP). To totiž umožňuje nastavit detekční pravidla, která se vyznačují vyšší abstrakcí a dokážou tak zachytit i případy, které se v textu vymykají běžným slovním vyjádřením. To v důsledku znamená, že techniky hlubokého učení mají v NER výsadní postavení a v současné době vykazují nejlepší výsledky (Xu et al., 2021).

Detekce a klasifikace osobních údajů je tedy typově specifickou aplikací NER a potažmo NLP na konkrétní sadu dokumentů, obsahujících osobní údaje. Problém v tomto případě vzniká nejasnou definicí toho, co to osobní údaj vlastně je, respektive tím, že se osobním údajem mohou stát i informace, které nejsou explicitně unikátní (Mihulková, 2018). V některých případech je tedy nutné tuto informaci (zda se jedná o osobní údaj) získat přímo z textu na základě širšího kontextu. To existující zahraniční vědecké práce, které se detekcí osobních údajů již zabývají spíše ignorují a soustředí se více na kvalitativní zpřesnění detekce identifikátorů na základě pokročilejších statistických modelů (Silva et al., 2020).

Ač se některé současné práce problematikou NER s ohledem na širší kontext textu zabývají, je tato funkčnost modelů hlubokého učení stále problematická a složitá (Williams, 2017). Místo hledání obecného řešení, které by dokázalo rozlišovat kontext ve všech typech textů, je v této situaci vhodnějším řešením tuto funkčnost aplikovat pouze na konkrétní typ dokumentu, čímž se sníží nutný stupeň abstrakce potřebný pro zpracování textu. Aplikací tohoto principu v případě detekce osobních údajů by tak mělo být možné získat kvalitnější výsledky.

Zároveň se současná řešení, která se detekcí osobních údajů zabývají nedají považovat jako dostatečně modulární tak, aby se dala aplikovat v různých doménách na konkrétní specifické typy jmenných entit a rozdílné druhy vstupních dokumentů. Klíčovým úkolem této práce je tedy zkoumat možnosti, jak vytvořit koncept nástroje, který se dá jednoduše upravit pro konkrétní využití.

Tato diplomová práce pracuje s předem vytvořeným konceptem modulárního nástroje a zkoumá možnost využití kontext-citlivé NER v oblasti detekce osobních údajů, a to na sadě dokumentů, které jsou běžně nahrávány do veřejného registru smluv. Pomocí state-of-the-art metod hlubokého učení je v ní navržen a implementován nástroj, který je volným rozšířením anonymizačního nástroje “Nástroj pro anonymizaci dokumentů” dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>). Pomocí tohoto nástroje lze anonymizovat nahrané dokumenty, lze to však provádět pouze manuálně. Nástroj, navržený v této práci, by tuto činnost po integraci do původního nástroje dokázal provádět i automaticky.

TODO rozšířit podle budoucího stavu

## Vymezení problému

Subjekty, které mají povinnost nahrávat smlouvy do veřejného registru smluv čelí při tomto procesu řadě těžkostem, které dokonce mohou v některých případech zamezit vkládání smluv pracovníkům bez dostatečného právnického a technického vzdělání.

Jedním z těchto problémů je povinnost anonymizovat osobní údaje v nahraných dokumentech a přílohách. Tato povinnost je stanovena dle 160/1999 Sb. Zákona o svobodném přístupu k informacím a následně i 101/200 Sb. Zákona o ochraně osobních údajů, který upravuje předpisy GDPR. Samotná činnost detekce osobních údajů pro jejich následnou anonymizaci je nejen jednotvárná a nezáživná, ale existuje zde velké riziko lidské chyby, ať už přehlédnutím důležitého údaje nebo i jeho špatnou klasifikací a interpretací.

Následující problémy jsou společné pro jakoukoliv detekci osobních údajů v nestrukturovaném textu. Tímto pojmem (nestrukturovaný text) se označuje jakýkoliv text, který nemá dopředu pevně stanovenou strukturu v podobě předem definovaných kapitol či povinných textových částí. Neobsahuje také žádná povinná textová pole, do kterých by byly zapisovány konkrétní informace (např. pole pro jméno a příjmení).

### Absence jednoznačné definice osobního údaje

Dle Zákona o zpracování osobních údajů lze termín osobní údaj definovat jako jakoukoli informaci, která se týká identifikované nebo identifikovatelné žijící osoby (Mihulková, 2018). Tato definice je z hlediska implementace anonymizačního nástroje velmi široká. Neexistuje žádný souhrnný seznam, který by definoval, co všechno se dá považovat za osobní údaj. Je proto třeba definovat jednotlivé osobní, případně citlivé osobní údaje pro konkrétní typ dokumentů, které bude nástroj zpracovávat. Tento seznam osobních údajů je rozdílný pro různé typy dokumentů. Pro funkční implementaci klasifikačního nástroje je ale tento seznam klíčový.

### Limitující závislost na jazyce trénovacích datasetů

Pro úspěšné využití jakékoliv techniky hlubokého učení je potřeba mít k dispozici kvalitní dataset, na jehož datech může být vznikající model natrénován. V oblasti NLP (Natural Language Processing) je tak klíčové mít buď k dispozici dataset v požadovaném jazyce nebo předtrénovaný multilingvní model, který by se dal následně metodou transfer learning přetrénovat pro požadované využití (Moberg, 2020). V případě detekce osobních údajů se navíc v existujících datových sadách obsahujících ukázky různých smluv logicky žádné osobní údaje nevyskytují, protože tyto smlouvy již anonymizací musely projít.

### Nemodulární implementace existujících nástrojů

Jednotlivé již existující nástroje, určené pro detekci identifikátorů v nestrukturovaném textu nejsou dostatečně modulární, aby se daly upravit pro konkrétní doménově specifickou aplikaci. Zároveň se většina existujících akademických řešení soustředí na detekci pouze několika druhů jmenných entit, což vzhledem k větší šíři detekce konkrétních údajů v doménově specifické aplikaci nestačí. Pro větší flexibilitu a modularitu vyvíjeného konceptu je tedy třeba rozšířit tradiční seznam jmenných entit tak, aby jejich šíře lépe odpovídala možnému rozšíření nástroje na konkrétní sadu dokumentů a identifikátorů.

Zároveň jsou vyvíjené NER modely často vytvářeny pouze jako akademické koncepty a necílí na žádnou konkrétnější doménu do které by měly spadat zpracovávané dokumenty. Také vzhledem k jejich abstraktnímu využití nejsou některé tyto modely veřejně dostupné.

## Účel a cíle práce

Hlavním cílem této 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.

Cílem této práce je vytvořit nástroj, který bude 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 bude konkrétní implementace nástroje přizpůsobena povinné anonymizaci osobních údajů při nahrávání dokumentů do veřejného registru smluv. Bude tedy sloužit jako rozšíření existujícího nástroje “Nástroj pro anonymizaci dokumentů” dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>) a pomáhat může všem subjektům povinným v tomto registru smlouvy zveřejňovat. Nástroj bude následně evaluován na vhodně vybrané sadě dokumentů.

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 bude vypracován model, který bude vyjadřovat míru potřeby anonymizace konkrétního identifikátoru v nestrukturovaném textu. Tento model bude sloužit 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. Dojde také k analýze možností kombinace těchto metod za účelem dosažení co nejlepšího výsledku detekce a klasifikace údajů.

## Omezení projektu

Vzhledem ke komplexnosti celé problematiky ochrany osobních údajů a kvůli specifickému charakteru osobních údajů v různých druzích textů je vyvíjený koncept navržen modulárně tak, aby se dal přizpůsobit konkrétní aplikaci klasifikace osobních údajů. V další fázi je však omezen jen na jednu konkrétní sadu dokumentů, a to na sadu textových dokumentů, které jsou běžně nahrávány do veřejného registru smluv.

Dále je práce omezena limitujícím počtem českých datasetů vhodných pro trénování vyvíjeného modelu. Vytváření nutně komplexnějších datových sad není součástí této práce. To může znamenat menší přesnost detekce než u modelů vyvíjených zahraničními výzkumníky a trénovaných na rozsáhlejších a lépe připravených textových datech v jiném jazyce.

Práce se také nesnaží zkoumat nové algoritmy hlubokého učení. Místo toho v ní jde o výběr state-of-the-art metod hlubokého učení a jejich aplikaci na danou problematiku. Znamená to, že přesnost detekce vyvíjeného nástroje bude s největší pravděpodobností maximálně stejně velká, jako přesnost detekce u existujících metod hlubokého učení, z kterých bude vycházet. Stejně tak i funkcionalita nástroje v oblasti rozpoznávání širšího kontextu textu bude omezená současným stavem poznání v této oblasti.

## Význam a přínos práce

Přínosem této práce je samotný koncept modulárního nástroje pro detekci osobních údajů v nestrukturovaných textech. Modularita nástroje by měla umožňovat ostatním výzkumníkům převzít část řešení a upravit ji pro jiný typ dokumentů a rozdílné druhy jmenných entit. Jinak řečeno tedy využít stejný postup pro vytváření požadavků jako v této práci a následně model přetrénovat na nové sadě dokumentů případně upravit existující klasifikátory tak, aby vyhovovaly konkrétnímu použití. Model pak bude vhodný k detekci osobních údajů i v jiném případě užití, jako je například anonymizace osobních údajů při digitalizaci veřejně přístupného archivu nebo při pseudonymizaci API requestů, obsahujících některé citlivější informace.

Dalším přínosem této práce je upravení modulárního nástroje pro konkrétní aplikaci, kterou je detekce osobních údajů v nestrukturovaných dokumentech běžně nahrávaných do veřejného registru smluv. Tato část slouží jako konkrétní proof of concept využitelnosti algoritmů hlubokého učení a kontext-citlivé NER při zpracování osobních údajů v doménově specifické oblasti.

Vyvinutý nástroj pomůže všem subjektům, povinným v tomto registru smlouvy zveřejňovat, v rychlejší a přesnější detekci osobních údajů v nahrávaných dokumentech za účelem jejich následné anonymizace. V ideálním případě bude moci být tento nástroj použit jako rozšíření existujícího nástroje “Nástroj pro anonymizaci dokumentů”, dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>). Přínosem se dá také označit fakt, že nástroj bude detekovat osobní údaje v textech v českém jazyce, což vzhledem ke složitosti českého jazyka a stavu vědeckého poznání popisovaného běžně na anglických příkladech není triviální záležitost.

V neposlední řadě se dá za přínos považovat i fakt, že model pro klasifikaci osobních údajů bude určen pro klasifikaci údajů v českém jazyce. Zároveň zde bude docházet k využití moderního state-of-the-art NLP frameworku spacy a ke customizaci open source nástroje presidio. Oba tyto nástroje bude upraveny pro efektivní užití v českém jazyce.

# Rešerše

## Rešeršní strategie

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. Nebyl vybrán žádný časový filtr, ale bylo nastaveno řazení článků od nejnovějších ke starším a následně i řazení dle relevance. Při každém hledání bylo analyzováno prvních 20 článků, přičemž největší důraz byl při tom kladen na praktické využití technik NLP při detekci textových identifikátorů co nejvíce podobných osobním údajům.

Při vyhledávání docházelo ke spojením několika klíčových slov logickými operátory.

**Klíčová slova využitá ve vyhledávacích řetězcích ACM a Google Scholar**

*NLP, Natural Language Processing, NER, Named Entity Recognition, multilingual, PII, personally identifiable information, Cognitive Data Capture, privacy, data privacy, private, personal, anonymization, pseudonymization, data security*

**Příklady vyhledávacích řetězců pro ACM a Google Scholar**

* (NLP OR Natural Language Processing) AND multilingual
* (NER OR Named Entity Recognition) AND (PII OR personally identifiable information)
* (NER OR Named Entity Recognition) AND (PII OR personally identifiable information) AND (NLP OR Natural Language Processing)
* (NLP OR Natural Language Processing) AND (Cognitive Data Capture)
* (NLP OR Natural Language Processing) AND privacy
* (NLP OR Natural Language Processing) AND data privacy
* (NLP OR Natural Language Processing) AND private
* (NLP OR Natural Language Processing) AND personal
* (NLP OR Natural Language Processing) anonymization

**Klíčová slova využitá ve vyhledávacích řetězcích UKAŽ**

*Ochrana osobních údajů, automatizovaný, GDPR*

**Příklady vyhledávacích řetězců pro UKAŽ:**

* jakékoliv pole obsahuje Ochrana osobních údajů
* jakékoliv pole obsahuje Ochrana osobních údajů AND jakékoliv pole obsahuje automatizovan
* jakékoliv pole obsahuje GDPR
* jakékoliv pole obsahuje GDPR AND jakékoliv pole obsahuje automatizovan

V nesystematické části rešerše došlo k prohledávání internetu se zaměřením na firmy, zabývající se digitalizací dokumentů a na konkrétní informace týkající se ochrany osobních údajů (například definice samotného osobního údaje). Pozorně prostudována byla například webová stránka Evropské komise, která srozumitelnou formou čtenáře seznamuje se základními pojmy v oblasti ochrany údajů v Evropské unii. (EK). Došlo také na odborné konzultace, ve kterých byla doporučena odborná literatura týkající se ochrany osobních údajů.

## Absence jednoznačné definice osobního údaje

Termín osobní údaj, definovaný již v kapitole Vymezení problému, může mít ze své podstaty mnoho významů. Informace je totiž za osobní údaj uvažovaná až ve chvíli, kdy vede k přímé a jednoznačné identifikaci jednotlivce. Často se tedy stává, že osobním údajem se údaj stává až ve chvíli, když se vyskytuje v textu společně s dalšími údaji, které dohromady dávají ucelenou informací o konkrétní osobě (Mihulková, 2018).

Do tohoto vymezení se mísí i koncepce práva na soukromí, která je ukotvena v evropských, kontinentálních právních kulturách. V této koncepci se jako osobní údaj uvažuje i jakákoliv informace, kterou jednotlivec sám o sobě nechce sdílet ve veřejném prostoru (Kubica, 2019).

Jako jednotlivá podkategorie osobních údajů se rozlišují tzv. zvláštní osobní údaje. Tyto údaje mohou být buď citlivějšího charakteru a vypovídat o rasovém původu, politických názorech nebo sexuálním vyznání a celkově vzato jsou klasifikovány tak, že mohou subjekt poškodit ve společnosti. Mezi tyto údaje potom patří i biometrické údaje, jakožto jednoznačné fyzické identifikátory subjektu (MVČR). Zpracování údajů spadajících do této podkategorie se řídí přísnějšími pravidly a při implementaci anonymizačního nástroje to musí být bráno v potaz.

Problém vzniká ve státní správě, kde může kolidovat 106/1999 Sb. Zákon o svobodném přístupu k informacím a 101/200 Sb. Zákon o ochraně osobních údajů. Občan má totiž právo přístupu k informacím, které nejsou omezené nařízeními o ochraně osobních údajů. Tyto informace potom může dostat v anonymizované formě. Ve chvíli, kdy ale údaj, jako je například křestní jméno není považován za osobní údaj, nespadá pod zákon o ochraně osobních údajů a měl by se dostat k žadateli v plné, neanonymizované podobě (Gealfow et al., 2019).

Na základě těchto tezí můžeme tušit, že pojem osobní údaj se liší podle konkrétní aplikace využití informací, a i podle sektoru, ve kterém ke zpracování dat dochází. Ostatně i například v (EK) a (Mihulková) se uvedené příklady osobních údajů liší.

## Limitující závislost na jazyce trénovacích datasetů

V oblasti strojového učení, a zejména pak v podoblasti NLP je často adresován problém nelokalizovaných vstupních dat. Vzhledem k dnešnímu stavu vědeckého světa se totiž výzkumné práce píšou ve většině případech v angličtině a v důsledku toho existuje i největší množství předtrénovaných modelů, které jsou založeny právě na anglických korpusech (Pod pojmem korpus se rozumí množina slov, která může být využita pro trénink modelu. Většinou jsou slova v korpusu spojená vazbami k doméně, ke které je korpus plánovitě vytvořen). Angličtina je také výhodná z pohledu Named Entity Recognition (NER), protože například oproti češtině se v ní nevyskytují ve větším množství pády, rody či větší množství nepravidelných tvarů množných čísel podstatných jmen.

Možným řešením problému s nedostatkem kvalitních korpusů různých jazyků mohou být moderní modely, jako je například model ELECTRA, který je vhodný pro trénování sítě transformerů i menším výpočetním výkonem a korpusem menší velikosti (Clark et al., 2020). Populárním řešením je také využití revolučního modelu BERT, publikovaného týmem výzkumníků z Google v roce 2018 (Devlin et al., 2019), zejména pak jeho multilingvní varianty M-BERT, která v současné době podporuje 104 jazyků, včetně češtiny (BERT GITHUB). Problém v tomto případě může být imbalance korpusu, na kterém byl model trénován (celosvětové záznamy Wikipedie). Kvůli rozdílnému poměru celkového množství textu každého jazyka v trénovacím datasetu se tak u modelu M-BERT dají u méně zastoupených jazyků čekat horší výsledky (Chau et al., 2020).

Jako další state-of-the-art multilingvní modely se dají označit modely XLM a jeho upravená varianta XLM-R, které mají ve srovnání provedeném Johnem Mobergem na 15 jazycích o několik procent lepší úspěšnost (Moberg, 2020).

## Nemodulární implementace existujících nástrojů

Jak již bylo dříve zmíněno, pro anonymizaci osobních údaje je vhodné zvolit metody NLP, které jsou součástí technik hlubokého učení (Qu el al., 2021; Silva et al., 2020; Ellman, 2018). Jednotlivé metody z této oblasti lze pro dosažení optimálních výsledků potom různě kombinovat. Většina existujících prací se však zabývá pouze zkoumáním a implementací těchto metod zvlášť.

Silva ve své práci využívá NER a porovnává v ní vhodné NLP toolkity, které se hodí pro klasifikaci osobních identifikátorů v smlouvách. S největším F1 score mu v jeho případě funguje toolkit Stanford CoreNLP (Silva et al., 2020). Jeho výsledek pak rozporuje např. Mendels, kterému v jeho případě nejlepší F1 score vychází u toolkitu Flair (Mendels, 2020).

Mendels sám uvádí další vhodné metody vhodné pro detekci identifikátorů, jako je využití regulárních výrazů uvnitř klasifikačních vrstev nebo vytvoření blacklistů s textovými řetězci s větší pravděpodobností výskytu hledaného výrazu.

Zajímavá je práce Mathiase Ellmanna, který se pomocí NLP snaží detekovat duplikáty „issue trackerů”. Narozdíl od předchozích autorů jde více do hloubky (vzhledem k tomu, že pracuje pouze s binárním klasifikátorem si to může dovolit) a tím dokáže lépe zpracovat samotnou extrakci příznaků z textu na základě sémantické analýzy textu originálního „issue“. Tím potom dochází k rozšířenému chápání širšího kontextu a menšímu zaměření na samotné identifikátory (Ellman, 2018).

Širším kontextem se zabývá také Jake Williams, který ve své práci vytváří „interní“ a „externí“ model. Interní model je určen pro lokální klasifikaci identifikátorů jako je název města nebo jméno. Externí model potom pracuje s již předzpracovaným textem připraveným interním modelem. Tento text obsahuje klasifikované identifikátory nahrazené klíčovými slovy, které by měly být z hlediska širšího kontextu relevantními pro lepší pochopení textu (Williams, 2017).

Chen Qu se zabývá konceptem typických architektur anonymizačních modelů a pokládá otázku, jak s anonymizovanými daty dokážou pracovat další vrstvy hlubokých neuronových sítí, které byly předučeny na neanonymizovaných datech (Qu el al., 2021).

Ve všech výše zmíněných pracích se vyskytuje jeden problém. Ač všechny trénované klasifikátory dosahují kvalitních výsledků, chybí zde kvalitnější uzpůsobení modelu konkrétní doméně. Silva například trénuje model, který dokáže klasifikovat osobní údaje ve smlouvách, jenže šíře kategorií osobních údajů je moc malá na to, aby došlo k detekci všech různých druhů osobních údajů (Silva et al., 2020).

Texty, na kterých byl model trénován totiž obsahují anotace pouze pro základní druhy jmenných entit, jako jsou jména, adresy, lokace a podobně. To pramení především z nejběžněji dostupných a v akademické sféře používaných veřejných datasetů, které jsou často anotovány dle standardu, který definuje pravidelná konference CONLL (Yadav et al., 2019).

Zároveň ve vědeckých pracích zcela chybí důraz na modularitu vyvíjených řešení. To je však většinově dáno tím, že práce se zabývají implementací konkrétního modelu, spíše než implementací konkrétního řešení.

## Shrnutí

Z provedené rešerše vychází, že zde existuje prostor pro vytvoření modulárního nástroje pro klasifikaci a anonymizaci osobních údajů, který bude možné upravovat pro konkrétní aplikační využití.

Při tom je nejprve je na základě provedené analýzy potřeba definovat požadavky na vývoj nástroje, zejména potom definovat konkrétní typy osobních údajů pro konkrétní aplikaci. Tyto typy osobních údajů se přitom mohou pro různé typy aplikací lišit.

Z rešerše také vychází, na co je třeba dávat si pozor při výběru modelu a při vytváření architektury nástroje a kombinaci různých NLP metod. Je také potřeba vybrat vhodný dataset, který musí reflektovat nejen jazyk analyzovaného textu při detekci osobních údajů, ale i vybraný model hluboké neuronové sítě, který tomu musí být přizpůsoben.

# Metodika práce

Vzhledem k charakteru této práce byla pro implementační část vybrána lehká metodika ML Project Checklist, vytvořená Aurélienem Géronem jako součást knihy *Hands-on Machine Learning with Scikit-Learn and TensorFlow,* která definuje fáze vývoje kvalitního modelu hluboké neuronové sítě v souvislosti s ideálním výběrem vhodných technik využitelných při vývoji požadované aplikace (Géron, 2017).

Samotný vyvinutý prototyp bude evaluován pomocí nejznámějších metrik využívaných při evaluaci modelů hlubokých neuronových sítí, jako je F1 score, Precision, Recall, aj. (Mishra, 2018). Evaluace bude probíhat na sadě dokumentů, které se běžně nahrávají do veřejného registru smluv, typicky tedy na jednotlivých smlouvách.

Analýza principů hlubokého učení, využitelných pro klasifikaci osobních údajů a také derivace požadavků na vyvíjený nástroj ze zákonů a zákonných opatření bude probíhat cestou systematické rešerše.

TODO rozšířit

TODO napsat k jednotlivým kapitolám jakou část metodiky plní

# Kategorizace osobních údajů

TODO zatím jen pracovní název

* Název by měl být jiný, nejde totiž o příznaky, jako spíš o kategorie pro hiearchical text classification
* Vypracovat nějaký vzorec pro určení toho kolik jednotlivých osobních údajů potřebuji k identifikaci osoby (nejednoznačných osobních údajů)

Paralela s ML checklistem – vlastně tím plníme první kapitolu

TODO

* Napsat na portál veřejného registru a poptat se jich s čím se tam potýkají
  + co je nejsložitější na zpracování a detekci
  + jaké dokumenty je možné nahrávat
  + jaká existují rizika v případě že nebudou osobní údaje správně anomymizovány
  + s jakými PII se nejčastěji setkávají a jaké se dají považovat za obtížněji detekovatelné
  + existují nějaké speciální PII, které jsou vyloženě spojené s registrem smluv?
  + Poslat jim můj seznam a zeptat se jestli by ho dokázali rozšířit
  + přiložit komunikaci do příloh
  + napsat i dalším tvůrcům anonymizačních nástrojů na jakém principu je jejich nástroj založen…jestli používají techniky DL
* co hrozí při špatném zpracování osobních údajů
* definovat základní pojmy zpracování údajů – správce, zpracovavatel, atd.
* srovnání již existujících řešení na českém trhu – nástroje na anonymizace údajů při správě dokumentů
  + důraz na to, které z těchto rešení nabízí přímo auto klasifikaci údajů
  + důraz na to kdo dokáže číst i dle kontextu

TODO předmluva

* oslí můstek k tomu abych se dostal k veřejnému registru smluv
* mělo by to být psané tak, aby to bylo více obecné a registr smluv je jenom konkrétní aplikace použitých technik

## Osobní údaje v kontextu Veřejného registru smluv

### Zákon o registru smluv

340/2015 Sb. Zákon o registru smluv, který nabyl platnosti 1. 7. 2016 a který byl 2x novelizován, nejprve v roce 2017 (zákonem 249/2017 Sb.), dále v roce 2019 (zákonem 177/2019 Sb.), hovoří o zvláštních podmínkách účinnosti některých smluv a o uveřejňování těchto smluv ve veřejném registru smluv (Bouda, 2019).

Tento zákon stanovuje povinnost k uveřejňování soukromoprávních smluv a smluv o poskytnutí dotace nebo návratné finanční výpomoci, jejichž alespoň jednou stranou je některý z okruhu vymezených povinných subjektů. Tímto subjektem je běžně míněn některý veřejnoprávní subjekt, který je definován v uzavřeném výčtu dle §2 tohoto zákona. Jsou zde také definované podmínky pro povinnost uveřejnění smlouvy, jako je například minimální výše hodnoty plnění smlouvy 50 000 Kč nebo plnění smlouvy převážně na území České republiky. Subjekt musí dle tohoto zákona také splňovat podmínky platnosti uveřejnění, jako je například nutnost uveřejnění 30 dnů od jejího uzavření nebo (Kraus, 2022).

TODO možné rozšíření…uvézt i některé výjimky uveřejnění?

### Registr smluv

TODO možné rozšíření…jaké smlouvy do registru smluv patří

Registr smluv je informačním systém veřejné správy, který slouží k uveřejňování smluv dle platného zákona. Správcem tohoto IS je Ministerstvo vnitra, které ale žádným způsobem neodpovídá za správnost nahraných smluv. Samotná kontrola správnosti smlouvy a především kontrola znečitelnění obsahu (neboli anonymizace či úplné odstranění některých údajů) spadá na subjekt, který smlouvu uveřejňuje (Kraus, 2022).

Smlouvy je možné do registru zasílat přes grafické rozhraní skrz elektronický formulář, dostupný na Portálu veřejné správy nebo přes systém datových schránek.

Prostřednictvím registru smluv není možné zveřejňovat informace, které nelze poskytnout dle 106/1999 Sb. Zákona o svobodném přístupu k informacím. Těmi jsou v kontextu této práce a možností algoritmického zpracování dat pomocí hlubokého učení osobní údaje a obchodní tajemství, jehož detekce však není cílem této práce.

Práce s osobními údaji je dále omezená 110/209 Sb. Zákonem o zpracování osobních údajů a potažmo tedy i nařízením Evropského parlamentu a rady EU GDPR. Vzhledem k tomu, že při uveřejnění smlouvy není k dispozici oprávněný zájem ani jiný právní titul, který by tuto operaci ospravedlňoval (ÚOOÚ, 2016) a získání souhlasu by znamenalo další interakci se subjektem smlouvy, je nutné všechny osobní údaje dle jejich platné definice v uveřejňovaných smlouvách řádně anonymizovat. V případě neexistující či nekompletní anonymizace může Úřad pro ochranu osobních údajů (ÚOOÚ) udělit odpovědnému subjektu pokutu nebo se případně poškozená fyzická osoba může domáhat náhrady způsobené újmy (Kraus, 2022).

### Nástroj pro anonymizaci dokumentů

Ministerstvo vnitra poskytuje přes Portál veřejné správy pro účely anonymizace textů uveřejňovaných smluv Nástroj pro anonymizaci dokumentů, který však mohou využívat pouze orgány veřejné moci. Tento nástroj je dostupný na adrese <https://anonymizace.gov.cz/crossroad/>.

Tento nástroj umožňuje řádným způsobem (znečitelněním všech vrstev textu) anonymizovat nahraný dokument. Dle textového popisu pro nepřihlášené uživatele však nenabízí automatickou detekci klíčových identifikátorů, a to ani v podobě jejich vyhledávání přes předdefinované šablony bez znalosti širšího kontextu textu. Model vyvíjený v další části této práce by se tedy mohl stát dalším rozšířením tohoto nástroje a pomoci tak pracovníkům orgánů veřejné moci lépe a rychleji anonymizovat nahrávané dokumenty.

Na obrázku 1 lze vidět část úvodní intrografiky dostupné na webu Portálu veřejné správy. Lze z něho vyčíst, že strojové zpracování textu, neboli Optical Character Recognition (OCR) při nahrání textu v obrazovém formátu nástroj již zvládá. Začernění osobních údajů a obchodního tajemství se ale stále musí provádět ručně.



Obrázek 1: Nástroj pro anonymizaci dokumentů – intrografika

TODO

* Omezení tohoto nástroje
* Vyjádření o nástroji dle mailu
* Existující externí nástroje? Několik jich uvést a napsat k nim jestli nabízí detekci pomocí hlubokého učení a jestli je detekce dostatečně domain specific

## Typy vstupních dokumentů

Smlouvy smí do registru smluv být nahrávané pouze v elektronické podobě v několika povolených formátech. Elektronický obraz textového obsahu smlouvy se do registru zveřejňuje v otevřeném a strojově čitelném formátu, případně v otevřeném formátu umožňujícím úplně strojové zpracování textového obsahu (Kraus, 2022).

Otevřený a strojově čitelný formát definuje zákon 106/1999 Sb. o svobodném přístupu k informacím.

Otevřený formát nesmí být závislý na konkrétním technickém a programovém vybavení. To obvykle splňují formáty, které jsou udržovány neproprietárními standardizačními organizacemi a které nemají právní omezení pro jejich používání (např. PDF, ODT).

Strojově čitelný formát obsahuje takovou vnitřní strukturu, která umožňuje při dalším elektronickém zpracování extrahovat ze souboru čitelný a nijak nepozměněný text od originálu.

Do 31. prosince 2023 je možné do registru smluv nahrát i dokumenty, které sice umožňují úplné strojové zpracování textového obsahu, ale které v sobě neobsahují další informace o vnitřní struktuře dat, jako jsou informace o nadpisech, podnadpisech nebo poznámkách pod čarou. Vždy je ale nutné, aby dokument obsahoval čitelnou textovou vrstvu.

Tato informace je důležitá pro implementaci modelu hlubokého učení, který bude text zpracovávat. Znamená totiž, že metadata u jednotlivých textů nemusí být vždy dostupná a nelze se na ně tedy spolehnout.

Do registru smluv se nahrává smlouva ve stejném formátu, v jakém byla v originále podepsána, a to včetně všech jejích dodatků, které by měly být součástí jednoho dokumentu. Subjekt, který do registru smlouvu nahrává musí vyplnit i metadata. Ta by měla obsahovat základní informace o smlouvě, jako je identifikace smluvních stran, vymezení předmětu smlouvy, datum jejího uzavření a případně, pokud jí lze určit, i cenu předmětu vyjednávání. Informace uvedené v metadatech však musí splňovat stejná nařízení o nezveřejnitelných informacích jako samotný text smlouvy.

Součástí nahrání smlouvy mohou být i přílohy. Obrazové přílohy obvykle nutnosti nahrávání nenaplňují. Textové přílohy musí splňovat stejné formální požadavky pro správný formát jako samotná smlouva.

## Kategorizace osobních údajů

Na základě předpokládaného výskytu osobních údajů, běžně se vyskytujících ve smlouvách nahrávaných do registru smluv podle Metodického návodu k aplikaci Zákona o registru smluv (Kraus, 2022) a na základě úvahy autora byl vytvořen seznam specifických osobních údajů, které se pravděpodobně budou vyskytovat ve zpracovávaných dokumentech (TODO tohle by mělo být lépe podloženo, mělo by to být minimálně ještě na základě e-mailové komunikace s portálem veřejné správy a na základě analýzy smluv v datasetu. Taky můžu udělat nějakou analýzu návrhových vzorů pro smouvy).

Údaje jsou roztříděné do několika základních kategorií primárně dle formátu, ve kterém se údaj v textu vyskytuje (sekvence čísel, znaků, apod.).

**Jmenné identifikátory**

* Jméno
* Příjmení
* Příjmení za svobodna
* E-mail osobní
* E-mail pracovní
* Přihlašovací údaje (login)

**Číselné identifikátory**

* Rodné číslo/číslo pojištěnce
* Číslo občanského průkazu (OP)
* Číslo pasu
* Číslo řidičského průkazu
* Číslo platební karty
* Číslo jiné osobní karty (např. členské karty)
* Daňové identifikační číslo (DIČ)
* Číslo bankovního účtu
* Sdružené inkaso plateb obyvatelstva (SIPO)
* IBAN
* Telefonní číslo osobní
* Telefonní číslo pracovní
* Datum narození
* Věk

**Znakové identifikátory**

* Státní poznávací značka (SPZ)
* MAC adresa
* IP adresa
* Internetová doména
* Podpis
* Elektronický podpis
* Elektronický klíč
* Biometrická data
* Identifikační číslo vozidla (VIN)
* Parcelní číslo
* Sériové číslo
* GPS souřadnice

**Nezařazené**

* Adresa
* Pohlaví
* Rasa
* Náboženské vyznání
* Sexuální vyznání
* Titul
* Pracovní zařazení
* Politická příslušnost
* Státní příslušnost
* Obchodní tajemství
* Cena zakázky

**Identifikátory nepodléhající anonymizaci**

* Název firmy
* Identifikační číslo osoby (IČO)
* Id datové schránky

Tento základní seznam ale není k dalšímu zpracování dostatečný, protože nereflektuje zvláštní situace, kdy některý údaj povinné anonymizaci může a nemusí podléhat. Zároveň nevystihuje jednotlivé vazby údajů mezi sebou a také v něm chybí rozpad na konkrétnější identifikátory nižší úrovně.

### Kompletní seznam

V další fázi tedy byla po hlubší analýze vytvořena tabulka, která obsahuje podrobnější seznam jednotlivých identifikátorů, rozepsaných více dopodrobna.

Každému identifikátoru bylo přiděleno unikátní číslo ID, název, příznak, zda je identifikátor primární nebo sekundární, seznam závislých identifikátorů a příznak, zda je identifikátor specifický využitím v dokumentech typu smlouva.

Pole **Typ identifikátoru** vyjadřuje, zda je daný identifikátor v kontextu vždy osobním údajem (v tomto případě je označen jako primární) nebo se osobním údajem stává až při vzniku vazby na další identifikátor (v tomto případě je označen jako identifikátor sekundární). To si lze jednoduše představit na jednoduchém příkladu křestního jména a příjmení. Samo o sobě není křestní jméno tak silným identifikátorem, aby dokázalo identifikovat konkrétní osobu na kterou odkazuje. Ve spojení s příjmením se ovšem množina lidí, která může být tímto identifikátorem označena značně zmenšuje a v tomto kontextu lze tedy již prohlásit, že se jedná o osobní údaj dle platné definice.

Existují samozřejmě výjimky, kdy se jedním celým jménem (křestní jméno + příjmení) dá označit více osob a naopak křestní jméno nebo příjmení může být tak unikátní, že je přiřazeno jen jedné osobě. Hranice může být v tomto případě velmi tenká. Tato práce se však ve své základní podobě zabývá standartními situacemi.

V poli **Závislé identifikátory (ID)** jsou v případě sekundárních identifikátorů uvedeny odkazy na další identifikátory, na kterých je v kontextu textu daný identifikátor závislý ve smyslu toho, že pokud daná vazba v textu existuje, stává se kombinace těchto identifikátorů osobním údajem.

V poli **Identifikátor specifický pro dokument typu smlouva** se nachází informace, zda je nutná detekce tohoto identifikátoru specifická přímo pro využití detektoru v kontextu smluv. Například identifikátor, který odkazuje na adresu bydliště je v kontextu běžného textu téměř vždy osobním údajem. V kontextu dokumentu typu smlouva zde však existuje závislost na tom, jestli se jedná o adresu-bydliště fyzické osoby či o adresu-sídlo právnické osoby. V druhém jmenovaném případě totiž tento údaj nepodléhá povinné anonymizaci.

To, jestli je identifikátor specifický jen pro smlouvy, je také důležité rozdělení jednotlivých detektorů do vhodných modulů tak, aby se v případě využití modulárního přístupu k vývoji aplikace daly některé obecné části znovu využít i v případě přepracování modelu na detekci osobních údajů v jiném kontextu.

V případě, kdy identifikátor nepodléhá povinné anonymizace, nejsou vyplněny hodnoty v žádném dalším sloupci kromě ID a název identifikátoru.

Kompletní seznam identifikátorů lze naleznout v příloze XXX. TODO zjistit jak správně vložit kompletní tabulku do přílohy a jak na ní správně odkazovat

## Mapa kategorií

V kompletním seznamu se pracuje s počtem 53 identifikátorů. Tento počet je v další fázi redukován pomocí postupného logického zjednodušování, kdy je bráno v potaz konkrétní využití aplikace na detekci osobních údajů na dokumentech typu smlouva. Také zde dochází k odstranění některých duplicitních příznaků u jednotlivých identifikátorů.

Kvůli své podstatě sekundárních identifikátorů vede tato mapa k následnému vytvoření n-rozměrné vazební matice binárních hodnot, kdy n je rovno celkovému počtu zpracovávaných identifikátorů.

Zjednodušenou mapu kategorií osobních údajů lze najít v tabulce 1.

TODO sekundárním identifikátorům přidat jinou jistotu určení a pak je v grafické aplikaci označovat zvlášť.

Taky by ještě možná nebylo od věci něco redukovat

|  |  |  |  |
| --- | --- | --- | --- |
| **Mapa kategorií osobních údajů** | | | |
|  |
| **ID** | **Identifikátor** | **Typ identifikátoru** | **Závislé identifikátory (ID)** |  |
| **Jmenné identifikátory** | | | |  |
| 1 | Křestní jméno | sekundární | 2, 7, 9, 53 |  |
| 2 | Příjmení | sekundární | 1, 7, 9, 53, 56 |  |
| 4 | Jméno + Příjmení | primární | 1, 2 |  |
| 6 | E-mail s unikátním jménem | primární |  |  |
| 7 | E-mail pracovní s obecným jménem | sekundární | 1, 2, 9, 53 |  |
| 9 | Přihlašovací údaje (login) | sekundární | 1, 2, 7, 10, 32, 33 |  |
| 10 | Název společnosti | - | - |  |
| **Číselné identifikátory** | | | |  |
| 12 | Rodné číslo/číslo pojištěnce | primární |  |  |
| 13 | Číslo OP | primární |  |  |
| 14 | Číslo pasu | primární |  |  |
| 15 | Číslo platební karty | primární |  |  |
| 16 | Číslo jiné karty | sekundární | vydavatel karty, 10 |  |
| 17 | Číslo bankovního účtu | sekundární | 59 |  |
| 19 | IBAN | sekundární | 59 |  |
| 21 | Telefonní číslo | primární |  |  |
| 23 | Číslo řidičského průkazu | primární |  |  |
| 24 | Věk (sec) | sekundární | 1, 2, 3, 53 |  |
| 25 | SIPO | sekundární | 59 |  |
| **Znakové identifikátory** | | | |  |
| 30 | SPZ | sekundární | 1, 2, záznam o jízdě |  |
| 31 | MAC adresa | primární |  |  |
| 32 | IP adresa | sekundární | 9, 10, 53 |  |
| 33 | Doména | sekundární | 9, 10, 53 |  |
| 38 | VIN | sekundární | 1, 2, záznam o jízdě |  |
| 40 | Sériové číslo | sekundární | 1, 2, 10 |  |
| 42 | Daňové identifikační číslo (DIČ) | sekundární | 59 |  |
| **Lokalizační identifikátory** | | | |  |
| 45 | Adresa – bydliště | primární |  |  |
| 46 | Adresa – sídlo PO nebo FOP | - | - |  |
| **Ostatní identifikátory** | | | |  |
| 49 | Náboženské vyznání | sekundární | nepřímá závislost |  |
| 52 | Titul | sekundární | nepřímá závislost |  |
| 53 | Pracovní zařazení | sekundární | 1, 2, 7, 10, 32, 33 |  |
| 54 | Politická příslušnost | sekundární | nepřímá závislost |  |
| 55 | Státní příslušnost | sekundární | nepřímá závislost |  |
| 56 | Datum narození | sekundární | 2, 3, 10 |  |
| 59 | Právnická osoba (PO)/Fyzická osoba podnikající (FOP) | - | - |  |

Tabulka : Mapa kategorií osobních údajů

### Jmenné identifikátory

V této skupině došlo ke sjednocení všech identifikátorů vyjadřujících jméno osoby do 3 kategorií: křestní jméno, příjmení a celé jméno, které vzniká spojením 2 předchozích. Kvůli tomu, že všechna celá jména budou v konečném důsledku stejně podléhat anonymizaci, identifikátor označující jednatele/zástupce firmy nedává v samostatném významu smysl.

E-mail byl zařazen do 2 kategorií: E-mail s unikátním jménem (ať už pracovní nebo osobní) a e-mail s obecným jménem, který si lze představit například jako info@vse.cz. Ten se dá za osobní údaj uvažovat jen pokud jsou známé dodatečné informace jako například příjmení osoby.

Název společnosti sám o sobě anonymizaci nepodléhá. Je ho ovšem potřeba brát v potaz kvůli vazbám dalších sekundárních identifikátorů, které jsou na tomto údaji závislé.

### Číselné identifikátory

Primární identifikátory uvedené v kompletním seznamu zůstávají principiálně ve stejné podobě, jen dochází ke sjednocení osobního a pracovního telefonního čísla a především k vytvoření vazby na identifikátoru *Právnická osoba (PO)/Fyzická osoba podnikající (FOP)*.

Identifikátory označující identifikační číslo osoby (IČO) byly zcela odstraněny, protože tento údaj v daném kontextu nepodléhá anonymizaci.

### Znakové identifikátory

Z této skupiny byl odstraněn identifikátor vyjadřující podpis. Vyvíjený program totiž nezpracovává obrazová data a k vyfiltrování podpisu v obrazové podobě by tak mělo dojít ještě v předchozím zpracování. Pokud je podpis uveden jen v textovém formátu, spadá již pod identifikátor vyjadřující jméno osoby.

Stejně tak byla vyřazena i textová reprezentace elektronického podpisu a klíče. U těchto údajů se totiž tento formát zobrazení nepředpokládá. Stejně tak jsou vyjmuta i biometrická data, která by se ve smlouvě vůbec neměla v žádné formě vyskytovat.

### Lokalizační identifikátory

Adresa obecně vždy podléhá anonymizaci, pokud není v kontextu smlouvy zároveň sídlem PO nebo i FOP. Z toho se dá tedy odvodit, že tento identifikátor je ve své základní podobě primární, avšak dá se na něj pohlížet i jako na sekundární, přičemž jako sekundární anonymizaci nepodléhá.

### Ostatní identifikátory

Identifikátor *Rasa* byl z této skupiny odstraněn kvůli velmi nepravděpodobnému výskytu ve smlouvách. To stejné platí i o identifikátoru *Sexuální vyznání*. *Pohlaví* bylo odstraněno kvůli malé identifikační vypovídající hodnotě a zároveň kvůli složitostem při detekci pohlaví v českém jazyce, kde se dá pohlaví určit jen podle tvaru slova.

Cena zakázky by měla být dle metodických doporučení uvedena přímo v předmětu smlouvy a anonymizaci tak podléhá jen ve velmi specifických případech, které jsou velmi složité a značně nekonkrétní. Cena může souviset i s obchodním tajemstvím, jehož detekce není v této verzi aplikace umožněna.

### Shrnutí

Po redukci kompletního seznamu identifikátorů se podařilo snížit počet identifikátorů z 53 na 34. Do vytvořené mapy se dají v budoucnu při rozšíření stávajícího využití nebo při novém stanovení typu analyzovaných dokumentů přidávat i další identifikátory, které mohou být navázané na již stávající. Díky modulární struktuře vyvíjeného modelu by mělo být možné relativně jednoduše rozšířit stávající funkcionalitu vyvíjeného klasifikátoru.

TODO Udělat nějakou paralelu s ML checklistem jako závěrečné zhodnocení kapitoly

TODO dále více rozřadit klasifikátory na generic a domain-specific z hlediska NER

# Použité techniky hlubokého učení

Úvod k hlubokému učení

* Obecná teorie od základu ML po větší zúžení a uvedení do oblasti NLP a NER.
* Zaměření na multilingvní modely a na modely zpracovávající kontext textu
* Model by měl být vytvořen z více modelů závislých na typu údaje a využití tak, aby byla možnost ho rozšiřovat o další modely více závislé na kontextu

TODO

* Přidat nějaké obrázky
* Rozdělit na více podnadpisů

## Hluboké učení

Hlubokým učením (deep learning) se nazývá podoblast strojového učení (machine learning). Hluboké neuronové sítě, tedy neuronové sítě patřící právě do oblasti hlubokého učení vynikají především větším stupněm abstrakce, která je způsobena větším počtem skrytých vrstev (hidden layers).

Známým příkladem, který ukazuje, jak větší počet skrytých vrstev dokáže ovlivňovat složitost požadované operace, je trénink neuronové sítě tak, aby dokázala provádět operaci XOR. Zatímco binární operace jako AND a OR jsou lineárně separovatelné v dvourozměrném prostoru jednou polorovinou, u XOR už toto tvrzení neplatí a pro úspěšné vyjádření této funkce je potřeba do neuronové sítě přidat skrytou vrstvu.

Zároveň je dokázáno, že jakákoliv funkce vyjádřená ve dvourozměrném prostoru může být neuronovou sítí simulována dostatečným počtem výpočetních jednotek (computation units) v pouze jedné skryté vrstvě, přičemž toto tvrzení vytváří paralelu s Fourierovou transformací. Tento přístup je však pro praktické použití nevýhodný, protože s narůstající přesností simulace funkce roste exponenciálně i počet výpočetních jednotek ve skryté vrstvě a tím pádem i celková výpočetní náročnost (Wang, 2017).

TODO obrázek ukázky vyšší abstrakce, možná přímo ten XOR?

Jako hluboké neuronové sítě jsou definovány takové neuronové sítě, které mají 3 a více skrytých vrstev a pracují s více úrovněmi reprezentace informací. Princip většího počtu skrytých vrstev je bližší fungování lidského mozku, který také tíhne k vytváření větší abstrakce a postupnému spojování dílčích informací (Bengio a Delalleau, 2011).

Dá se tedy říci, že zatímco mělké neuronové sítě (s jednou skrytou vrstvou) se budou při úloze detekce automobilu v obrázku snažit hledat automobil jako celek, hluboké neuronové sítě budou v jednotlivých vrstvách nejdříve detekovat dílčí části automobilu, jako jsou kola, karoserie, a podobně.

Eldan a Shamir ve své práci *The Power of Depth for Feedforward Neural Netoworsk* (2015) zjistili a dokázali, že více skrytých vrstev neuronové sítě dosahuje z hlediska výpočetní náročnosti v poměru s přesností modelu lepší výsledky. Zároveň však mohou nastat situace, kdy je přidávání dalších vrstev spíše na škodu, protože model může v důsledku ve vyšších vrstvách vytvářet i abstraktní vazby takové úrovně, že z hlediska řešeného problému nemusejí dávat smysl.

Zároveň obecně platí, že čím je model neuronové sítě hlubší, tím je i náročnější na trénování algoritmem zpětné propagace chyby (backpropagation algorithm), protože trénování se značně zpomaluje a zároveň se zvyšuje problém uváznutí v lokálních minimech kvůli špatné volbě počátečních vah. Také je v těchto případech pro trénování potřeba většího množství trénovacích dat (Hinton, 2006).

Tento problém adresoval v roce 2006 ve své práci *A fast learning algorithm for deep belief nets* Geoffrey Hinton, který přišel s greedy algoritmem, který umožňuje trénovat jednotlivé vrstvy odděleně, a tím otevřel cestu k řadě dalším výzkumům v oblasti hlubokého učení pomocí jemného ladění (fine tuning) ve vyšší vrstvách, bez přístupu k obrovskému množství trénovacích dat.

* Co to je, jak to navazuje na strojové učení, co se v něm využívá, co to vlastně je, jaký je matematický princip a jaké jsou běžné typy modelů – svm, neuronovka,…

## Natural Language Processing

Techniky patřící do oblasti Natural Language Processing (dále NLP) se zabývají strojovým zpracováním jazyka. Jedná se tedy o činnost, která je pro člověka relativně jednoduchá (ve většině případech), ale v případě strojového zpracování zde věda naráží na obtížnost vyjádření myšlenky, kontextu či emocí pomocí matematických pravidel (Johri et al., 2021).

NLP samo o sobě není přímou součástí hlubokého učení. Výzkum strojového zpracování jazyka začal ještě dávno před tím, než se objevily první záznamy o hlubokém učení a o hlubokých neuronových sítích. Nejprve se svět NLP dlouho zabýval vytvořením automatického překladače, který by dokázal spolehlivě automaticky překládat mezi různými jazyky. To vedlo k vytvoření časově náročného projektu, který si dal za cíl vytvořit překladové strojové slovníky. To se podařilo Georgesi Artstrounimu, který si nechal roku 1933 patentovat zařízení s názvem „Mechanical brain“, které sloužilo jako mechanický překladač (machinetranslate.org, nedatováno).

Tento přístup se ale z praktického používání v reálném světě ukázal jako nepoužitelný, protože tradiční slovníkový přístup si nedokáže poradit s gramatikou. V roce 1957 přišel Noam Chomsky s teorií o lingvistických strukturách, která měla obecně definovaným setem pravidel definovat matematické vztahy mezi jednotlivými slovy ve větách. Později se ale ukázalo, že tento přístup nedokáže nikdy komplexně gramatiku vystihnou kvůli rozdílům v jednotlivých jazycích (Johri et al., 2021).

Dalším evolučním krokem byla tokenizace slov ve větě, jejich klasifikace a zpracovávání kontextu dle pořadí tokenů ve stromové struktuře představující větu. Hojně se při tom využívala rekurze, kdy docházelo k postupnému zpracovávání všech tokenů, dokud nebyl rozpoznán význam textu. Taková technika NLP se nazývá Augmented Transition Network (ATN) a byla popsána roku 1970 Williamem Woodsem (Woods, 1970).

Všechny výše popsané metody měly jednu zásadní nevýhodu: všechna pravidla pro zpracování textů musela být přesně definována a algoritmicky popsána, což z tohoto úkolu dělalo nesmírně časově a intelektuálně náročnou činnost (nehledě na to, že způsob, jak některá pravidla programaticky vyjádřit ani nebyl nalezen).

V 80. letech 20. století se oblast NLP začala pomalu prolínat se světem strojového učení, které pomocí například rozhodovacích stromů dokázalo přesněji interpretovat jednotlivá gramatická a syntaktická pravidla na základě tréninku modelu na reálných datech. Stále zde ale existovala velká oblast nevyřešených problémů s určováním významu slov na základě kontextu textu nebo rozdílné chápání textu na základě vztahu pisatele.

Nejnovějšími technikami NLP se tedy staly techniky hlubokého učení, které umožňují právě vyšší abstrakci chápání textu.

Klíčovou technikou se stalo vytváření matic reprezentujících vztahy mezi jednotlivými tokeny (těmi můžou být jak jednotlivá slova, tak například i dvojice slov, atd.). Samotné vytvoření této matice se provádí trénováním modelu na korpusu slov, které by se měly týkat stejné či příbuzné problematiky jako na kterou bude využitý i výsledný NLP model.

Běžnými technikami jsou například TF-IDF, One-Hot Encodings, Word Embeddings, Word2Vec a další. Cílem je vytvořit takovou n-rozměrnou matici, aby slova která jsou si významem a využitím ve větách příbuzná, měla k sobě co nejbližší vzdálenost (Nigam, 2021).

Základních typů neuronových sítí, které jsou v současné době považované jako nejvhodnější pro použití v NLP je několik. V první řadě jde o rekurentní neuronovou síť (Recurrent Neural Network – RNN), která dokáže zpracovat vstupní data (jedno slovo/jeden token) s ohledem na data zpracovaná při předchozím průchodu. Tím pádem dokáže při zpracování např. věty brát v potaz i kontext předchozích slov. Problém RNN spočívá v tom, že fungují pouze jednosměrně. To znamená, že slova následující ve větě později už nemohou ovlivnit zpracovaný význam slov předcházejících dříve ve větě. Zároveň zde existuje problém takzvaného mizejícího gradientu, kdy v delších textech postupně mizí vazba na předchozí zpracovaná data (Goldberg, 2017).

V současné době se tedy spíše využívají modifikované verze RNN, které tyto problémy adresují. Nejznámějšími jsou modely Gated Recurrent Unit (GRU) a Long Short Term Memory (LSTM), které řeší problém mizejícího gradientu představením takzvaných bran (gate), které umožňují lepší kontrolu nad oddělenou pamětí, představující kontext předchozích tokenů (Goldberg, 2017).

Problém nemožnosti vytvoření vazby na tokenech zpracovávaných později ve větě je řešen obousměrným RNN (Bidirectional RNN – BRNN). V tomto případě probíhá zpracovávání tokenů obousměrně, takže při zpracování konkrétního tokenu BRNN pracuje s error gradienty obou RNN (Goldberg, 2017).

* Nějaký obrázek který představuje třeba ve stromu rozložení věty
* Jaké jsou základní úlohy

## Named Entity Recognition

Named Entity Recognition (NER) je jedním z hlavních dílčích úkolů NLP, který představuje detekci a klasifikaci jmenných identifikátorů v analyzovaném textu. Tato technika má potom využití jak sama o sobě, tak i jako předzpracování textu před dalšími NLP úlohami jako je porozumění textu, vytvoření shrnutí textu nebo strojového překladu (Li et al., 2020).

Jmenné identifikátory (Named Entities – NE) jsou obecně rozdělovány do dvou skupin: generické jmenné identifikátory (generic NE) a jmenné identifikátory specifické pro konkrétní kontext textu (domain-specific NE).

V literatuře lze identifikovat 4 hlavní přístupy v řešení NER (Li et al., 2020). Prvním je rule-based approach, založený na definici konkrétních pravidel pro každý typ jmenné entity. V tomto případě se často využívají prostředky jako jsou slovníky (dictionaries), regulární výrazy, syntakticky-lexikální vzory nebo ručně vytvořená syntaktická pravidla (Petasis et al., 2001). Systémy vytvořené tímto způsobem se vyznačují vysokou přesností (v terminologii strojového učení pojem *precision* = poměr true positives/(true positives + false positives)), ale malým recallem (poměr true positives/(true positives+ false negatives)) (Li et al., 2020).

Dalším přístupem je tzv. *Unsupervised Learning* neboli učení bez učitele, konkrétně potom využití clusteringu, kdy dochází k segmentaci jmenných identifikátorů do různých skupin dle obecně se vyskytujících lexikálních vzorů ve větším textovém korpusu (Nadeau, 2007). V případě unsupervised learning není potřeba anotovat data v trénovacím datasetu. Tím pádem existuje menší práce s přípravou trénovacích dat a tím i větší flexibilita při aplikování této techniky pro konkrétní doménu.

Přístup aplikující *Supervized Learning* neboli učení s učitelem je ve velké míře závislý na vektoru příznaků. Příznaky mohou být mnoha druhů. Některý vektor může označovat všechna velká písmena nebo číslice v textu, jiný speciální znaky či fráze s větší důležitostí (Settles, 2004), apod. Tyto vektory pak slouží jako abstraktní vrstva nad vektorem reprezentujícím samotný text. Klasifikátor je potom vytvořen jako multi-class (neuronová síť umožňuje klasifikaci několika jmenných entit) některým z tradičních algoritmů strojového učení, jako jsou rozhodovací stromy (decision trees), SVM (Support Vector Machine) nebo CRF (Conditional Random Fields) (Li et al., 2020).

Posledním přístupem je využití algoritmů hlubokého učení. V tomto případě se jedná (ve většině případech) také o učení s učitelem. Oproti předchozímu přístupu se však využití hlubokého učení liší ve vytváření příznakových vektorů (tato část se běžně označuje jako feature engineering). Hluboká neuronová síť totiž dokáže odhalovat příznaky již sama. To značně zmenšuje nutnou odbornost v doméně zpracování daného textu, protože odpadá práce s odvozováním toho, jaké příznaky nejvíce souvisí s analyzovaným textem a jak je z textu derivovat.

Běžně využívaná architektura moderního NER implementovaného pomocí hlubokého učení je následující (jednotlivé části jsou uvedeny od vstupní vrstvy představující analyzovaný text po výstupní vrstvu, která představuje označené jmenné identifikátory v textu) (Li et al., 2020):

* Vrstva představující distribuované reprezentaci vstupních dat (distributed vector representation)
* Context encoder
* Tag decoder

Pod vrstvou představující **distribuovanou reprezentaci vstupních dat** si lze představit N-rozměrnou matici, která je uspořádána takovým způsobem, který reprezentuje vztahy mezi významy jednotlivých slov. Slova, která jsou si příbuzná, pak mají mezi sebou kratší vzdálenost. N-1 potom představuje množství příznaků.

Rozdíl mezi nedistribuovanou a distribuovanou reprezentací dat ukazuje obrázek 2: *Rozdíl mezi distribuovaným a nedistribuovaným rozdělením vstupních dat*. Vlevo je vidět řídká matice s nedistribuovanou reprezentací dat (one-hot vektor), ve které neexistuje žádný vztah mezi vstupními daty (geometrické obrazce). Vpravo se nachází matice s distribuovanou reprezentací, kde již lze vztahy mezi obrazci odvodit.

Diagram

Description automatically generated

Obrázek : Rozdíl mezi distribuovaným a nedistribuovaným rozdělením vstupních dat (Ganesh, 2019)

Matice představující distribuované rozdělení vstupních dat se v kontextu NLP a NER nazývá word embeddings. Vrstvu word embeddings je nutné před zařazením do architektury NER neuronové sítě vytrénovat na vhodném korpusu tak, aby došlo ke správné distribuci reprezentace slov.

Správně vytrénovaná vrstva word embeddings se ukazuje v posledním vývoji NER jako klíčová. Moderní NER modely tedy pracují s word embeddings předtrénovaným na obrovských korpusech, které je následně pomocí techniky fine-tuning doučeno na konkrétnějších datech více závislých na dané doméně (Li et al., 2020).

**Context encoder**, jak již název této části NER architektury naznačuje, se zabývá zpracováním kontextu. To znamená, že se tato část hluboké neuronové sítě stará o vyšší abstrakci nad textem předzpracovaným pomocí word embeddings. Vytvářený model v tomto případě vyjadřuje syntaktické, gramatické či významové závislosti, které se neuronová síť učí skrz supervised learning z anotovaných jmenných identifikátorů v textu.

Nejběžněji využívanými architekturami pro tuto část jsou obousměrné rekurentní neuronové sítě (BRNN), společně s jejich odvozenými variantami LSTM nebo GRU. Zajímavou alternativou (postavený na této architektuře je například populární model BERT) je využití transformerů, které jsou v současné době považovány za state-of-the-art v oblasti NLP. Pomocí transformerů a techniky fine-tuning lze funkci předtrénovaných modelů založených na transformerech poupravit pro využití v NER (Giacaglia, 2021).

Výstupní vrstvou typické moderní NER architektury je **Tag decoder**. Ten si bere jako vstup kontextově závislou reprezentaci textových dat a produkuje původní text proložený anotovanými jmennými identifikátory.

Typicky se používá jednoduchý multi-layer perceptron se softmax vrstvou, případně pak Conditional Random Fields (CRF). Některé novější práce pak pracují i s RNN, případně potom s upravenou RNN nazývající se Pointer Network (Li et al., 2020).

TODO na základě této hrubé architektury popsat části mého vyvinutého modelu

TODO rozšíření

* Více rozvést jednotlivé varianty word embeddings, zároveň i oddělit word-level representation, character-level a hybrid…možná to více rozvést v části o datasetu a přípravě dat?
* V případě že to budu používat i zmínit, že do word embeddings se dají přidávat i vlastní příznaky, pokud do toho chci nějak zasahovat
* Více rozepsat konkrétní modely pro context encoder

## NER v českém jazyce

Většina výzkumů týkajících se NER a NLP se tradičně provádí v angličtině (model je trénován na datasetech v anglickém jazyce). Dá se tedy říct, že jako state-of-the-art stav poznání současného NER se dají považovat modely, které se běžně evaluují na anglických korpusech.

Anglický jazyk se však dá svojí složitostí z pohledu sémantiky považovat jako spíše jednodušší. Celkové množství tvarů a variací jednotlivých slov (neboli morfologická složitost) je zde výrazně menší než v češtině. Stejně tak i například skladba vět je v češtině typicky složitější a vyžaduje větší povědomí o širším kontextu. Příkladem může být větší návaznost na rody podstatných jmen a na odkazování se na ně ve složitějších větách.

Konkrétním příkladem může být porovnání celkového množství použitých slov v jejich různých tvarech na 2 korpusech, obou zhruba o velikosti 10 milionů slov, přičemž jeden z těch to korpusů je turecký a druhý anglický. Turečtina je přitom morfologickou složitostí češtině podobná. V anglickém korpusu lze dohledat celkové množství různých morfémů (tvarů slov) 97 734. V tureckém je to potom 474 957. Pokud se ale morfémy převedou do svého základního tvaru, vychází v obou těchto korpusech počet morfémů na hodnotu kolem 90 000. Průměrně na jeden základní tvar slova v turečtině vychází 5 různých morfémů (Demir, 2014).

Při trénování NER klasifikátoru operujícím na morfologicky bohatých jazycích, do kterých čeština (společně třeba se zvýše zmiňovanou turečtinou) patří, tak nabývá na důležitosti struktura trénovacích dat.

Ta totiž může ke každému jednotlivému slovu v datasetu nést i dodatečné informace, které mohou v konečném důsledku zpřesnit klasifikaci jmenných entit právě v případě výskytu většího množství formátů daného slova (Konkol a Konopík, 2014).

### Stematizace a lematizace

Jednou z užitečných dodatečných informací, kterou mohou jednotlivá slova v datasetu nést je tzv. lemma. Toto označení popisuje slovo ve svém základním slovníkovém tvaru, tedy například lemmou výrazů *jsi, jsem* nebo *byli jste* je slovo *být*. Lematizací se potom nazývá proces vytváření lemm z původních morfémů.

Jako stematizace se označuje algoritmy jednodušší proces převádění morfémů do základních tvarů, avšak bez dodatečné morfologické analýzy. Typicky se tedy jedná o odstraňování přípon a předpon slov a následné převádění do slovníkové podoby

Lematizace i stematizace jsou složitými operacemi, které se řeší například Porterovým algoritmem (použitelným pro angličtinu), rule-based algoritmy nebo různými statistickými modely. Jako příklad state-of-the-art nástroje využitelného pro lematizaci v češtině lze uvést nástroj MorphoDiTa (Konkol a Konopík, 2014).

### POS

Part-of-speech tag, neboli POS označuje, je-li v datasetu dostupný, slovní druh každého morfému. I v tomto případě je většina dnešních taggerů (modelů umožňujících určování POS) implementována pomocí modelů strojového učení. Největší výzvu v tomto případě představuje fakt, že různá slova ve stejném tvaru mohou být v závislosti na kontextu rozdílným slovním druhem (například slovo *jez* může být v různém kontextu podstatným jménem nebo slovesem).

Ve světě NLP se typicky pracuje s POS tagy ve stejném, standardizovaném formátu, definovaném rámcem Universal Dependencies. Tento rámec se snaží celosvětově sjednotit značení morfologických a sémantických příznaků v anotovaných datasetech.

* Vlastní model budu trénovat (oproti tomu kdybych využil třeba již hotový NameTag) proto, abych mohl architekturu udělat modulární a připravit jí pro nové klasifikátory pro jiné využití.
* Stav poznání v Česku
  + Zmínit že v cs je NER opravdu na úrovni
  + Zaměření na osobní údaje neexistuje, vše je obecné
  + Modely nejsou dobře rozšiřitelné, jejich upravení pro specifickou doménu a použití není snadné
  + Část před transformery
  + Část po transformerech

## Multilingvní modely

Využití multilingvních modelů, přesněji řečeno pak cross-lingual word embeddings je částečným řešením problému s nedostatečně velkými a kvalitními datasety v cílovém jazyce.

Typický přístup k trénování NLP modelu spoléhá na anotovaná data pro danou úlohu, kterou má výsledný model vykonávat. To vede k technice *supervised learning* (viz nadpis 5.3 Named Entity Recognition). Trénovací anotovaná data jsou potom typicky poskytována v jednom jazyce (Conneau et al., 2018).

V případě, kdy má model fungovat pro více jazyků současně (což je případ mezinárodně dostupných produktů jako například webových vyhledávačů) je ale potřeba, aby byl model natrénován zároveň pro více jazyků najednou. Vzhledem k tomu, že rozdílné jazyky, a především pak skupiny jazyků se vyznačují rozdílnou vnitřní logikou a sémantickými pravidly, mělo by při správném trénování tohoto modelu docházet k vytváření jazykově nezávislých příznaků (language-independent features) (Tsai et al., 2016).

Stejný přístup se dá využít i při trénování NLP modelu v jazyce, v kterém není dostupné větší množství anotovaných dat pro danou NLP úlohu (např. NER). Váhy multilingvního modelu potom mohou být využity v počátku trénování monolingvního jazykově specifického modelu a tím omezit celkově potřebné množství dat potřebných pro úspěšné natrénování modelu (Tsai et al., 2016).

Kapitolou samou o sobě je potom trénování multilingvních word embeddings. Tato běžně využívaná část modelu dokáže zachytit logické vztahy mezi jednotlivými slovy trénovacího datasetu. Velkou výhodou je, že word embeddings se dá trénovat pomocí techniky *unsupervised learning*, tedy učením bez učitele. Tím vznikl prostor pro trénink word embeddings na obrovských neanotovaných korpusech, jako jsou například kompletní záznamy z wikipedie ve všech na wikipedii dostupných jazycích (Conneau et al., 2018).

Dále existuje rozdíl mezi kontextově závislými a nezávislými word embeddings. Modely jako Word2Vec nebo GloVe totiž vytvářejí kontextově nezávislou reprezentaci mezi jednotlivými zpracovávanými slovy. Modely typu BERT, ELmo či podobné dokážou pracovat i s kontextem textu na úrovni vět a kontinuálních textových částí (Miaschi et al., 2020). Například pokud bude některé slovo znít stejně, ale bude mít více významů (tedy bude homonymem), ve word embeddings vygenerovaném modelem Word2Vec bude toto slovo stejnou vektorovou reprezantaci. Ve word embedding vygenerovaném modelem BERT však může mít stejné slovo v závislosti na významu takových reprezentací víc.

Z hlediska úlohy NER, ve které je širší kontext textu klíčový se tedy tato práce dále zaměřuje na modely, které pracují s kontextově závislými word embeddings.

### mBERT

Poslední verze (6. 5. 2022) modelu mBERT, neboli multilingual BERT (Bidirectional Encoder Representations from Transformers) podporuje 104 jazyků (mezi nimi i češtinu), skládá se z 12 vrstev transformerů se společným počtem 768 skrytých vrstev.

Tento model byl vytrénován na korpusu obsahujícím kompletní texty z internetové encyklopedie Wikipedie (Bert Github, 2022).

Tento model (převážně tedy jeho ne-multilingvní varianta BERT) je považován jako současný state-of-the-art ve většině NLP úloh. BERT využívá ve své architektuře principu MLM, neboli Masked Language Modeling, kdy záměrně maskuje některá vstupy tak, aby model dokázal zvyšovat svojí schopnost abstrakce a vyhnul se tak jeho přeučení – viz obrázek 3: Zpracování vstupu pomocí modelu mBERT (Devlin et al., 2019).

TODO dopsat o problémech s inbalancí datasetu + doplnit obrázek <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>

A screenshot of a computer

Description automatically generated with medium confidence

Obrázek 3: Zpracování vstupu pomocí modelu mBERT (Devlin et al., 2019)

### XLM

Model XLM byl již vyvinut jako cross-lingual, tedy s cílem objevovat jazykově nezávislé příznaky. Mimo principu MLM ve své architektuře využívá i dalšího principu označeného jako Translation Language Modeling (TLM) – viz obrázek 4:Zpracování vstupu modelem XML. To ve své podstatě znamená, že model může při predikování využít dostupný další nezamaskovaný token, který je ovšem dostupný pouze v jiném jazyce. Tím vzniká větší provázanost mezi těmito jazyky (Lample a Conneau, 2019).

Tento model byl trénován na stejném datasetu jako předchozí model mBERT.

Nevýhoda tohoto modelu spočívá v tom, že jeho trénink není možný kompletně bez učitele. Vyžaduje totiž mít v datasetu označenou závislost mezi paralelními příklady tak, aby dokázal pomocí TLM zaměňovat jednotlivé tokeny.

Graphical user interface, application, table, Excel

Description automatically generated

Obrázek 4: Zpracování vstupu modelem XML (Lample a Conneau, 2019)

### XML-RoBERTa

Model XML-RoBERTa dělá oproti XML krok zpátky a zavrhuje TLM. Dělá to kvůli podpoře kompletního učení bez učitele.

V podstatě se tedy jedná o podobný princip jako u modelu mBERT, ovšem model dle svého jména používá architekturu RoBERTa. Ta funguje na stejném principu jako BERT, ovšem je řádově košatější. Dle původní práce *RoBERTa: A Robustly Optimized BERT Pretraining Approach* totiž původní model BERT podléhá underfittingu. RoBERTa tedy tento model rozšiřuje a trénuje na řádově větším datasetu. Zároveň však zvětšuje jeho velikost a zvyšuje výpočetní náročnost pro jak trénování modelu, tak i následnou inferenci. Celková přesnost modelu se však dá označit oproti modelu BERT jako vyšší (Liu et al., 2019).

Model XML-RoBERTa je tak po vzoru modelu RoBERTa vytrénován na datasetu CommonCrawl, který obsahuje až 250 000 unikátních tokenů (slov) (Conneau et al., 2020).

TODO dodělat porovnání těchto modelů dle https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models

## Context-based modely

* Context based modely jsou vlastne deep learning NER modely…
* Nejprve by zde měla být část o tradiční detekci údajů bez hlubokého učení a tím i bez možnosti zpracování kontextu
* Je součástí NER – zaměřit se na modely, které dokáží udržet delší kontext, ne jen třeba 1 větu
* Naše zpracování kontextu je v případě naší aplikace spíš provázání příznaků dle mapy a na tom navázaná nutnost anonymizace, není to typické zachycování kontextu při NER

## Transfer learning

### Fine-tuning

* Napsat o závislosti na deep learningu – viz Deep Learning book, str. 526 Greedy Layer-Wise …

## Sekvenční modely

## Meta klasifikátory

# Výběr technologií

Vzhledem k tomu, že cílem této práce není vytváření nových architektur NLP a NER modelů, je třeba vybrat nejvhodnější již existující technologie pro vytváření modelů hlubokého učení, které se pro tyto úlohy nejvíce hodí.

Existuje mnoho frameworků, které se používají pro vývoj deep learning modelů. Jako příklad některých nejpopulárnějších lze uvést například TensorFlow, Keras, Caffe či PyTorch. Tyto frameworky umožňují relativně jednoduše vytvářet, trénovat a evaluovat jakékoliv modely založené na hlubokém učení bez vlastní implementace nízko úrovňové logiky těchto modelů. Tato implementace totiž může být velmi složitá a je ve velké míře závislá na vhodně zvolených optimalizačních technikách či upravení celého trénovacího workflow na konkrétní HW.

Společnosti, které tyto frameworky vyvíjí, se do nic snaží neustále pronášet nejmodernější poznatky z oblasti AI, a tím pádem lze obecně říci, že pokud není vyvíjený model velmi specifického charakteru, je lepší některý z těchto frameworků použít, protože to znamená využít všechny state-of-the-art principy, které jsou zrovna dostupné.

## NLP frameworky

Oblast NLP je v rámci hlubokého učení specifická typickým využívaním prvků, které se v ostatních oblastech (např. oblasti počítačového vidění) nevyužívají. V důsledku toho existují knihovny specificky zaměřené právě na NLP, které utilizují sílu předchozích zmiňovaných deep learning frameworků a přidávají ke své funkcionalitě i dodatečné funkce, které se zaměřují na NLP specifické části modelů, jako jsou například Word Embeddings či na sémantickou analýzu textu.

### Kritéria pro výběr NLP frameworku

Při výběru NLP frameworku použitého pro vývoj prototypu anonymizačního nástroje jsou autorem práce určena následující kritéria:

* Framework musí mít dostatečně kvalitní a obsáhlou dokumentaci.
* Framework musí být volně dostupný (a pokud možno open source) pro nekomerční využití.
* Framework musí podporovat programovací jazyk C++ nebo Python (a v ideálním případě, pokud je framework open source by měla být většina jeho částí i v těchto jazycích implementována).
* Framework by mělo být možné využít ve vývojovém prostředí Google Colab out-of-the-box, tzn. měl by být dostupný jako samostatná knihovna, kterou není nutné pro toto vývojové prostředí samostatně kompilovat.
* Framework musí obsahovat co nejrozsáhlejší podporu úlohy NER.
* Framework musí podporovat rozšířenou znakovou sadu UTF-16 či podobnou, aby v něm bylo možné natrénovat model založený na českém jazykovém korpusu.

### Stanford CoreNLP (Stanford NER)

Knihovna Stanford CoreNLP obsahuje řadu nástrojů pro různé úlohy pojící se s NLP. Pro oblast NER, která je pro tuto práci nejvíce relevantní pak Stanford CoreNLP nabízí úžeji zaměřenou samostatně využitelnou knihovnu Stanford NER (která je jinak součástí větší CoreNLP knihovny).

Samotná knihovna Stanford NER je implementována v jazyce Java, nativně tedy podporuje tento programovací jazyk. Existuje však řada implementací nadřazených programových rozhraní pro řadu dalších programovacích jazyků, včetně Pythonu, pro který existuje samostatná knihovna Stanza, která je vyvíjená týmem zodpovědným i za Stanford NER a která využívá v pozadí právě knihovny Stanford NER.

Stanford NER podporuje již v základu fine-grained hierarchickou klasifikaci entit (viz 7.1.1 Granularita NER datasetů).

### NLTK

NLTK je NLP knihovna využitelná s programovacím jazykem Python (je součástí i balíčkovacího systému pip). Oproti Stanford CoreNLP je tato knihovna snadněji využitelná, ale méně vhodná například pro vytváření modelů s experimentálními architekturami.

Chybí v ní podpora některých pokročilejších technik využitelných pro NER, jako je syntaktická analýza. U NLTK je ovšem dostupná kvalitní rozsáhlá dokumentace včetně mnoha příkladů konkrétního využití.

### Flair

Flair je jednou z novějších NLP knihoven. Je vyvinutá vývojovým týmem společnosti Zalando a v práci Sergeye Vychegzhanina a Evgenye Kotelnikova (2019), která porovnává přesnost klasifikace jmenných entit různých NLP knihoven při využití defaultních out-of-the-box modelů, dosahuje celkově nejkvalitnějších výsledků ze všech porovnávaných knihoven.

Flair je psaná v programovacím jazyce Python a důraz klade na možnost rozsáhlejších úprav vrstvy Word Embeddings. Zároveň poskytuje speciální podporu například pro zpracování biomedicínských datasetů.

Bohužel, dokumentace k této knihovně není moc rozsáhlá ani kvalitní a spíše připomíná ukázku využití jejich základního přiloženého modelu.

### SpaCy

Knihovna SpaCy cílí na zákazníky, kteří tuto knihovnu mohou využít jako end-to-end řešení pro implementaci komerčních projektů (nicméně knihovna samotná je distribuována jako open source). Nabízí tedy i další funkcionality potřebné pro nasazení NLP modelu do komerčního prostředí, jako je relativně jednoduché vytváření kompletních NLP pipelines či rozsáhlou knihovnu předtrénovaných modelů využitelných out-of-the-box v komerčních aplikacích. Dá se tak říct, že SpaCy je celý ekosystém v kterém je vše potřebné pro implementaci komplexního NLP řešení.

SpaCy nabízí programovací rozhraní v jazyce Python, přičemž knihovna samotná je pak implementována kombinací Pythonu a Cythonu, což je programovací jazyk s Python-like syntaxí, ovšem s větším důrazem na optimalizaci a memory management po vzoru programovacího jazyka C. Tím SpaCy dosahuje proti předchozím zmiňovaným knihovnám větší rychlosti (Vychegzhanin a Kotelnikov, 2019).

Dokumentace této knihovny je rozsáhlá a dostatečně kvalitní. Zároveň je zde kladen velký důraz na upravitelnost konkrétní implementace. Tato customizovatelnost je možná na různých úrovních znalostí programátora. Je tedy možné využít některé předdefinované komponenty, které je možné i upravit konkrétnímu use case a zároveň i vytvořit zcela nový model hlubokého učení dle vlastní experimentální architektury, který lze přímo vytvořit v závislé knihovně PyTorch.

Zajímavá je také integrace s databází hlubokých modelů společnosti HuggingFace. SpaCy tak umožňuje využít nejnovější state-of-the-art modely a upravit jejich funkčnost pro doménově specifické použití.

### Porovnání

Porovnání frameworků z pohledů přesnosti klasifikace jmenných entit je obtížné. Ač se některé práce touto problematikou zabývají (Vchegzhanin a Kotelnikov, 2019; Al Omran et al., 2017), výsledek těchto porovnání nemá velkou vypovídající hodnotu. Porovnávají se zde totiž defaultní modely poskytnuté knihovnami, které nutně nemusí být totožné (např. spaCy v defaultním stavu nevyužívá architekturu transformerů). Zároveň se v těchto pracech nepočítá s zpřesňujícími příznaky jako POS, LEMMAs či syntaktickou analýzou, a to z důvodu že ne všechna tato rozšíření vstupních dat nejsou všemi frameworky podporovány.

Obecně lze ale z těchto prací říct (ač zde platí zmiňovaná omezení), že framework spaCy je považován z pohledu klasifikace jmenných entit za nejrychlejší, flair za nejpřesnější a StanfordNLP za nejvíce používaný ve vědeckých kruzích.

Z pohledu této práce se jako nejvhodnější pro implementaci prototypu nástroje pro anonymizaci jeví framework spaCy, který se od ostatních odlišuje nejkvalitnější dokumentací a nejrozsáhlejšími možnostmi z hlediska úpravy kompletní klasifikační pipeline.

Na druhém místo lze dosadit framework flair, který je velmi nadějným adeptem pro další rozvoj a který podporuje řadu NER specifických funkčností (například kombinovatelné Word Embeddings), ale který je v tuto chvíli špatně dokumentovaný. Alternativou také může být framework stanza.

Někam hodit že jazyk implementace tedy bude python? Popsat HW a další věci stejně jako Ondra?

Presidio

* Podporované modely – spacy +?

## Vývojové prostředí

Jako vývojové prostředí bylo zvoleno Google Colab. Toto cloudové řešení je nástrojem pro vývoj aplikací pracujících se strojovým učením, který nabízí připojení na vzdálené servery s dostatečným výkonem jak pro trénink nových modelů hlubokého učení, tak pro jejich následné využití (inferenci).

Google Colab nabízí připojení k virtuálním strojům, které v závislosti na zvolené variantě poskytují buď pouze CPU, kombinaci CPU a GPU nebo dokonce i TPU (Tensor Processing Unit). Toto prostředí má v sobě již předinstalované některé nástroje (např. knihovnu TensorFlow) vhodné pro hluboké učení, takže je příprava pracovní instance o to jednodušší.

Samotné IDE je potom téměř totožné s populárním IDE Jupyter Notebook a nabízí tedy bohaté formátování společně s možností spouštět kód po jednotlivých buňkách. Prostředí lze propojit s Google osobním účtem a osobním úložištěm ve službě Google Drive, do kterého se dají při trénování modelů ukládat hotové prototypy.

Limitací tohoto prostředí je omezený čas dostupný pro neplatícího uživatele, pro který je přidělen výpočetní čas vzdáleného virtuálního stroje. Tento čas není nikde definován a je vypočítáván dle vytížení zdrojů a času připojení k virtuálním stroji (a to ne jednorázově, ale v delším časovém horizontu).

TODO napsat o různých druzích GPU dostupných v různých druzích předplatných

# Dataset

Výběr vhodného datasetu pro trénování a evaluaci vyvíjeného modelu je nejen nezbytnou, ale i klíčovou a velmi důležitou činností pro správné konečné fungování celého nástroje.

V kontextu této práce se jedná o dataset, který bude obsahovat anotované jmenné entity v českém jazyce a který tedy bude možné využít pro trénování modelu s učitelem. Distribuce kategorií jmenných entit by v nejlepším případě měla být v datasetu rovnoměrná, přičemž by dataset neměl obsahovat žádné chybně kategorizované entity (false positives).

Vzhledem k co největší snaze o modularitu vytvářeného nástroje je vhodné najít dataset s co nejširším záběrem definovaných jmenných entit. V tom případě se totiž vytvářené klasifikátory budou moct později lépe upravovat i bez dalšího přetrénování modelu pro upravení nástroje pro konkrétní aplikaci. Kategorie entit by také měly být svým charakterem co nejblíže kategoriím osobních údajů (viz kapitola 4 Kategorizace osobních údajů).

V neposlední řadě by vhodný dataset měl obsahovat i dodatečné informace o jednotlivých morfémech, jako dříve zmiňované lemmy či POS tagy.

## Anotace NER datasetů

### Granularita NER datasetů

V případě běžně využívaných a veřejně dostupných datasetů, používaných pro úlohu Named Entitity Recognition lze datasety rozdělit do 2 základních kategorií dle granularity klasifikace jmenných entit.

Datasety s nižší granularitou jmenných entit se označují jako coarse-grained. Tento způsob anotace se dá obecně označit jako více běžný, nese však s sebou typicky problém nedostatečné možnosti rozlišení jmenných entit ve více specifikovaných aplikacích. Zpracování výstupu modelu, který je natrénován na coarse-grained datasetu je však jednodušší jak na logiku, tak na výpočetní náročnost.

Datasety s vyšší granularitou jmenných entit se označují jako fine-grained, neboli také hierarchické datasety. Tento způsob nabízí efektivnější rozdělení kategorií do funkčních celků a zároveň nabízí možnost kaskádování dalších modelů pro každý supertyp. Označení supertyp potom označuje hlavní kategorii, která může mít pod sebou ve stromové struktuře další příbuzné atomické jmenné entity (např. supertyp OSOBA může mít pod sebou další kategorie jako je křestní jméno nebo příjmení). Problém v tomto případě nastává, pokud jeden druh jmenné entity patří do více kategorií zároveň (Mai et al., 2018).

### Způsoby anotace jmenných entit

Jmenné enity se běžně v datasetech anotují 2 způsoby. Prvním z nich je BIO (někdy označován taky jako IOB), což je zkratka pro Begin, In, Out. Tento způsob rozlišuje, zda je daný token jmennou entitou (pokud ne, je tokenu přiřazen tag O) a dále jestli je prvním tokenem jmenné entity (B) anebo pokračujícím tokenem (I).

Dalším způsobem je BILUO, což značí Begin, In, Last, Unit a Out. Tento způsob mimo předchozího značí i poslední token jmenné entity tagem I. Pokud se jmenná entita skládá pouze z jednoho tokenu, je jí potom přiřazen tag U. Dle práce Rationova a Rotha (2009) je tento způsob anotace vhodnější pro efektivnější trénování modelu, avšak novější práce jako je například článek *Named Entity Recognition Systém for Postpositional languages* od Kamrana a Mansoora (2016) tuto hypotézu vyvrací.

## Dostupné datasety

### Czech Named Entity Corpus (CNEC)

Asi nejznámější český dataset jmenných entit je v postupném vývoji již od roku 2007, kdy byl vytvořen jako součást práce *Named Entities in Czech: Annotating Data and Developing NE Tagger* (Ševčíková et al., 2007). Tento originální dataset je označován jako CNEC 1.0 a obsahuje 2000 manuálně anotovaných českých vět.

Od té doby došlo k rozšíření datasetu o další věty, POS tagy, lemmy i dodatečné morfologické informace. Poslední dostupná verze datasetu, označená jako CNEC 2.0 obsahuje 8993 manuálně anotovaných českých vět s celkovým množstvím 35220 jmenných entit (ÚFAL, nedatováno).

Tento dataset je proti ostatním českým NER datasetům specifický tím, že je hierarchický. Definované jmenné entity jsou v něm rozdělené do 8 supertypů, přičemž každý ze supertypů má pod sebou ještě další podkategorie. Celkově potom tento dataset obsahuje 46 kategorií atomických jmenných entit.

CNEC 2.0 je dostupný pod licencí Creative Commons BY-NC-SA 3.0, což znamená, že je možné ho volně použít pro jakékoliv nekomerční účely.

### BSNLP

BSNLP je dataset, který je periodicky aktualizován před výročním workshopem *Balto-Slavic Natural Language Processing*, který tradičně pořádá soutěž o nejlepší multilingvní NER model, který dokáže pracovat s pobaltskými slovanskými jazyky. Ty čítají bulharštinu, češtinu, polštinu, ruštinu, slovenštinu a ukrajinštinu. Poslední vydaná verze datasetu je BSNLP-2021.

BSNLP-2021 rozlišuje pouze 5 základních typů jmenných entit, dodatečně však obsahuje i lemmy (BSNLP Workshop 2021, nedatováno).

Dataset BSNLP-2021 je dostupný pod licencí Creative Commons BY-NC-SA 4.0.

### SumeCzech-NER

Dataset SumeCzech-NER je odvozený od velkého datasetu SumeCzech 1.0, který obsahuje až 1 milion českých novinových článků společně s jejich titulky a který je určen pro úlohu sumarizace textu.

SumeCzech-NER odkazuje na texty datasetu SumeCzech a přidává k nim jmenné entity, přičemž při tom rozlišuje 7 jejich kategorií (které se shodují se supertypy definovanými v CNEC 2.0 kromě typu Number expressions).

Anotace SumeCzech-NER není vytvořena manuálním způsobem, ale programaticky, což otevírá prostor pro více chybně definovaných či nenalezených jmenných entit v textu. Zároveň tento dataset neobsahuje žádné dodatečné informace jako jsou lemmy či POS tagy.

### Shrnutí

Z dostupných českých anotovaných datasetů, vhodných pro úlohu NER se pro využití z hlediska této práce jeví jako nejvhodnější dataset CNEC 2.0., který odpovídá všem předem stanoveným kritériím. Zároveň je pro tuto práci vhodnější než řada běžně používaných anglických datasetů, a to právě díky většímu množství anotovaných kategorií jmenných entit.

Dataset je také neustále udržován Ústavem formální a aplikované lingvistiky (ÚFAL), což zároveň do jisté míry i zaručuje jeho kvalitu, především pak tedy jeho přesnost a komplexitu co se týče nejednostranného doménově specifického zaměření.

## Zpracování datasetu

Vybraný dataset CNEC 2.0 je dostupný v několika formátech:

* **Html** – formát vhodný pro vizuálně efektivnější zobrazení anotovaných textů. Neobsahuje žádné dodatečné informace.
* **Plaintext** – formát, který obsahuje pouze anotované entity bez dodatečných informací. Tokeny, které nejsou jmennými entitami zde nejsou nijak značeny.
* **XML** – je anotován stejně jako plaintext, pouze v notaci XML.
* **Treex** – proprietární formát určený pro zpracování pomocí nástroje Treex, vyvíjeného ÚFAL. Obsahuje dodatečné informace.

Speciální varianta datasetu CNEC\_extended (pozor, dataset navzdory názvu o žádné další entity ani informace rozšířený není) je potom dostupná ve formátu **CONLL**, což je ve světě běžně používaný formát, který je spojený s výroční konferencí se stejným jménem. Tento formát obsahuje jak informaci o kategorii každého z tokenů, tak i dodatečné informace jako jsou lemmy. Problém je v tom, že CNEC\_extended rozlišuje pouze coarse-grained kategorie, tedy vlastně jednotlivé supertypy definované v CNEC 2.0.

Knihovna spaCy vyžaduje pro trénování modelu buď vlastní proprietární binární formát (.spacy) nebo dataset převedený do JSONu podle definovaných pravidel. Současně ve verzi knihovny spaCy v3.0 je podporovaný binární formát. JSON formát je potom spojený s verzí knihovny v2.0, nicméně je novější verzí knihovny stále podporován.

SpaCy nabízí nástroje pro konvertování běžně se vyskytujících formátů datasetů do jejich zpracovávatelného formátu. Mezi těmito formáty se vyskytuje CONLL-2003 i CONLL-U, který obsahuje rozšířené morfologické informace.

### Načtení CNEC\_extended

Pro další zpracování byl do .spacy binárního formátu načten pomocí nástroje spacy convert dataset CNEC\_extended (anotovaný způsobem BIO), který obsahuje menší množství kategorií jmenných entit (pouze supertypy).

### Načtení CNEC 2.0

TODO přidat sem i do zobrazení nějaký ukázky kódu

Kvůli potřebě většího počtu kategorií pro dosáhnutí větší modularity vyvíjeného nástroje byl dále dataset CNEC 2.0 (anotovaný způsobem BIO, obsahující hierarchické fine-grained kategorie) pomocí skriptu *treex2conll2003.pl*, dostupného jako součást práce *Neural Networks for Featureles Named Entity Recognition in Czech* (Straková et al., 2016) konvertován do formátu CONLL-2003.

Tento dataset ve formátu CONLL-2003 byl pomocí nástroje spacy convert s parametrem -c conll převeden do binární .spacy podoby, avšak bez dodatečných informací jako je lemma či POS. Spacy convert totiž tyto dodatečné informace při konvertování z CONLL-2003 neumí do svého formátu převést.

Pro načtení lemm je však možné postupně převést CNEC 2.0 dataset v CONLL-2003 formátu pomocí spacy convert s parametrem -t json do staršího spaCy JSON formátu a potom algoritmicky (pomocí funkce transform\_lemmas) opravit chybně konvertovaná data tak, aby byly v JSON struktuře lemmy dostupné.

Zároveň je při této operaci potřeba odstranit speciální úpravy na lemmách, které jsou specifické pro CNEC 2.0 dataset a které dodatečně morfologicky rozšiřují lemmu i například o různé vysvětlivky či reference (Prague Dependency Treebank, nedatováno). Tyto dodatečné informace sice mohou být pro trénování modelu potenciálně velmi cenné, avšak jsou jazykově specifické a knihovna spaCy není na zpracování těchto zvláštních morfologických atributů připravena.

Zpětným konvertováním pomocí spacy convert již v tomto případě lze dojít k použitelnému datasetu ve vyžadovaném binárním formátu .spacy s dodatečnou informací navíc ve formě potřebných lemm.

Při požadavku ke zpracování všech dostupných a pomocí knihovny spaCy zpracovatelných dodatečných informací, které dataset CNEC 2.0 nabízí, je potřeba načíst i POS tagy. Ty však nelze načíst při konverzi pomocí spacy nástrojů, pokud se zdrojový dataset nachází ve formátu CONLL-2003. Z tohoto důvodu byl dataset programaticky transformován do rozšířeného formátu CONLL-U. CNEC 2.0 poskytuje i POS tagy ve zvláštní rozšířené formě (stejně jako lemmy). Ta však není nijak standardizovaná (rámcem Universal Dependencies) a tedy ani dále zpracovatelná neproprietárními nástroji, vázanými přímo k CNEC 2.0 datasetu. Dalším krokem tedy došlo k namapování standartních POS tagů na CNEC 2.0 specifické tagy.

Bohužel, v posledním kroku, kterým mělo být převedení CNEC 2.0 v CONLL-U formátu pomocí spacy convert do .spacy binárního formátu, byla nalezena chyba ve zdrojovém kódu konverzního spacy nástroje, která zabraňuje správnému převedení BIO (i BILUO) anotací při využití zdrojového formátu CONLL-U.

Tuto chybu lze (k 8. 5. 2022) dohledat v spacy repositáři a to v souboru spacy/spacy/training/converters/conllu\_to\_docs.py (Spacy GitHub, 2022). Nachází se v něm konstanta MISC\_NER\_PATTERN, která v sobě nese regulární výraz, který by měl sloužit pro detekci jmenných entit, ale který funguje jen pro detekci tokenu O, tedy pro token, který nese informaci, že daný morfém není jmennou entitou.

TODO případně dopsat o konvertování a zpracování POS pokud se to nakonec nějak podaří

## Analýza dat

### Kategorie klasifikovaných jmenných entit

Diagram

Description automatically generated

Obrázek 5:Hierarchie jmenných entit CNEC 2.0

Vybraný dataset CNEC 2.0 rozlišuje 46 kategorií jmenných entit, které jsou hierarchicky přiřazené pod 8 supertypů (viz obrázek 5).

V této práci však bylo rozhodnuto o využití coarse-grained struktury, tedy struktury která je plochá a lze u modulárního nástroje lépe využít pro další přiřazení nadřazených recognizerů (viz kapitola 8 Vývoj modelu). Využití fine-grained struktury by přicházelo v úvahu při implementaci hierarchických modelů, které by bylo možné načítat dle nadřazených recognizerů, to ale není součástí této práce.

Při analýze datasetu bylo zjištěno, že některé jmenné entity jsou zastoupené s minimální četností, a tedy u trénování modelu pro jejich klasifikaci může docházet buď k přeučení modelu nebo naopak k vytváření chybné reprezentace. Tento problém konkrétně nastává u jmenné entity mi, tf nebo i\_, což bylo ověřeno i při evaluaci prvního natrénovaného modelu. Zároveň zde nastává problém s rozdělením datasetu na trénovací, testovací a evaluační část, kdy se absolutní hodnoty ještě více zmenšují. Konkrétní hodnoty absolutního a procentuálního zastoupení jmenné entity v datasetu lze vidět na obrázku 6.

### Úprava datasetu

Jmenná entita *mi* vyjadřuje identifikované URL adresy. Celkově se však v datasetu vyskytuje pouze ve 22 případech, což je velmi málo pro správné vytrénování modelu. Tato jmenná entita (URL) je pro detekci osobních údajů důležitou, proto ji nelze zcela odstranit. Při trénování NER modelu však není tento dataset z tohoto pohledu dostatečně kompletní, proto je entita *mi* z datasetu pro další zpracování vyřazena. Potřeba klasifikace URL z pohledu anonymizace osobních údajů je potom tento problém adresován přidáním dalšího nadřazeného recognizeru (viz kapitola 8).

Jmenná entita *pp*, která představuje označení mýtické bytosti je při dalším zpracování z datasetu odstraněna. Pro žádané využití modelu je tento údaj spíše irelevantní a mohl by způsobovat chyby při tréninku modelu.

Jmenné entity kategorie *tf* (entita představující názvy svátků a významných dnů) ani kategorie i\_ (která představuje instituce, které nejsou blíže specifikované) z datasetu odstraněny nejsou, a to kvůli možnému využití ve vyšších vrstvách nástroje. Také se u těchto entit předpokládá, že budou ve vyšších vrstvách sjednoceny s dalšími entitami pro detekci jednoho typu osobního údaje.

Seznam kategorií jmenných entit s kódovým označením jednotlivých kategorií, názvy kategorií, procentuálním zastoupením výskytu po úpravě datasetu a s doplňujícím vysvětlujícím popisem nebo příklady slov, které do dané kategorie patří je vyobrazen v tabulce 2.

Graf s procentuálním rozložením jmenných entit v upraveném datasetu lze potom vidět na obrázku 7. Z obrázku lze vidět, že rozložení stále není rovnoměrné. To je nicméně v případě datasetů určených k trénování NER modelů zcela běžná záležitost a pramení z podstaty nerovnoměrného výskytu určitých informací ve zdrojových textech.

Některé názvy kategorií jsou pojmenovány podle původní práce, ve které byl dataset navržen (Ševčíková et al., 2007). Další názvy a popisy jsou autorem upraveny nebo rozšířeny dle aktuálního stavu datasetu CNEC 2.0 v roce 2022.

Table

Description automatically generated

Obrázek 6:Zastoupení jmenných entit v datasetu CNEC 2.0

|  |  |  |  |
| --- | --- | --- | --- |
| Označení kategorie | Název kategorie | Procentuální zastoupení | Doplňující popis |
| tf | Svátky | 0.09 | Názvy různých svátků (Vánoce, Halloween,…) |
| i\_ | Nespecifikované instituce | 0.15 | Instituce nespecifikovaného konkrétního druhu |
| gh | Vodní útvar | 0.20 | Název jakéhokoli vodního útvaru (název řeky, jezera, moře,…) |
| ms | Stanice | 0.28 | Jméno TV nebo radio stanice |
| pm | Druhé jméno | 0.30 | Část jména osoby, která není křestním jménem či příjmením |
| gt | Kontinent | 0.31 |  |
| pd | Titul | 0.37 | Titul, navázaný na jméno osoby |
| g\_ | Nespecifikovaný geografický název | 0.39 |  |
| me | E-mail | 0.39 |  |
| gl | Přírodní oblasti/útvary | 0.40 | Např. Krkonoše, Sibiř,… |
| na | Věk | 0.40 |  |
| az | PSČ | 0.46 | Poštovní směrovací číslo |
| ns | Sportovní výsledek | 0.52 | Např. 2:0, vítězství Sparty,… |
| or | Předpis, norma | 0.53 | Např. sbírka zákonů, ISO 20000 |
| gq | Místní název | 0.58 | Např. část obce, místní pojmenování místa,… |
| o\_ | Nespecifikovaný název věci | 0.62 |  |
| ni | Položka seznamu | 0.66 |  |
| ah | Číslo popisné | 0.70 |  |
| gr | Menší územní celky | 0.77 | Např. kraje, oblasti,… |
| om | Měna | 0.79 |  |
| at | Telefonní číslo | 0.87 | V datasetu se vyskytují i čísla faxu |
| ia | Přednáška, konference, soutěž | 0.90 |  |
| mn | Periodikum | 0.91 | Může vyjadřovat i název redakce či mediální společnosti |
| pc | Obyvatelské jméno | 0.95 | Např. Afričan, Pražan,… |
| nb | Sekce | 1.03 | Např. kapitola, strana, vyjádření části celku |
| gs | Ulice | 1.14 | Názvy ulic a náměstí |
| no | Číslovky | 1.20 |  |
| p\_ | Nespecifikované jméno osoby | 1.20 |  |
| op | Výrobek, produkt | 1.53 |  |
| oe | Jednotka | 1.80 |  |
| td | Den | 1.95 |  |
| tm | Měsíc | 2.18 |  |
| n\_ | Nespecifikované číselné vyjádření | 2.26 |  |
| io | Státní nebo politická instituce | 2.95 |  |
| if | Firma | 3.20 |  |
| gc | Stát | 4.10 |  |
| ty | Rok | 4.90 |  |
| th | Čas | 5.43 |  |
| ic | Vzdělávací, kulturní nebo sportovní instituce | 5.58 |  |
| oa | Umělecké dílo | 6.72 | Kniha, obraz, film,… |
| gu | Město, hrad, zámek | 7.11 | Praha, Konopiště,… |
| nc | Číselné vyjádření | 7.65 |  |
| pf | Křestní jméno | 10.98 |  |
| ps | Příjmení | 14.95 |  |

Tabulka : Seznam kategorií jmenných entit upraveného datasetu CNEC 2.0

Chart, histogram

Description automatically generated

Obrázek : Graf rozložení výskytu jmenných entit v upraveném datasetu CNEC 2.0

Víceméně jde o kapitoly 2-4 v ML checklistu – projít.

* Zpracování datasetu
  + Seznam vytvořených datasetů
  + Popsat spacy nástroje na konvertování
  + Popsat spacy specifický formát ukládání datasetu a porovnat ho s předchozím JSONem
  + Zobrazit jak jsou jednotlivé kategorie v datasetu zastoupené, případně odstranit nepotřebné kategorie s málo příklady (mythical person?)
  + Zpracování lemmas, vynechání pos - odůvodnění
  + Zpracování outliners, minimálně jejich zhodnocení
  + Zjistit jestli a jak často se které NE opakují
  + Grafy
    - Viz <file:///C:/Users/ondra/Documents/V%C5%A0E/DP/Zdrojaky/others/Czech_Named_Entity_Corpus_2.0/cnec2.0/doc/techrep-ne-2007.pdf>
    - Kvantitativní vlastnosti – počty entit + procentuálně, počet délky instance jmenné entity
    - Kvalitativní vlastnosti – popsat způsob anotace 2 anotátory a porovnání výstupů
    - Rozložení entit v datasetu
    - Velikost – porovnání např. s anglickým conll
    - Zkusit objevit některé vztahy mezi jmennými entitami – viz <https://colab.research.google.com/drive/1yos8M0xIXpyOEHd9JmfEuY1I86SzZ_H8#scrollTo=vKGElU-hllwH>
  + Rozdělení na trénovací, validační a testovací dataset
    - K tomu hodit odkazy na nějaký současný stav poznání

TODO dodělat dataset smluv, který bude sloužit ke konečné evaluaci

# Vývoj prototypu

V následující kapitole bude postupně popsán proces návrhu a implementace prototypu anonymizačního nástroje. Tento proces se skládá z dvou hlavních částí.

V první části jde o vytrénování dostatečně kvalitního NER modelu založeného na technologii hlubokého učení, který bude sloužit jako hlavní komponenta anonymizačního nástroje.

V druhé části se tento model integruje do samotného anonymizačního nástroje. Tento nástroj bude umožňovat namapování doménově specifických jmenných entit relevantních pro danou aplikaci. Tímto způsobem je tento nástroj možné využít pro více druhů dokumentů i pro více kategorií jmenných entit. V případě této práce dojde k namapování na osobní údaje identifikované v kapitole 4 a vyobrazené v tabulce *1 Mapa kategorií osobních údajů*.

Pro dokonalejší funkci nástroje budou dále vyvinuty další samostatné klasifikátory, které budou dále utilizovat vyšší přesnost a kvalitu anonymizačního nástroje. Zároveň bude tímto ukázána modularita nástroje, která dokáže, že je možné funkčnost tohoto prototypu přenést na další druhy dokumentů a kategorií doménově specifických jmenných entit.

TODO tady na to navázat přesnější use casy a požadavky na nástroj i model?

Napsat vhodné praktiky, které směřují k implementaci nástroje tak, aby architektura byla dostatečně otevřená a výsledný nástroj tak byl co nejvíce modulární

## Trénování NER modelu

Pro trénování NER modelu bylo využito dříve zmiňovaného frameworku spaCy, který se dá nainstalovat přes Python balíčkovací systém pip včetně všech svých dependencies. To však platí pouze pro verzi knihovny, která utilizuje trénování pouze pro CPU. Vzhledem k tomu, že se dále v této práci využívá jak trénování jen pomocí CPU, tak ale i pomocí GPU, bylo potřeba doinstalovat i verzi spaCy s podporou CUDA a korespondující verzí knihovny PyTorch. V případě této práce byla vybrána verze CUDA 11.1 a k tomu kompatibilní verze PyTorch 1.10.1+cu111. Knihovna spaCy pak byla nainstalována s podporou modelů s architekturou transformer. Následně muselo dojít k exportu některých globálních proměnných pro další správné fungování frameworku.

!pip install torch==1.10.1+cu111 torchvision==0.11.2+cu111 torchaudio==0.10.1 -f https://download.pytorch.org/whl/torch\_stable.html

Zdrojový kód 8.1.1: Instalace CUDA + PyTorch

!pip install -U spacy[cuda111,transformers]

!export CUDA\_PATH=/usr/local/cuda-11.1

!export PATH=/usr/local/cuda-11.1/bin${PATH:+:${PATH}}

!export LD\_LIBRARY\_PATH=/usr/local/cuda-11.1/lib64${LD\_LIBRARY\_PATH:+:${LD\_LIBRARY\_PATH}}

!pip install cupy-cuda111

Zdrojový kód 8.1.2: Instalace spaCy pro GPU

K vytvoření funkčního kompletního spaCy modelu připraveného k použití je potřeba vytvořit spaCy pipeline. Pipeline si lze představit jako řadu navazujících komponent, které postupně zpracovávají vstupní data (tedy nestrukturovaný text) a generují různé výstupy. Tyto výstupy se pak dají použít jako vstupy pro další komponenty, které jsou součástí pipeline nebo jako součásti konečného výstupu celé pipeline. Tímto způsobem může spaCy sloužit jako jediná knihovna potřebná pro vytváření kompletních komerčních end-to-end řešení.

SpaCy nabízí řadu mnoho hotových již předtrénovaných pipelines, které lze jednoduše využít pro řadu jednodušších use cases. Například v angličtině spaCy nabízí pipeline en\_core\_web\_sm, která obsahuje komponenty typu tagger, parser, atrribute\_ruler, lemmatizer a ner a je vhodná pro většinu běžných NER úloh v anglickém jazyce. Pro stroje, které dokážou utilizovat GPU pak nabízí pipeline en\_core\_web\_trf, která má v sobě navíc komponentu transformer a dokáže využívat modelů založených na architektuře transformer.

Žádnou předtrénovanou pipeline pro český jazyk spaCy nenabízí, bylo proto potřeba ji vhodně navrhnout a natrénovat od začátku na datasetu zpracovaném v předchozí kapitole.

Návrh pipeline lze provést ve verzi spaCy v3 pomocí speciálního konfiguračního souboru config.cfg, který si spaCy knihovna načítá před spuštěním trénovacího skriptu. V tomto souboru je nutné nadefinovat všechny klíčové parametry potřebné pro trénování modelu. To platí jak u každé jednotlivé komponenty modelu tak i u globálních parametrů (těmi je například cesta k trénovacímu a testovacímu datasetu nebo batch size). Struktura konfiguračního souboru je tvořena řadou sekcí a podsekcí, které dohromady vytvářejí konkrétní architekturu všech jednotlivých částí pipeline.

Na obrázku 8 lze vidět zvolenou architekturu spaCy pipeline zvolenou pro trénování modelů utilizujících pouze CPU. Na obrázku 9 pak architekturu spaCy pipelinou modelů utilizujících GPU. Tyto pipelines se liší v komponentě tok2vec a transformer. Tato představuje vrstvu word embeddings, přičemž v GPU modelu vrstva transformer využívá architektury modelu transformer a kontextuálního word embeddings (tedy fakticky spíše jen embeddings).



Obrázek : CPU spaCy pipeline



Obrázek : GPU spaCy pipeline

Vytvořené modely, trénované na datasetu bez dostupných lem pak ve své architektuře neobsahují komponentu lemmatizer (čímž se značně snižuje čas potřebný pro jejich vytrénování).

Konkrétní konfigurační soubory v plné délce, použité pro trénování jednotlivých modelů jsou dostupné v příloze jako součást zdrojového kódu. (TODO definovat v jaké části přílohy) V následující části tak budou popsány jen některé jejich části. Každý konfigurační soubor pak obsahuje řadu dalších zajímavých vlastností pipeline a jednotlivých komponent, které dále nejsou popsány, a to z důvodu zjednodušení přehlednosti a srozumitelnosti následujícího popisu.

V následujícím zdrojovém kódu (Zdrojový kód 8.1.3) lze vidět definované absolutní cesty ke zdrojovému datasetu. Také je zde vidět, že nebyly použité žádné předtrénované váhy pro komponentu tok2vec.

[paths]

train = "/content/drive/MyDrive/DP/datasets/fine\_grained\_conll/spacy\_lemmas/train\_transformed.spacy"

dev = "/content/drive/MyDrive/DP/datasets/fine\_grained\_conll/spacy\_lemmas/dtest\_transformed.spacy"

vectors = null

init\_tok2vec = null

Zdrojový kód 8.1.3: Cesty k externím vstupním datům

Následující zdrojový kód obsahuje specifikaci, že trénink modelu bude probíhat s využitím GPU a knihovny PyTorch (Zdrojový kód 8.1.4)

[system]

gpu\_allocator = "pytorch"

seed = 0

Zdrojový kód 8.1.4: Definice podřízené ML knihovny

Dále je potřeba specifikovat, že dochází k trénování pipeline v českém jazyce. Tím pak spaCy může vytvářet slovník pouze s českými výrazy. Zároveň je zde definováno pořadí jednotlivých komponent a batch size (počet vzorků využitých pro trénink jedné iterace modelu), která je v případě následujícího konfiguračního souboru určeného pro trénink modelu na CPU stanovena na 1000 vzorků (Zdrojový kód 8.1.5).

[nlp]

lang = "cs"

pipeline = ["tok2vec","trainable\_lemmatizer","ner"]

batch\_size = 1000

…

Zdrojový kód 8.1.5: Ukázka některých NLP globálních nastavení

Vlastnosti každé komponenty modelu je potom nutné dále detailněji specifikovat. Vše lze vidět na příkladu definice komponenty *ner*. V té je třeba definovat například objekt *scorer*, který je odpovědný za evaluaci dané iterace modelu (jde tedy přeneseně řečeno o objekt, který spravuje cost funkci).

Každé komponentě modelu je také třeba přiřadit architekturu podřazeného modelu. U *ner* je to například spacy.TransitionBasedParser.v2, tedy knihovnou spaCy předdefinovaná architektura, která se stará o správné propojení jednotlivých tokenů (které již mohou být předzpracovány předchozími komponentami) tak, aby došlo ke kontext citlivé klasifikaci jmenných entit. Je také nutné definovat velikost skryté vrstvy, která byla v případě této komponenty stanovena na 64 jednotek (neuronů) (Zdrojový kód 8.1.6).

[components.ner.model]

@architectures = "spacy.TransitionBasedParser.v2"

state\_type = "ner"

extra\_state\_tokens = false

hidden\_width = 64

…

Zdrojový kód 8.1.6: Ukázka přiřazení vlastností podřazeného modelu *ner*

V komponentě *ner*, konkrétně potom v její embedding vrstvě, která slouží ke zpracování vstupních vektorů je potřeba definovat s jakými příznaky bude tato vrstva pracovat. Jde tedy vlastně o vytvoření spojení s nižšími vrstvami (v případě této práce s vrstvou *lemmatizer*), které mohou předzpracovávat vstupní tokeny (jednotlivá slova vstupního textu). U každého tohoto příznaku se musí stanovit rozměr vstupního vektoru, který je typicky stejný jako rozměr výstupního vektoru z předřazené komponenty. Dá se zjednodušeně říct, že tímto nastavením se určují dostupné morfologické či syntaktické vlastnosti tokenu, které mají být brány v potaz při konečné klasifikaci jmenných entit komponentou *ner*. V následující ukázce je tedy vidět, že atribut „LEMMA“ je společně s atributem „NORM“, který označuje původní tvar tokenu (slova), příznakem při trénování komponenty *ner (*Zdrojový kód *8.1.7)*.

[components.ner.model.tok2vec.embed]

@architectures = "spacy.MultiHashEmbed.v2"

width = ${components.tok2vec.model.encode.width}

attrs = ["NORM","LEMMA","PREFIX","SUFFIX","SHAPE","SPACY"]

rows = [5000,5000,1000,2500,2500,50]

include\_static\_vectors = false

Zdrojový kód 8.1.7: Nastavení *embedding* vrstvy komponenty *ner*

Další zajímavou částí je popis samotného trénovacího procesu. V této části se dá pracovat například s hodnotou dropout rate, která zabraňuje přeučení modelu nebo s hyperparametrem learn\_rate, který definuje velikost jednotlivých kroků při optimalizaci cost funkce. Dá se lze také nastavit optimalizační fukce (v případě této práce byl použitý optimizer Adam) (Zdrojový kód 8.1.8).

[training]

accumulate\_gradient = 3

dropout = 0.1

patience = 1600

max\_steps = 20000

eval\_frequency = 200

…

[training.optimizer]

@optimizers = "Adam.v1"

beta1 = 0.9

beta2 = 0.999

L2\_is\_weight\_decay = true

L2 = 0.01

grad\_clip = 1.0

use\_averages = false

eps = 0.00000001

[training.optimizer.learn\_rate]

@schedules = "warmup\_linear.v1"

warmup\_steps = 250

total\_steps = 20000

initial\_rate = 0.00005

Zdrojový kód 8.1.8: Ukázka nastavení některých trénovacích parametrů aplikovaných pro celou pipeline

V neposlední řadě je zde potřeba zmínit načtení již předtrénovaného multilingvního modelu založeného na transformer architektuře v případě trénování pipeline utilizující GPU. V následující ukázce dojde k načtení modelu „bert-base-multilingual-uncased“. Tento model, více popsaný v části 5.5.1 mBert, je předtrénovaný na vícejazyčných textech z Wikipedie. Knihovna spaCy strukturu i váhy tohoto modelu stahuje z dříve zmiňované databáze modelů HuggingFace. Tato vrstva v případě této práce není uzamčená, a tedy při trénování pipeline dochází i k dalšímu trénování tohoto modelu na vstupním českém datasetu (Zdrojový kód 8.1.9).

[components.transformer.model]

@architectures = "spacy-transformers.TransformerModel.v3"

name = "bert-base-multilingual-uncased"

mixed\_precision = false

Zdrojový kód 8.1.9: Definice předtrénovaného modelu mBert v komponentě *transformer*

Po doinstalování dodatečného balíčku spacy-lookups-data je možné využít některé již vytvořené lookup tabulky, které se mohou týkat lematizace slov, ale i jiných fenoménů. V případě této práce je načtena i lookup tabulka pro normalizaci, tzn. tabulka, která dokáže sjednotit různé tvary jednoho slova se stejným významem (tedy např. gymnázium a gymnasium) (Zdrojový kód 8.1.10).

[initialize.lookups]

@misc = "spacy.LookupsDataLoader.v1"

lang = "cs"

tables = ["lexeme\_norm","lemma\_lookup"]

Zdrojový kód 8.1.10: Načtení externích lookup tabulek

V první fázi trénování modelů byly vytrénovány modely utilizující pouze CPU a tedy modely, které nejsou založené na architektuře transformer. Ačkoli ve finální verzi anonymizačního nástroje bude použitý model utilizující GPU, na těchto modelech, méně výpočetně náročných pro trénování bylo provedeno vyhodnocení připravených datasetů.

Natrénován byl tedy CPU model na coarse-grained datasetu derivovaném z CNEC\_extended, model nevyužívající dodatečné morfologické informace lemma a model, který informace o lemmách využívá. Tyto 2 modely byly přitom vytrénovány na transformované vyčištěné verzi datasetu CNEC2.0.

Jako způsob pro evaluaci modelu vybráno kritérium F1 score, které vyjadřuje vztah mezi dalšími základními ML evaluačními kritérii Precision a Recall. Toto kritérium je při evaluování modelů velmi populární a dobře vyjadřuje obecnou kvalitu modelu, a to s menší závislostí na imbalanci datasetu než třeba základní kritérium Accuracy.

TODO doplnit podrobnější popis precision a recall?

Bylo vybráno 5 předtrénovaných modelů založených na archiktektuře transformer. Každý z těchto předtrénovaných modelů byl vytrénován na jiném textovém korpusu.

Vybrané modely **GPU\_bert\_uncased** a **GPU\_bert\_cased** používají známý model mBERT vytrénovaný na textech Wikipedie a jsou tedy multilingvními modely. Varianta modelu GPU\_bert\_cased je oproti GPU\_bert\_uncased předtrénována na datasetu obsahujícím velká i malá písmena. To je výhodné například z hlediska lepší detekce různých jmen.

Model **GPU\_small\_e\_czech** používá v pozadí model Small-E-Czech, který je architekturou modelem ELECTRA a je předtrénovaný na datasetu Czech web corpus, který je spravován společností Seznam.cz. Tento model i dataset jsou poměrně nové, článek o nich byl publikován v prosinci 2021 (Kocián et al., 2021).

Model **GPU\_robeczech** je předtrénovaný na textech z různých českých periodik a článků o celkové velikosti až 450 milionů tokenů. Svou architekturou se jedná o model RoBERTa. Tvůrci tohoto modelu tvrdí, že svojí kvalitou při použití pro české texty překonává známé neanotované multilingvní i další české datasety (Straka et al., 2021).

Model **GPU\_czert\_b\_based** je založený na modelu Czert-B-base-cased. Ten je vytvořený na kombinaci modelů BERT a ALBERT a natrénovaný je na menším, ale zato kvalitním datasetu, který obsahuje 340 000 českých vět. Dle autorů by měl ve většině NLP úloh předčít většinu multilingvních modelů (Sido et al., 2021).

V konfiguračních souborech všech trénovaných pipelines byla stanovena pevná hranice pro celkový maximální počet kroků (max\_steps) 20 000 a posuvná hranice pro zastavení tréninku modelu v případě absence dalšího zlepšování (patience) 1600 kroků. Maximální počet epoch nebyl definován.

Pojmem jeden krok (step) se v tomto případě označuje 1 kompletní průchod neuronovou sítí pro danou várku (batch) z trénovacího datasetu. Pojem epocha potom označuje 1 kompletní průchod neuronovou sítí pro všechny vzorky z trénovacího datasetu.

V prostředí Google Colab nelze dopředu vyčíst, jak dlouho lze využít připojení k jednomu virtuálnímu stroji. Může se tak proto stát, že dojde v průběhu tréninku k odpojení. Aby se v tomto případě zabránilo ztrátě již vypočtených vah modelu, je model v intervalech automaticky průběžně ukládán na Google Disk. Zároveň byla maximální délka trénování jednoho modelu stanovena na 8 hodin. Dále prezentovaná evaluace kvality jednotlivých vytrénovaných modelů je tedy prováděna vždy na iteraci modelu, vykazujícího největší kvalitu při tréninku, přičemž platí následující pravidla:

* Došlo k dokončení tréninku modelu podle výše uvedených parametrů.
* Trénink modelu byl zastaven po 8 hodinách.

Evaluace každého modelu probíhá na speciálním evaluačním datasetu se vzorky nevyužitými pro samotné trénování modelu.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluace NER modelů – porovnání** | | | | |
| Model | lemmatizer accuracy | ner precision | ner recall | ner F1 score |
| CPU\_coarse |  | 64,43 | 62,21 | 63,30 |
| CPU\_fine\_nomorph |  | 64,42 | 63,56 | 63,99 |
| CPU\_fine\_lemmas | 91,49 | 65,70 | 61,40 | 63,48 |
| GPU\_bert\_uncased | 92,39 | 79,08 | 79,53 | 79,30 |
| **GPU\_bert\_cased** | **94,43** | **78,76** | **82,89** | **80,77** |
| GPU\_small\_e\_czech | 86,82 | 66,39 | 66,28 | 66,34 |
| GPU\_robeczech | - | - | - | - |
| GPU\_czert\_b\_based | 93,96 | 77,47 | 81,12 | 79,25 |

Tabulka : Evaluace NER modelů – porovnání

V tabulce 3 lze vidět výsledky evaluace jednotlivých modelů. Model GPU\_robeczech se nepodařilo správně vytrénovat, a tedy konkrétní hodnoty evaluace tohoto modelu v této tabulce chybí.

Z výsledků evaluace lze vyčíst, že nejlépe uspěly multilingvní typy modelů GPU\_bert\_uncased a GPU\_bert\_cased, oba založené na modelu mBERT. Z těchto dvou pak lépe vychází varianta GPU\_bert\_cased, což vzhledem k charakteru klasifikovaných kategorií jmenných entit, které jsou často názvy institucí či jmény, dává smysl. GPU\_bert\_uncased má oproti GPU\_bert\_cased vyšší hodnotu metriky precision a menší hodnotu metriky recall. To znamená, že GPU\_bert\_uncased dokáže klasifikovat detekovanou jmennou entitu s vyšší úspěšností správného rozeznání konkrétní entity. GPU\_bert\_cased má potom větší schopnost samotné detekce jmenné entity. Z hlediska využití v anonymizačním nástroji je pro tuto práci důležitější parametr recall, protože je spíše důležité danou entity detekovat než správně kategorizovat. Krom toho byla jako metrika určená pro porovnání kvality všech modelů zvolená F1 score, které má vyšší model GPU\_bert\_cased.

Dalším v pořadí je model GPU\_czert\_b\_based, který je následován modelem GPU\_small\_e\_czech. V tomto případě je tento výsledek taky spíše očekávaný. GPU\_small\_e\_czech totiž využívá architektury modelu ELEKTRA, který je spíše modelem lehčím a známým spíše pro jeho rychlost. Ta v kontextu této práce není brána v potaz, a proto není v tabulce 4 uvedena (jde o vývoj prototypu nástroje s co největší přesností detekce a klasifikace). Přesto z dalšího porovnání vychází, že model GPU\_small\_e\_czech je z vytrénovaných modelů nejrychlejší (metrikou je počet slov za vteřinu) a má nejlepší poměr ner F1 score a rychlosti.

O mnoho méně přesné jsou potom modely utilizované pro CPU. Jejich F1 score se pohybuje kolem hodnoty 63,5.

doplnit konkrétní výsledky + potom i s GPU modely vyplnit tabulku, dopsat model s kterým se dále pracuje ve streamlit části.

Popsat proč má hned na začátku lemmatizer větší přesnost – lexeme table

Diskutovat výsledky evaluace, podívat se jaké entity různé modely umějí dobře detekovat, apod.

## Návrh anonymizačního nástroje

Pro vývoj anonymizačního nástroje byl použit open source software development toolkit (SDK) Presidio. Ten nabízí v programovacím jazyce python nástroje na vytvoření aplikace sloužící k anonymizaci osobních údajů (TODO hodit sem odkaz na presidio repo).

Presidio nabízí 3 moduly: *analyzer*, *anonymizer* a *image redactor*.

*Analyzer* slouží k detekci osobních údajů ve vstupním textu. K této operaci přitom používá řadu definovaných python objektů, nazývajících se recognizery. Každý z těchto recognizerů dokáže v textu klasifikovat jednu či několik jmenných entit. Recognizery mohou fungovat na různých principech, jako je například detekce entit pomocí regulárních výrazů nebo pomocí slovníků. Presidio nabízí k využití buď několik již hotových recognizerů od jeho tvůrců (vhodných pro jazykově nezávislé entity typu IP adresa) nebo možnost vytvořit si recognizery vlastní, dle potřeby doménově specifického využití vyvíjené aplikace.

Klíčovým typem recognizeru je potom recognizer založený na hlubokém učení. Ten je potřeba vytvořit tak, aby mohl využívat v pozadí některého z podporovaných NLP frameworků. V případě této práce jde tedy o jakýsi wrapper na NER model vytrénovaný v předchozí části, který využívá *AnalyzerEngine* pro spouštění klasifikace každého vstupního tokenu knihovnou spaCy.

Dále je nutno zmínit, že Presidio nemá žádnou podporu českého jazyka. Ačkoli je tedy možné využít některé předdefinované jazykově nezávislé recognizery, je potřeba naimplementovat několik dalších jazykově závislých a aplikačně specifických recognizerů tak, aby byl splněn požadavek na anonymizaci důležitých osobních údajů běžně se vyskytujících ve smlouvách nahrávaných do Veřejného registru smluv.

*Anonymizer* slouží k anonymizaci osobních údajů, poskytnutých předchozím modelem *analyzer*. Nabízí přitom různé možnosti anonymizace, jako je prosté začernění údajů či pseudonymizace. V kontextu této práce je v navrženém nástroji použitá pouze základní funkčnost tohoto modulu, která nahrazuje ve vstupním textu konkrétní jmennou entitu jménem klasifikované kategorie.

*Image redactor* je v tuto chvíli spíše experimentální částí Presidia, která slouží k anonymizaci osobních údajů v obrázcích. V kontextu této práce není tento modul využit.

Před využitím Presidia je potřeba nejprve nainstalovat pomocí balíčkovacího systému pip moduly presidio\_analyzer a presidio\_anonymizer. Následně, aby SDK Presidio mohl správně referencovat vytrénovaný NER model, je potřeba z vytrénované pipeline vygenerovat pomocí nástroje spacy package python balíček, které je nutné dále znovu pomocí balíčkovacího systému pip nainstalovat (Zdrojový kód 8.2.1).

!pip install presidio\_analyzer

!pip install presidio\_anonymizer

!python -m spacy package /content/drive/MyDisk/PIIAnonymizer/models/CPU\_fine\_lemmas ./packages --name CPUFineLemmas

!pip install /content/packages/cs\_CPUFineLemmas-0.0.0

Zdrojový kód 8.2.1: Instalace Presidia a vytvoření balíčku z vytrénované NER pipeline

Pro integraci vytrénovaného NER modelu do SDK Presidio byl vytvořen wrapper *SpacyRecognizerCustom*, který zachytává výsledky klasifikace jmenných entit z podřazeného spaCy modelu a přiřazuje jednotlivým kategoriím správné konečné labely na základě analýzy typů osobních údajů na začátku této práce. Přiřazení fine-grained kategorií lze vidět na ukázce (Zdrojový kód 8.2.2). V případě využití tohoto nástroje pro jinou doménově specifickou aplikaci (tedy ne detekce osobních údajů ve smlouvách) lze tuto tabulku upravit tak, aby odpovídala konkrétním analyzovaným dokumentům. Tím je dosaženo jednoho z pilířů modularity tohoto nástroje.

CHECK\_LABEL\_GROUPS = [

    ({"PERSON"}, {"pd", "pf", "pm", "ps"}),

    ({"EMAIL\_ADDRESS"}, {"me"}),

    ({"LOGIN\_NICK"}, {"p\_"}),

    ({"iNSTITUTION"}, {"ia", "ic", "if", "io", "i\_"}),

    ({"PHONE\_NUM"}, {"at"}),

    ({"MEDIA\_NAME"}, {"mn", "ms"}),

({"NUMBER\_EXPR"}, {"nb", "nc", "ni", "no", "ns", "n\_"}),

    ({"DOMAIN"}, {"mi"}),

    ({"LOCATION"}, {"ah", "az", "gc", "gh", "gl", "gq", "gr", "gs", "gt", "gu", "g\_"}),

    ({"PRODUCT"}, {"op"}),

    ({"DATE\_TIME"}, {"td", "tf", "th", "tm", "ty"}),

    ({"OTHER"}, {"oa", "or", "o\_", "pc"})

]

Zdrojový kód 8.2.2: Ukázka možného nastavení konečné kategorizace jmenných entit v závislosti na vytrénovaném NER modelu

Společně s recognizerem *SpacyRecognizerCustom* byly vyvinuty jako ukázka možného dalšího modulárního rozšiřování nástroje i další doménově specifické recognizery. Tyto doménově specifické recognizery mohou sloužit i jen jako doplňující prvek anonymizačního nástroje, který zvyšuje konečnou přesnost klasifikace. Například vyvinutý recognizer *CSRCRecognizer*, který slouží pro klasifikaci rodného čísla ve své podstatě není pro aplikaci na klasifikaci osobních údajů ve smlouvách nutný. Rodná čísla totiž dokáže klasifikovat i zmiňovaný předchozí recognizer *SpacyRecognizerCustom* a to s labelem „NUMBER\_EXPR“. Pro zvýšení přesnosti a pro zvýšení rozlišovací schopnosti anonymizačního nástroje však byl tento recognizer jako součást prototypu naimplementován a dokazuje tím existenci druhého pilíře modularity tohoto nástroje. Tedy možnosti relativně jednoduše vytvářet doménově specifické recognizery pro konkrétní aplikaci anonymizačního nástroje.

*CSRCRecognizer* využívá definované délky a formy rodného čísla a využívá pro jeho spolehlivou detekci regulární výraz (Zdrojový kód 8.2.3).

PATTERNS = [

    Pattern("rodne cislo (high)", r"\d{2}(0[1-9]|1[0-2]|5[1-9]|6[0-2])(0[1-9]|1[0-9]|2[0-9]|3[0-1])\/?\d{3,4}", 0.5)

]

Zdrojový kód 8.2.3: Regulární výraz použitý pro detekci rodného čísla

V případě, ve kterém by se vyvíjený anonymizační nástroj stal komerčním produktem, by pravděpodobně došlo k dalšímu vývoji řady dalších doménově specifických recognizerů, které by si konečný uživatel (tedy někdo, kdo potřebuje uspokojit potřeby anonymizace osobních údajů) mohl volitelně zapnout. Zároveň by měl stejný koncový uživatel možnost vytvořit si vlastní pravidla a implementovat si tak zcela vlastní a zcela specifické recognizery.

Ve vyvíjeném nástroji byly dále použité některé předdefinované jazykově nezávislé recognizery, které nabízí SDK Presidio. Těmito recognizery jsou:

* CreditCardRecognizer - detekce čísel platebních karet
* CryptoRecognizer - detekce adres bitcoinových peněženek
* EmailRecognizer - detekce e-mailových adres
* IbanRecognizer – detekce čísla IBAN
* IpRecognizer – detekce IP adresy

Například využití předdefinovaného Presidio *EmailRecognizer*u je v případě anonymizačního nástroje vhodnější, a to proto, že ve zdrojovém datasetu, na kterém byl trénován NER model není výskyt jmenných entit vyjadřujících e-mailovou adresu dostatečný. Zároveň lze e-mailová adresa typicky dobře detekovat přes správně vytvořený regulární výraz.

V tabulce 4 lze vidět kategorie jmenných entit, které byly sumarizovány v části 4.4 Mapa kategorií doplněny o konkrétní jim přidělené recognizery a výsledné labely, kterými budou tyto údaje označovány v anonymizačním nástroji. Jak je z tabulky vidět, ke všem identifikovaným osobním údajům je přiřazen některý recognizer a výsledný label, s výjimkou identifikátoru 53 – pracovní zařazení. Tento typ jmenné entity by byl vhodný pro klasifikaci pomocí recognizeru pracujícího se strojovým učením, bohužel ale zdrojový dataset neobsahoval žádné anotace, které by tuto jmennou entitu popisovaly.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mapa kategorií osobních údajů – výsledná kategorizace** | | | | |
|  |
| **ID** | **Identifikátor** | **Typ identifikátoru** | **Přiřazený Presidio recognizer** | **Přiřazený výsledný label** |  |
| **Jmenné identifikátory** | | | | |  |
| 1 | Křestní jméno | sekundární | SpacyRecognizerCustom | PERSON |  |
| 2 | Příjmení | sekundární | SpacyRecognizerCustom | PERSON |  |
| 4 | Jméno + Příjmení | **primární** | SpacyRecognizerCustom | PERSON |  |
| 6 | E-mail s unikátním jménem | **primární** | EmailRecognizer | EMAIL\_ADDRESS |  |
| 7 | E-mail pracovní s obecným jménem | sekundární | EmailRecognizer | EMAIL\_ADDRESS |  |
| 9 | Přihlašovací údaje (login) | sekundární | SpacyRecognizerCustom | LOGIN\_NICK |  |
| 10 | Název společnosti | - | SpacyRecognizerCustom | INSTITUTION |  |
| **Číselné identifikátory** | | | | |  |
| 12 | Rodné číslo/číslo pojištěnce | **primární** | CSRCRecognizer | CS\_RC |  |
| 13 | Číslo OP | **primární** | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 14 | Číslo pasu | **primární** | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 15 | Číslo platební karty | **primární** | CreditCardRecognizer | CREDIT\_CARD |  |
| 16 | Číslo jiné karty | sekundární | CreditCardRecognizer | CREDIT\_CARD |  |
| 17 | Číslo bankovního účtu | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 19 | IBAN | sekundární | IbanRecognizer | IBAN\_CODE |  |
| 21 | Telefonní číslo | **primární** | SpacyRecognizerCustom | PHONE\_NUM |  |
| 23 | Číslo řidičského průkazu | **primární** | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 24 | Věk (sec) | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 25 | SIPO | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| **Znakové identifikátory** | | | | |  |
| 30 | SPZ | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 31 | MAC adresa | **primární** | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 32 | IP adresa | sekundární | IpRecognizer | IP\_ADDRESS |  |
| 33 | Doména | sekundární | SpacyRecognizerCustom | DOMAIN |  |
| 38 | VIN | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 40 | Sériové číslo | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 42 | Daňové identifikační číslo (DIČ) | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| **Lokalizační identifikátory** | | | | |  |
| 45 | Adresa - bydliště | **primární** | SpacyRecognizerCustom | LOCATION |  |
| 46 | Adresa - sídlo PO nebo FOP | - | SpacyRecognizerCustom | LOCATION |  |
| **Ostatní identifikátory** | | | | |  |
| 49 | Náboženské vyznání | sekundární | SpacyRecognizerCustom | OTHER |  |
| 52 | Titul | sekundární | SpacyRecognizerCustom | PERSON |  |
| 53 | Pracovní zařazení | sekundární | - | - |  |
| 54 | Politická příslušnost | sekundární | SpacyRecognizerCustom | OTHER |  |
| 55 | Státní příslušnost | sekundární | SpacyRecognizerCustom | OTHER |  |
| 56 | Datum narození | sekundární | SpacyRecognizerCustom | NUMBER\_EXPR |  |
| 59 | Právnická osoba (PO)/Fyzická osoba podnikající (FOP) | - | Může být vyčteno z metadat |  |  |

Tabulka : Mapa kategorií osobních údajů - výsledná kategorizace

Aby byl vyvíjený anonymizační nástroj dobře prezentovatelný, byla společně s ním vyvinuta jednoduchá grafická webová aplikace, ve které je možné naživo sledovat proces anonymizace vstupního textu.

Pro vývoj této grafické webové aplikace byl použit open source framework Streamlit, ve kterém je možné psát v programovacím jazyce Python jednoduché grafické aplikace s použitím poskytovaných jednoduchých grafických komponent. K tomu, aby šel Streamlit framework použít v prostředí Google Colab, je potřeba nainstalovat správnou verzi 1.7.0, která používá shodnou verzi pythonu jako Google Colab (updatovat v Google Colab verzi pythonu není dobře možné, protože je potom vyžadován restart virtuálního stroje) (Zdrojový kód 8.2.4).

!pip install streamlit==1.7.0 pandas

Zdrojový kód 8.2.4: Instalace frameworku Streamlit do prostředí Google Colab

Následné spuštění Streamlit aplikace probíhá v prostředí Google Colab přes tunelování portu 8501 skrze utilitu localtunnel. Ta umožňuje jednoduše sdílet webovou aplikaci běžící na localhost na vzdálenou serverovou instanci, ke které je následně přiřazená URL adresa (Zdrojový kód 8.2.5).

!streamlit run /content/DP/src/streamlit\_app/presidio\_streamlit.py & npx localtunnel --port 8501

Zdrojový kód 8.2.5: Spuštění Streamlit aplikace v prostředí Google Colab

TODO doplnit konkrétními ukázkami kódu + trochu rozšířit a upravit a samotný skript

Samotná Streamlit aplikace má dostupné zdrojové kódy v příloze v části /src/streamlit\_app. Nejdříve dojde k importování dependencies. Mezi nimi jsou třeba předem nainstalované knihovny jako spaCy nebo Presidio, ale i běžnější python moduly jako pandas.

Následně dojde k načtení NER pipeline a definování všech výše zmiňovaných recognizerů, včetně vlastních doménově specifických.

Je nutné vytvořit instanci NlpEngineProvider, která slouží jako mezivrstva mezi spaCy NLP frameworkem a frameworkem Presidio. Stejně tak je potřeba i vytvořit instanci AnalyzerEngine, která si vezme jako parametr kolekci všech vytvořených instancí definovaných recognizerů.

V poslední části se definují zobrazované grafické prvky, jako je postranní panel (sidebar) nebo textová pole, která slouží pro zápis vstupního analyzovaného nestrukturovaného textu.

V přiloženém skriptu je ke spuštění na výběr z 2 možných variant výsledného anonymizačního nástroje. První z variant je určena pro spuštění na virtuálním stroji bez dostupné GPU. Tato varianta používá model CPU\_fine\_nomorph, a to především kvůli rychlosti tohoto modelu. Druhá varianta, vhodná pro spuštění pouze na GPU využívá nejlepšího z vytrénovaných modelů GPU\_bert\_cased.

* Obrázek toho jak to vypadá s popisem
* Popsat výběr modelu, kritéria pro výběr,…
* Svázat s ML checklistem
* Vytvořit algoritmus na spojování jmenných identifikátorů dle tabulky v části 1
* <https://spacy.io/usage/embeddings-transformers> - zajimavy treba pretraining – proc se nehodi při vyuziti transformeru
* Popsat propojeni s Hugging face knihovnou, jak se tyto modely transformuji do spacy
* Provest analyzu ruznych architektur modelu a vybrat která se hodi nejvíce (GPT2, BERT, Roberta, Electra,…)
* Napsat aspoň odstavec o kazdem recognizeru – jakym zpusobem k detekci dochazi, co recognizer dela (napr. PhoneRecognizer = <https://github.com/daviddrysdale/python-phonenumbers>
* Popsat presidio druhy dostupných recognizeru
* Popsat problém s cachovanim – vyresen odstanenim @st.cache z presidio\_streamlit.py
* Vytvořit obrázek architektury – jak jsou na sebe navázané jednotlivé modely, jak s nimi pracuje presidio,…
* Vyzdvihnout zajímavé technologie + vytvoření cs pipeline v spacy
* Přidat fotky GUI + popsat co to umí a jak se s ním dá pracovat

# Evaluace modelu

TODO

* Tady by mělo dojít i k porovnání výsledků vůči ostatním modelům, které nejsou založené na NLP a tudíž nemohou zpracovávat i kontext

Udelat evaulace všech variant modelu

* Coarse grind všechny typy
* Fine grind
  + Preddefinovane entity
  + Custom entity
* Nějaké info o ideální metrice měření
  + Vyžadujeme spíš vysoký recall než precision
  + Problém s imbalanced datasetem je běžný v NLP úlohách
* Li et al., 2020 – NER Evaluation Metrics

## Hodnocení naměřených kvalitativních indikátorů

## Automatická evaluace

## Evaluace člověkem

## Zhodnocení výsledků a porovnání

Závěr

Nápady na další iterace případného vývoje

* Doplnit mechanismus, který by dokázal vytvářet závislosti hlavně sekundárních identifikátorů
* pouzit i dalsi datasety pro trenink dalsich modelu, ktere by nasledne byly dodany do presidia jako dalsi analyzery

Použitá literatura

BERT GITHUB, bert/multilingual.md at master · google-research/bert. *GitHub* [online] [vid. 2022-01-09]. Dostupné z: <https://github.com/google-research/bert>

CHAU, Ethan C., Lucy H. LIN a Noah A. SMITH, 2020. Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. *arXiv:2009.14124 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/2009.14124>

CLARK, Kevin, Minh-Thang LUONG, Quoc V. LE a Christopher D. MANNING, 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/2003.10555>

DEVLIN, Jacob, Ming-Wei CHANG, Kenton LEE a Kristina TOUTANOVA, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/1810.04805>

EK, Ochrana údajů. *Evropská komise - European Commission* [online] [vid. 2022-01-09]. Dostupné z: <https://ec.europa.eu/info/law/law-topic/data-protection_en>

ELLMANN, Mathias, 2018. Natural language processing (NLP) applied on issue trackers. In: *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering* [online]. New York, NY, USA: Association for Computing Machinery, s. 38–41 [vid. 2022-01-09]. NL4SE 2018. ISBN 978-1-4503-6055-5. Dostupné z: doi:[10.1145/3283812.3283825](https://doi.org/10.1145/3283812.3283825)

GEALFOW, John Altair a Christian MAY, 2019. Anonymizace osobních údajů v soudních rozhodnutích. *Revue pro právo a technologie* [online]. **10**(19), 3–39. ISSN 1805-2797. Dostupné z: doi:[10.5817/RPT2019-1-1](https://doi.org/10.5817/RPT2019-1-1)

GRISHMAN, Ralph a Beth SUNDHEIM, 1996. Message Understanding Conference-6: a brief history. In: *Proceedings of the 16th conference on Computational linguistics - Volume 1* [online]. USA: Association for Computational Linguistics, s. 466–471 [vid. 2022-01-09]. COLING ’96. Dostupné z: doi:[10.3115/992628.992709](https://doi.org/10.3115/992628.992709)

GÉRON, Aurelien, 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1st vyd. B.m.: O’Reilly Media, Inc. ISBN 978-1-4919-6229-9.

KUBICA, Jan, 2019. *Vybrané problémy technologické realizace evropské ochrany osobních údajů* [online]. B.m. [vid. 2022-01-09]. Univerzita Karlova. Dostupné z: <http://invenio.nusl.cz/record/397848>

MENDELS, Omri, 2020. NLP approaches to data anonymization. *Medium* [online] [vid. 2022-01-09]. Dostupné z: <https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929>

MIHULKOVÁ, Jitka, 2018. *Co je, co není a co bude osobní údaj podle GDPR - Frank Bold Advokáti* [online] [vid. 2022-01-09]. Dostupné z: <https://www.fbadvokati.cz/cs/clanky/541-co-je-co-neni-a-co-bude-osobni-udaj-podle-gdpr>

MISHRA, Aditya, 2018. Metrics to Evaluate your Machine Learning Algorithm. *Medium* [online] [vid. 2022-01-09]. Dostupné z: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>

MOBERG, John, 2020. A deep dive into multilingual NLP models. *Peltarion* [online] [vid. 2022-01-09]. Dostupné z: <https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models>

MVČR. *Zvláštní kategorie osobních údajů - Ochrana osobních údajů* [online] [vid. 2022-01-09]. Dostupné z: <https://www.mvcr.cz/gdpr/clanek/zvlastni-kategorie-osobnich-udaju.aspx>

NADEAU, David a Satoshi SEKINE, 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* [online]. **30**(1), 3–26. ISSN 0378-4169, 1569-9927. Dostupné z: doi:[10.1075/li.30.1.03nad](https://doi.org/10.1075/li.30.1.03nad)

QU, Chen, Weize KONG, Liu YANG, Mingyang ZHANG, Michael BENDERSKY a Marc NAJORK, 2021. Natural Language Understanding with Privacy-Preserving BERT. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* [online]. New York, NY, USA: Association for Computing Machinery, s. 1488–1497 [vid. 2022-01-09]. ISBN 978-1-4503-8446-9. Dostupné z: <http://doi.org/10.1145/3459637.3482281>

SILVA, Paulo, Carolina GONÇALVES, Carolina GODINHO, Nuno ANTUNES a Marilia CURADO, 2020. Using natural language processing to detect privacy violations in online contracts. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing* [online]. New York, NY, USA: Association for Computing Machinery, s. 1305–1307 [vid. 2022-01-09]. ISBN 978-1-4503-6866-7. Dostupné z: <http://doi.org/10.1145/3341105.3375774>

WILLIAMS, Jake a Giovanni SANTIA, 2017. Context-Sensitive Recognition for Emerging and Rare Entities. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text* [online]. Copenhagen, Denmark: Association for Computational Linguistics, s. 172–176 [vid. 2022-01-09]. Dostupné z: doi:[10.18653/v1/W17-4423](https://doi.org/10.18653/v1/W17-4423)

XU, Hanchen, Zhenxiang CHEN, Shanshan WANG a Xiaoqing JIANG, 2021. Chinese NER Using ALBERT and Multi-word Information. In: *ACM Turing Award Celebration Conference - China ( ACM TURC 2021)* [online]. New York, NY, USA: Association for Computing Machinery, s. 141–145 [vid. 2022-01-09]. ISBN 978-1-4503-8567-1. Dostupné z: <http://doi.org/10.1145/3472634.3472667>

BOUDA, Petr, 2019. *Aktuálně k registru smluv | Právní prostor* [online] [vid. 2022-02-13]. Dostupné z: <https://www.pravniprostor.cz/clanky/spravni-pravo/aktualne-k-registru-smluv>

KRAUS, Martin, 2022. *Metodický návod k aplikaci ZRS*. 5. leden 2022. B.m.: MVČR.

ÚOOÚ, 2016. *Ochrana osobních údajů v souvislosti se zákonem o registru smluv: Názory a rozhodnutí Úřadu: Úřad pro ochranu osobních údajů* [online] [vid. 2022-02-14]. Dostupné z: <https://www.uoou.cz/ochrana-osobnich-udaju-v-souvislosti-se-zakonem-o-registru-smluv/d-20322/p1=1099>

WANG, Haohan a Bhiksha RAJ, 2017. On the Origin of Deep Learning. *arXiv:1702.07800 [cs, stat]* [online]. [vid. 2022-03-03]. Dostupné z: <http://arxiv.org/abs/1702.07800>

BENGIO, Yoshua a Olivier DELALLEAU, 2011. On the Expressive Power of Deep Architectures. In: Jyrki KIVINEN, Csaba SZEPESVÁRI, Esko UKKONEN a Thomas ZEUGMANN, ed. *Algorithmic Learning Theory* [online]. Berlin, Heidelberg: Springer, s. 18–36. Lecture Notes in Computer Science. ISBN 978-3-642-24412-4. Dostupné z: doi:[10.1007/978-3-642-24412-4\_3](https://doi.org/10.1007/978-3-642-24412-4_3)

ELDAN, Ronen a Ohad SHAMIR, 2015. The Power of Depth for Feedforward Neural Networks [online]. [vid. 2022-03-05]. Dostupné z: <https://arxiv.org/abs/1512.03965v4>

HINTON, Geoffrey E., Simon OSINDERO a Yee-Whye TEH, 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* [online]. **18**(7), 1527–1554. ISSN 0899-7667, 1530-888X. Dostupné z: doi:[10.1162/neco.2006.18.7.1527](https://doi.org/10.1162/neco.2006.18.7.1527)

JOHRI, Prashant, Sunil Kumar KHATRI, Ahmad AL-TAANI, Munish SABHARWAL, Shakhzod SUVANOV a Avneesh CHAUHAN, 2021. Natural Language Processing: History, Evolution, Application, and Future Work. In:  [online]. s. 365–375. ISBN 9789811597114. Dostupné z: doi:[10.1007/978-981-15-9712-1\_31](https://doi.org/10.1007/978-981-15-9712-1_31)

machinetranslate.org [b.r.]. Georges Artsrouni. *Machine Translate* [online] [vid. 2022-03-13]. Dostupné z: <https://machinetranslate.org/georges-artsrouni>

WOODS, W. A., 1970. Transition network grammars for natural language analysis. *Communications of the ACM* [online]. **13**(10), 591–606. ISSN 0001-0782. Dostupné z: doi:[10.1145/355598.362773](https://doi.org/10.1145/355598.362773)

NIGAM, vibhor, 2021. Natural Language Processing: From Basics, to using RNN and LSTM. *Analytics Vidhya* [online]. [vid. 2022-03-13]. Dostupné z: <https://medium.com/analytics-vidhya/natural-language-processing-from-basics-to-using-rnn-and-lstm-ef6779e4ae66>

GOLDBERG, Yoav, 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies* [online]. **10**(1), 1–309. ISSN 1947-4040. Dostupné z: doi:[10.2200/S00762ED1V01Y201703HLT037](https://doi.org/10.2200/S00762ED1V01Y201703HLT037)

LI, Jing, Aixin SUN, Jianglei HAN a Chenliang LI, 2020. A Survey on Deep Learning for Named Entity Recognition. *arXiv:1812.09449 [cs]* [online]. [vid. 2022-03-20]. Dostupné z: <http://arxiv.org/abs/1812.09449>

PETASIS, Georgios, Frantz VICHOT, Francis WOLINSKI, Georgios PALIOURAS, Vangelis KARKALETSIS a Constantine D. SPYROPOULOS, 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* [online]. USA: Association for Computational Linguistics, s. 426–433 [vid. 2022-03-20]. ACL ’01. Dostupné z: doi:[10.3115/1073012.1073067](https://doi.org/10.3115/1073012.1073067)

SETTLES, Burr, 2004. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)* [online]. Geneva, Switzerland: COLING, s. 107–110 [vid. 2022-03-20]. Dostupné z: <https://aclanthology.org/W04-1221>

GANESH, Prakhar, 2019. Distributed Vector Representation : Simplified. *Medium* [online] [vid. 2022-03-21]. Dostupné z: <https://towardsdatascience.com/distributed-vector-representation-simplified-55bd2965333e>

GIACAGLIA, Giuliano, 2021. Transformers. *Medium* [online] [vid. 2022-03-21]. Dostupné z: <https://towardsdatascience.com/transformers-141e32e69591>

DEMIR, Hakan a Arzucan OZGUR, 2014. Improving Named Entity Recognition for Morphologically Rich Languages Using Word Embeddings. In: *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014* [online]. Dostupné z: doi:[10.1109/ICMLA.2014.24](https://doi.org/10.1109/ICMLA.2014.24)

KONKOL, Michal a Miloslav KONOPÍK, 2014. Named Entity Recognition for Highly Inflectional Languages: Effects of Various Lemmatization and Stemming Approaches. In: Petr SOJKA, Aleš HORÁK, Ivan KOPEČEK a Karel PALA, ed. *Text, Speech and Dialogue* [online]. Cham: Springer International Publishing, Lecture Notes in Computer Science, s. 267–274 [vid. 2022-05-06]. ISBN 978-3-319-10815-5. Dostupné z: doi:[10.1007/978-3-319-10816-2\_33](https://doi.org/10.1007/978-3-319-10816-2_33)

YADAV, Vikas a Steven BETHARD, 2019. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *arXiv:1910.11470 [cs]* [online]. [vid. 2022-05-06]. Dostupné z: <http://arxiv.org/abs/1910.11470>

CONNEAU, Alexis, Guillaume LAMPLE, Ruty RINOTT, Adina WILLIAMS, Samuel R. BOWMAN, Holger SCHWENK a Veselin STOYANOV, 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv:1809.05053 [cs]* [online]. [vid. 2022-05-06]. Dostupné z: <http://arxiv.org/abs/1809.05053>

TSAI, Chen-Tse, Stephen MAYHEW a Dan ROTH, 2016. Cross-Lingual Named Entity Recognition via Wikification. In: *CoNLL 2016*: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* [online]. Berlin, Germany: Association for Computational Linguistics, s. 219–228 [vid. 2022-05-06]. Dostupné z: doi:[10.18653/v1/K16-1022](https://doi.org/10.18653/v1/K16-1022)

MIASCHI, Alessio a Felice DELL’ORLETTA, 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In: *Proceedings of the 5th Workshop on Representation Learning for NLP* [online]. Online: Association for Computational Linguistics, s. 110–119 [vid. 2022-05-06]. Dostupné z: doi:[10.18653/v1/2020.repl4nlp-1.15](https://doi.org/10.18653/v1/2020.repl4nlp-1.15)

BERT Github., 2022. *BERT* [online]. Python. B.m.: Google Research [vid. 2022-05-06]. Dostupné z: <https://github.com/google-research/bert/blob/eedf5716ce1268e56f0a50264a88cafad334ac61/multilingual.md>

LAMPLE, Guillaume a Alexis CONNEAU, 2019. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]* [online]. [vid. 2022-05-06]. Dostupné z: <http://arxiv.org/abs/1901.07291>

LIU, Yinhan, Myle OTT, Naman GOYAL, Jingfei DU, Mandar JOSHI, Danqi CHEN, Omer LEVY, Mike LEWIS, Luke ZETTLEMOYER a Veselin STOYANOV, 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]* [online]. [vid. 2022-05-06]. Dostupné z: <http://arxiv.org/abs/1907.11692>

CONNEAU, Alexis, Kartikay KHANDELWAL, Naman GOYAL, Vishrav CHAUDHARY, Guillaume WENZEK, Francisco GUZMÁN, Edouard GRAVE, Myle OTT, Luke ZETTLEMOYER a Veselin STOYANOV, 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]* [online]. [vid. 2022-05-06]. Dostupné z: <http://arxiv.org/abs/1911.02116>

ŠEVČÍKOVÁ, Magda, Zdeněk ŽABOKRTSKÝ a Oldřich KRŮZA, 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In: Václav MATOUŠEK a Pavel MAUTNER, ed. *Text, Speech and Dialogue* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, Lecture Notes in Computer Science, s. 188–195 [vid. 2022-05-07]. ISBN 978-3-540-74627-0. Dostupné z: doi:[10.1007/978-3-540-74628-7\_26](https://doi.org/10.1007/978-3-540-74628-7_26)

ÚFAL., [b.r.]. *Czech Named Entity Corpus 2.0 | ÚFAL* [online] [vid. 2022-05-07]. Dostupné z: <https://ufal.mff.cuni.cz/cnec/cnec2.0>

BSNLP Workshop 2021., [b.r.]. *BSNLP 2021 Workshop* [online] [vid. 2022-05-07]. Dostupné z: <http://bsnlp.cs.helsinki.fi/shared-task.html>

MAI, Khai, Thai-Hoang PHAM, Minh Trung NGUYEN, Tuan Duc NGUYEN, Danushka BOLLEGALA, Ryohei SASANO a Satoshi SEKINE, 2018. An Empirical Study on Fine-Grained Named Entity Recognition. In: *COLING 2018*: *Proceedings of the 27th International Conference on Computational Linguistics* [online]. Santa Fe, New Mexico, USA: Association for Computational Linguistics, s. 711–722 [vid. 2022-05-07]. Dostupné z: <https://aclanthology.org/C18-1060>

RATINOV, Lev a Dan ROTH, 2009. Design Challenges and Misconceptions in Named Entity Recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)* [online]. Boulder, Colorado: Association for Computational Linguistics, s. 147–155 [vid. 2022-05-07]. Dostupné z: <https://aclanthology.org/W09-1119>

KAMRAN, Muhammad a Syed MANSOOR, 2016. Named Entity Recognition System for Postpositional Languages: Urdu as a Case Study. *International Journal of Advanced Computer Science and Applications* [online]. **7**(10) [vid. 2022-05-07]. ISSN 21565570, 2158107X. Dostupné z: doi:[10.14569/IJACSA.2016.071019](https://doi.org/10.14569/IJACSA.2016.071019)

STRAKOVÁ, Jana, Milan STRAKA a Jan HAJIČ, 2016. Neural Networks for Featureless Named Entity Recognition in Czech. In: Petr SOJKA, Aleš HORÁK, Ivan KOPEČEK a Karel PALA, ed. *Text, Speech, and Dialogue* [online]. Cham: Springer International Publishing, Lecture Notes in Computer Science, s. 173–181 [vid. 2022-05-08]. ISBN 978-3-319-45509-9. Dostupné z: doi:[10.1007/978-3-319-45510-5\_20](https://doi.org/10.1007/978-3-319-45510-5_20)

PRAGUE DEPENDENCY TREEBANK., [b.r.]. *2.1. Lemma structure* [online] [vid. 2022-05-08]. Dostupné z: <https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s01.html>

SPACY GITHUB,HONNIBAL, Matthew, Ines MONTANI, Sofie VAN LANDEGHEM a Adriane BOYD, 2020. *spaCy: Industrial-strength Natural Language Processing in Python* [online]. Python [vid. 2022-05-08]. Dostupné z: doi:[10.5281/zenodo.1212303](https://doi.org/10.5281/zenodo.1212303)

VYCHEGZHANIN, Sergey a Evgeny KOTELNIKOV, 2019. Comparison of Named Entity Recognition Tools Applied to News Articles. In:  [online]. s. 72–77. Dostupné z: doi:[10.1109/ISPRAS47671.2019.00017](https://doi.org/10.1109/ISPRAS47671.2019.00017)

AL OMRAN, Fouad Nasser A a Christoph TREUDE, 2017. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. In: *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*: *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)* [online]. Buenos Aires, Argentina: IEEE, s. 187–197 [vid. 2022-06-02]. ISBN 978-1-5386-1544-7. Dostupné z: doi:[10.1109/MSR.2017.42](https://doi.org/10.1109/MSR.2017.42)

KOCIÁN, Matěj, Jakub NÁPLAVA, Daniel ŠTANCL a Vladimír KADLEC, 2021. *Siamese BERT-based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset* [online]. 3. prosinec 2021. B.m.: arXiv. [vid. 2022-06-05]. Dostupné z: <http://arxiv.org/abs/2112.01810>

STRAKA, Milan, Jakub NÁPLAVA, Jana STRAKOVÁ a David SAMUEL, 2021. RobeCzech Base [online]. [vid. 2022-06-05]. Dostupné z: <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3691>

SIDO, Jakub, Ondřej PRAŽÁK, Pavel PŘIBÁŇ, Jan PAŠEK, Michal SEJÁK a Miloslav KONOPÍK, 2021. *Czert -- Czech BERT-like Model for Language Representation* [online]. arXiv:2103.13031. B.m.: arXiv [vid. 2022-06-05]. Dostupné z: doi:[10.48550/arXiv.2103.13031](https://doi.org/10.48550/arXiv.2103.13031)

Poznámka: Za koncem Použité literatury nutné nesmazat *konec oddílu*, aby fungovalo různé číslování hlavního textu práce a příloh.

Přílohy

1. Název první přílohy
2. Název druhé přílohy