LICENČNÍ SMLOUVA

Níže uvedeného dne, měsíce a roku uzavřely smluvní strany

Západočeská univerzita v Plzni
Veřejná vysoká škola
Univerzitní 8, 30100 Plzeň
IČ 49777513
Zastoupená: **doc. Ing. Luděk Hynčík, Ph.D.**, prorektor pro výzkum a vývoj

(dále jen „**poskytovatel licence**")

a

Česká republika - Ústav pro studium totalitních režimů
Organizační složka státu
Siwiecova 2428/2, 130 00, Praha 3
IČ: 75112779
zastoupená: Mgr. Zdeněk Hazdra, Ph.D., ředitel

(dále jen „**nabyvatel licence**")

v souladu s ust.§ 2358 a násl. zákona č. 89/2012 Sb., občanský zákoník, v platném znění tuto licenční smlouvu.

Preambule

Poskytovatel prohlašuje, že je oprávněn tuto smlouvu uzavřít a uzavřením této smlouvy nebude neoprávněně zasaženo do práv jiných osob.

I.
Předmět smlouvy

1. Touto smlouvou **poskytovatel licence** uděluje **nabyvateli licence** oprávnění (licenci) k výkonu práva užít software s názvem „HIDOAR", vytvořeného v rámci projektu programu NAKI č. DG16P02B048, dále jen „**autorské dílo**", a to ke všem způsobům užití, v rozsahu neomezeném.

2. Bližší specifikace **autorského díla** je uvedena v příloze č. 1, která je nedílnou součástí této smlouvy.

3. **Nabyvatel licence** podepsáním této smlouvy potvrzuje převzetí **autorského** díla před podpisem této smlouvy.

II.
Způsoby užití díla

1. Licence se uděluje ke všem známým způsobům užití **autorského díla**.

2. Územní rozsah licence není omezen.

3. Časový rozsah licence není omezen.

4. Množstevní rozsah licence není omezen.

5. Nabyvatel licence se zavazuje užívat **autorské dílo** způsobem nesnižujícím jeho hodnotu.

6. Nabyvatel licence není povinen licenci využít.

7. Nabyvatel licence není oprávněn upravit či jinak měnit **autorské dílo**, jeho název nebo označení poskytovatele licence.

8. **Nabyvatel licence** není oprávněn užívat **autorské dílo** za účelem přímého nebo nepřímého hospodářského nebo obchodního prospěchu.

9. **Nabyvatel licence** má ošetřena práva přístupu k datům, která byla využita pro nastavení vnitřních parametrů **autorského díla** (data uvedena v příloze č. 1).

10. Poruší-li **nabyvatel licence** povinnost dle odstavce 7 a/nebo 8, je **nabyvatel licence** povinen zaplatit **poskytovateli licence** smluvní pokutu ve výši Kč 10.000,- za každý případ porušení.

11. Ujednáním o smluvní pokutě není dotčeno právo **poskytovatele licence** na náhradu škody.

III.
Nevýhradní licence

1. Licence podle této smlouvy se uděluje jako licence nevýhradní.

IV.
Podlicence, postoupení licence

1. **Nabyvatel licence** není oprávněn poskytnout podlicenci třetí osobě.

2. **Nabyvatel licence** nesmí licenci postoupit ani zcela ani zčásti třetí osobě.

V.
Utajení

1. **Nabyvatel licence** je povinen utajit před třetími osobami předané podklady a sdělení, jichž se mu od **poskytovatele licence** v souvislosti s uzavřením této smlouvy dostalo, ledaže **nabyvatel licence** s těmito třetími osobami uzavře smlouvu o mlčenlivosti, která bude obsahovat stejné závazky třetích osob, jaké má **nabyvatel licence** podle této smlouvy. Podklady a sděleními se rozumí zejména specifikace **autorského díla.**

2. Porušením povinnosti podle odst. 1 vzniká **nabyvateli licence** povinnost uhradit **poskytovateli licence** smluvní pokutu ve výši Kč 25.000,-, a to za každý jednotlivý případ porušení stanovené povinnosti. Ujednáním o smluvní pokutě není dotčeno právo **poskytovatele licence** na náhradu škody v plné výši.

VI.
Odměna

1. Smluvní strany se dohodly tak, že **licence se poskytuje bezúplatně.**

VII.
Doba trvání smlouvy

1. Tato smlouva se uzavírá na dobu neurčitou.

VIII.
Vyloučení záruky a omezení odpovědnosti

1. **Poskytovatel licence** prohlašuje a **nabyvatel licence** s tím souhlasí, že **poskytovatel licence** vytvořil **autorské dílo** s odbornou péčí, avšak nenese odpovědnost za případné chyby **autorského díla** týkající se jeho charakteru a jeho technických omezení.

2. **Poskytovatel licence** nezaručuje, že **autorské dílo** je vhodné pro jiný účel, než pro jaký byl stanoven **poskytovatelem licence**, a dále nezaručuje, že **autorské dílo** je kompatibilní s jakýmkoliv jiným dílem, systémem, přístrojem anebo produktem se kterým **autorské dílo nabyvatel licence** spojí či do kterého jej **nabyvatel licence** zařadí.

3. **Poskytovatel licence** nenese odpovědnost za případné škody vzniklé v důsledku užití **autorského díla nabyvatelem licence.**

IX.
Ukončení smlouvy

1. Tato smlouva zaniká:

   a) dohodou smluvních stran,
   b) zánikem **nabyvatele licence**,

2. Smluvní strany se dohodly, že ust. § 2382 zákona č. 89/2012 Sb., se nepoužije.

X.
Závěrečná ustanovení

1. Tato smlouva se v otázkách neupravených řídí občanským zákoníkem a autorským zákonem.

2. Veškeré spory vzniklé z této smlouvy budou smluvní strany přednostně řešit smírnou cestou. Nebude-li smírnou cestou dosaženo dohody, spory smluvních stran vyplývající z této smlouvy bude projednávat věcně a místně příslušný soud.

3. Tato smlouva se vyhotovuje ve dvou originálech; každá smluvní strana obdrží po jednom.

4. Změny smlouvy vyžadují písemnou formu a souhlas smluvních stran.

5. Nabyvatel bere na vědomí, že poskytovatel je subjektem povinným uveřejňovat smlouvy dle zákona č. 340/2015 Sb., a pokud tato smlouva splňuje podmínky pro uveřejnění dané zákonem, poskytovatel tuto smlouvu uveřejnění v registru smluv.
6. Smlouva nabývá platnosti dnem jejího uzavření, tj. dnem podpisu smlouvy oprávněnými zástupci obou smluvních stran. Smlouva nabývá účinnosti dnem jejího uzavření, jde-li o smlouvu podléhající uveřejnění v registru smluv dle zákona č. 340/2015 Sb., pak teprve dnem uveřejnění v registru smluv.
7. Smluvní strany prohlašují, že si tuto smlouvu před jejím podpisem přečetli, že byla uzavřena po vzájemném projednání podle jejich pravé a svobodné vůle, určitě, vážně a srozumitelně, nikoli v tísni a nikoli za nápadně nevýhodných podmínek. Autentičnost této smlouvy potvrzují svými podpisy.

V PLZNI dne 10. 03. 2020

poskytovatel licence
Západočeská univerzita v Plzni
prorektor pro výzkum a vývoj

V PRAZE dne 14-02-2020

nabyvatel licence

Příloha č. 1 : Specifikace autorského díla:

Technická a uživatelská dokumentace

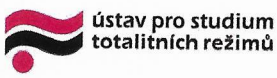
HIDOAR

- software pro polouatomatické zpracování a zpřístupnění textových a zvukových nahrávek v integrovaném archivu pramenů

(hlavní výsledek projektu – typ R)

Autoři: Jan Švec, Josef V. Psutka, Aleš Pražák, Petr Stanislav, Petr Neduchal, Marek Hrúz, Daniel Soutner, Jan Zelinka, Zbyněk Zajíc, Pavel Ircing, Martin Popel, Jan Hajič, Luděk Müller

Projekt DG16P02B048: Systém pro trvalé uchování dokumentace a prezentaci historických pramenů z období totalitních režimů



Obsah

Projekt DG16P02B048: Systém pro trvalé uchování dokumentace a prezentaci historických pramenů z období totalitních režimů

1 Úvod

Hlavním účelem předkládaného softwaru HIDOAR je převést zvukové nahrávky bilančních rozhovorů s pamětníky a relevantní listinné dokumenty obsahující text do takové strojově čitelné podoby, která umožní uživatelům v těchto materiálech efektivně vyhledávat zadaná klíčová slova či krátké fráze. Při vývoji tohoto softwaru byl kladen důraz nejen na efektivitu (tj. rychlost a přesnost vyhledávání), ale také na uživatelsky přívětivé rozhraní.

2 Technická dokumentace

2.1 Základní princip fungování softwaru

Klíčovými komponentami softwaru HIDOAR jsou modul pro rozpoznávání mluvené řeči (v tomto dokumentu budeme dále používat zavedenou zkratku ASR – z angl. *Automatic Speech Recognition*) a modul pro optické rozpoznávání znaků (OCR – *Optical Character Recognition*). Při bližším pohledu je zřejmé, že oba tyto moduly produkují na svém výstupu elektronický text – vstupem je pak buďto zvukový záznam řeči (u ASR) nebo grafický zápis přirozeného jazyka (OCR). Nicméně tento výsledný elektronický text není vhodné ukládat v neupravené podobě a to ze dvou důvodů - zaprvé se jeho chybovost pohybuje kolem 30% (tj. přibližně tři slova z deseti jsou chybně rozpoznána) a "surový" výstup rozpoznávacích modulů tedy není dobře čitelný a zadruhé následné vyhledávání v takovém textu není dostatečně efektivní. Proto používáme tzv. index, což je databázová reprezentace umožňující rychlé vyhledávání libovolného slova či krátké fráze. Tento index je vytvořen předem ze všech nahrávek a listinných dokumentů v dané sadě a při vlastním vyhledávání tedy již k rozpoznávání řeči či tištěných znaků nedochází.

Pro vyhledávání pak slouží webová aplikace nabízející komfortní uživatelské rozhraní. Takové řešení má celou řadu výhod:

- dostupnost odkudkoliv, kde je připojení k internetu
- uživateli stačí běžný moderní webový prohlížeč
- zařízením, na kterém bude uživatel pracovat, může být jak PC, tak např. tablet
- není třeba instalovat cokoliv na uživatelském zařízení (kromě samotného webového prohlížeče, který je však již předinstalován na téměř všech zařízeních)
- multiplatformnost (aplikace funguje nezávisle na OS zařízení)
- všichni uživatelé mají k dispozici stejnou verzi aplikace
- aktualizace aplikace probíhá na serveru (pokud je tedy aplikace aktualizována, mají všichni uživatelé k dispozici tuto aktualizovanou verzi)

2.2 Parametry použitých modulů

2.2.1 ASR modul

Samotný ASR modul obsahuje dvě komponenty, z velké části na sobě nezávislé - akustický a jazykový model. Akustický model – jak název napovídá – se snaží co nejlépe zachytit akustické vlastnosti zpracovávaných nahrávek. Jeho parametry se trénují metodami strojového učení s využitím doslovných přepisů řečových nahrávek. Při anotaci nahrávek z archivu ÚSTR byl využíván také software pro zarovnání

nahrávek s existujícími přepisy ALIGN, vyvinutý v rámci aktuálního projektu v první etapě řešení. Výsledný akustický model použitý v softwaru HIDOAR je natrénován s použitím dat od 35 řečníků z archivu nahrávek ÚSTR.

Jazykový model pak pomocí pravděpodobností modeluje způsob, jakým jsou jednotlivá slova řazena za sebou. Tyto pravděpodobnosti se opět odhadují (trénují) metodami strojového učení, tentokrát však jako trénovací data postačí pouze texty v elektronické podobě. Zatímco v případě akustického modelu potřebujeme, aby trénovací data byla akusticky podobná nahrávkám, které výsledný systém bude zpracovávat, v případě jazykového modelu vyžadujeme tématickou podobnost trénovacích textů. Výsledný jazykový model je natrénován na přepisech rozhovorů z archivu ÚSTR (1.4 mil. slov) a elektronických publikací obsahujících svědectví pamětníků totalitních režimů (260 tis. slov).

2.2.2 OCR modul

Pro vlastní rozpoznávání znaků byl použit volně dostupný program Tesseract, který patří mezi nejlépe hodnocené OCR nástroje. Naše testy ukázaly, že velmi dobře funguje i jeho verze pro češtinu. Jelikož kvalita OCR značně závisí na správně zvolené technice předzpracování vstupního "surového" obrazu, bylo otestováno téměř 50 metod předzpracování. Dále byly úspěšně aplikovány metody korekce nežádoucího natočení dokumentů během skenování a metody odstranění „optického" šumu na naskenovaných dokumentech.

2.2.3 Index

Software využívá univerzální platformu pro indexaci řečových dat vyvinutou na pracovišti hlavního řešitele projektu. Jde o distribuovanou (cloudovou) architekturu s následujícími komponenty:

- *MongoDB* - dokumentová databáze s podporou databázových clusterů, ukládání všech vstupních souborů a mezivýsledků
- *Python* - implementační jazyk pro server-side framework
- *Tornado* a *Motor* (*MongoDB Tornado*) - framework pro event-driven (asynchronní) programování vysoce propustných webových aplikací (server-side), Motor pro asynchronní interakci s databází
- *iSCSI* - protokol pro připojení MongoDB úložiště na výkonné výpočetní stroje
- *SpeechTech LVCSR technologie* - pro rozpoznávání řeči v reálném čase podporující výstup v podobě slovní/fonémové mřížky

Pro potřeby softwaru HIDOAR bylo – jak vyplývá z výše uvedeného popisu – třeba dodat akustický a jazykový model připravený na míru zpracovávaným datům. Dále bylo nutné uvedenou indexovací platformu podstatným způsobem modifikovat tak, aby byla schopna využívat softwarový balík Tesseract při indexaci listinných dokumentů. Uživatelské rozhraní bylo vyvinuto zcela na míru uživatelům a bude dále ve spolupráci s partnery projektu upravováno.

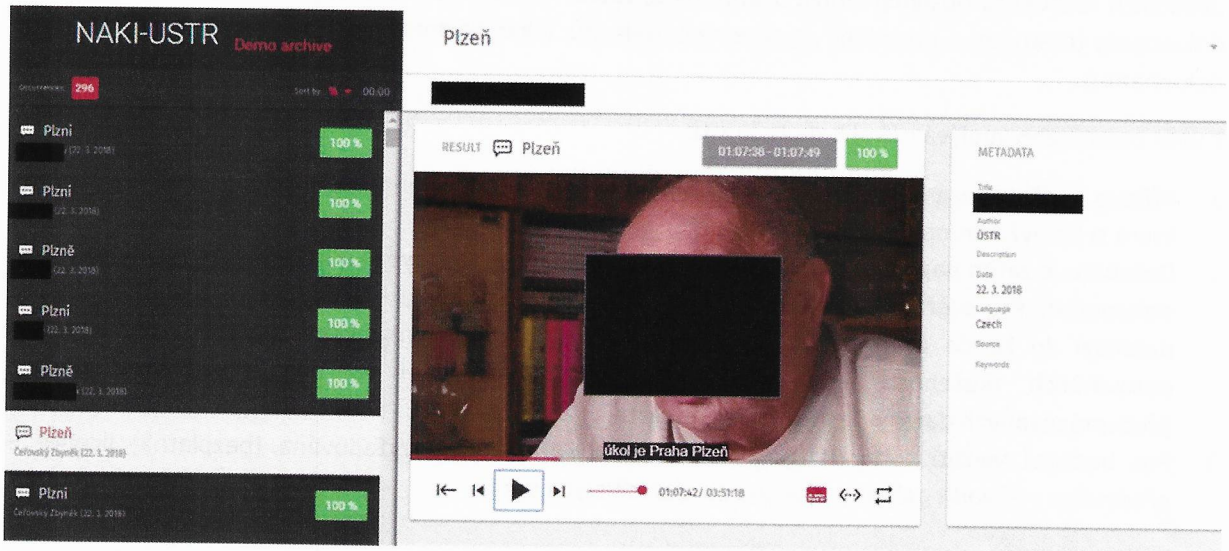
2.3 Hardwarové požadavky

- CPU: Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz
- RAM: 128GB
- 2x600GB pevný disk
- 10Gb ethernet

3 Uživatelská dokumentace

3.1 Návod na použití programu

Podoba stránky s výsledky je znázorněna na Obrázku 1. Stránka obsahuje čtyři prvky důležité pro ovládání aplikace. V horní části je to zejména pole pro zadání hledaného výrazu, v levé části seřazený seznam s výsledky pro vyhledaný dotaz. Největší část zaujímá detail výsledku s přehrávačem a výpisem dostupných metadat.



Obrázek 1 - Ukázka rozhraní softwaru HIDOAR

Zadávací pole slouží k textovému zápisu hledaného výrazu. Stejně jako vyhledávače od Googlu či Seznamu podporuje speciální operátory umožňující přesnější specifikaci hledaných slov. Hlavním účelem tohoto pole je možnost zadat seznam hledaných slov. Po iniciaci vyhledávání jsou automaticky vyskloňována. V některých situacích je automatické skloňování nežádoucí, a proto je možné hledané slovo „obalit" pomocí uvozovek. Všechna slova (skloňovaná i neskloňovaná) jsou tzv. **povinná.** To znamená, že všechna musí nalezena v určitém časovém intervalu, aby byl výsledek považován za platný. Tento interval je možné nastavit v nabídce rozšířeného vyhledávání. Označení slova jako nepovinného je možné pomocí znaku + umístěného na začátek slova.

Nalezené výsledky jsou zobrazeny v seznamu v levé části aplikace. Jednotlivé položky zobrazují nalezený text, název zdrojového videa a míru důvěry. Ta představuje číselné ohodnocení „kvality" výsledku. Zobrazené výsledky je možné řadit podle míry důvěry nebo data a času. Po kliknutí na libovolnou položku v seznamu je zobrazen odpovídající detail výsledku.

Hlavní funkcí přehrávače je přehrání vybraného úseku (odpovídajícího nalezenému výsledku). Po přehrání úseku se automaticky pokračuje přehráním následujícího výsledku. Pořadí přehrávání odpovídá řazení výsledků v seznamu s výsledky. Přehrávač disponuje interaktivní časovou osou, na které je zobrazena aktuální pozice v rámci videa a také další výsledky v tomto videu.

V současné verzi jsou slova hledána jak v přepisu mluvené řeči (v seznamu na levé straně okna jsou tyto nalezené výskyty označeny ikonou ), tak v listinných dokumentech (ikona )

3.2 Podmínky využití

Software HIDOAR sice využívá některé licencované technologie (*SpeechTech LVCSR)*, jelikož jsou ale tyto technologie použité pouze na serveru při indexaci a uživatel k nim nemá přístup, není tato skutečnost na překážku volnému využívání popisovaného softwaru.

Problém však nastává s korektním režimem práce s daty obsaženými v indexu. Zpracovávané nahrávky bilančních rozhovorů obsahují citlivá osobní data, totéž – spíše ještě ve větší míře – platí i pro listinné dokumenty (často jde o protokoly z policejních výslechů, lékařské zprávy, osobní dopisy a další důvěrné dokumenty).

Z výše uvedeného vyplývá několik skutečností:

1. Přístup k výše popsané webové aplikaci HIDOAR musí být chráněn uživatelským jménem a heslem, které si lze vyžádat na e-mailové adrese ircing@kky.zcu.cz
2. Databáze je zatím naplněna pouze několika málo desítkami rozhovorů a několika stovkami listinných dokumentů. Pro ověření funkčnosti softwaru je to zcela dostačující - na řádově větší datové sadě, jíž databázi do konce projektu plánujeme naplnit, by kvalita vyhledávání měla být podle našich dosavadních zkušeností srovnatelná a rychlost vyhledávání se díky efektivnímu uložení předzpracovaných dat v indexu sníží jen nepatrně.
3. Pro budoucí využití softwaru ostatními subjekty bude vždy vyžadována (bezplatná) licence – především z důvodu ochrany výše uvedených citlivých dat.