Prepis hovorenej reči, podpora slovenského jazyka

Autor: Samuel Baran, 3Ib

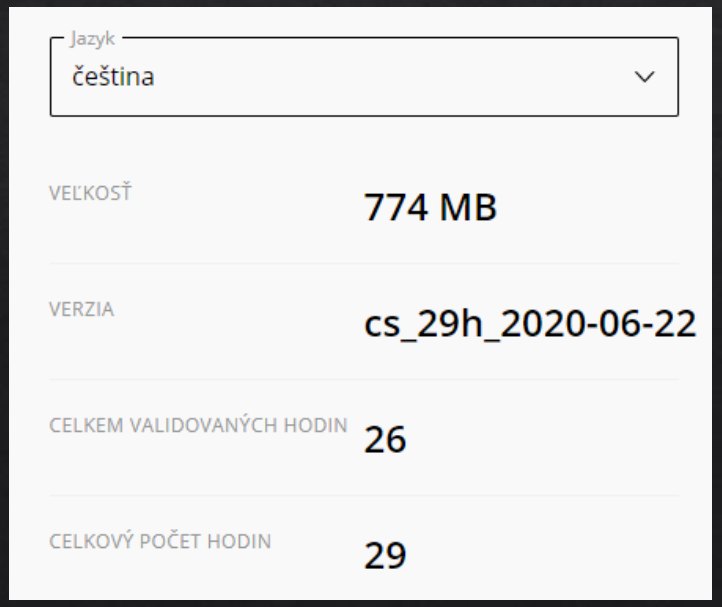
Vedúci: RNDr. Erik Bruoth, PhD.

Ciele

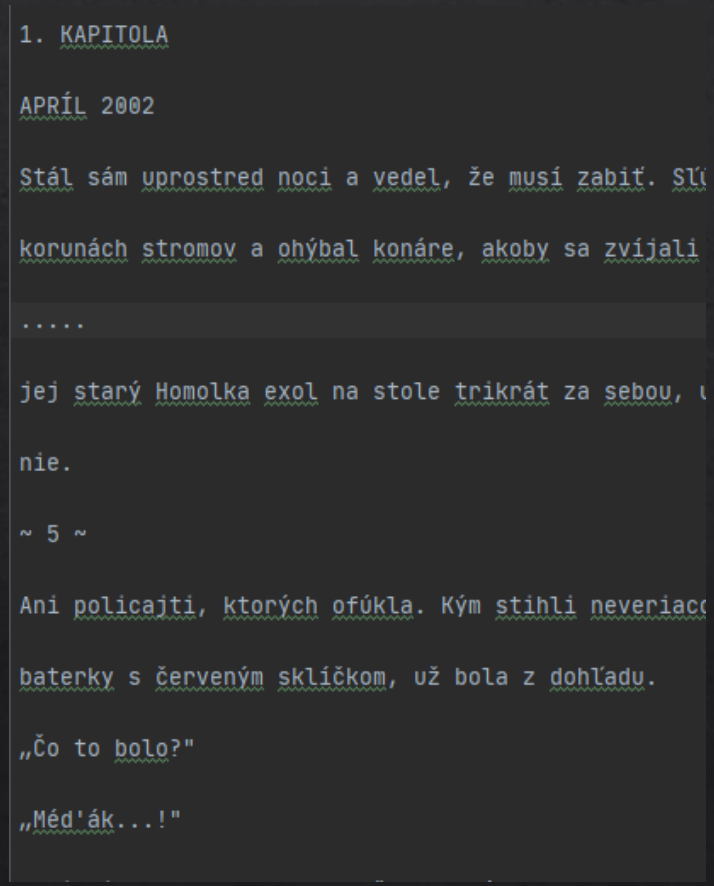
- Analyzovať existujúce prístupy STT
- Vytvoriť tréningovú sadu pre úlohu pomocou dostupných zdrojov
- Identifikovať open source implementácie modelov (min. 2) rozpoznávania reči (STT)
- Navrhnúť stratégiu tréningu modelov a ladenia hyperparametrov
- Základné porovnanie modelov na základe vybraných metrík resp. popis hlavných problémov

Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

- Datasety iniciatívy Mozilly Common Voice
  - 24 aktívnych jazykov (aj čeština) ďalšie pripravujú (slovenčinu :/)



Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov



```
1. KAPITOLA

APRÍL 2002

Stál sám uprostred noci a vedel, že musí zabiť. Slú

korunách stromov a ohýbal konáre, akoby sa zvíjali

.....

jej starý Homolka exol na stole trikrát za sebou, u

nie.

~ 5 ~

Ani policajti, ktorých ofúkla. Kým stihli neveriac

baterky s červeným sklíčkom, už bola z dohľadu.

„Čo to bolo?"

„Méd'ák...!"
```

- spracovanie audiokníh
- 3 knihy ~ 50 hodín (e-kniha + audiokniha)
  - e-kniha epub
  - audiokniha
    - mp3 pre každú kapitolu
- pooloautomatizované spracovanie
  - extrahovanie raw textu z epub publikácie knihy
    - online converter (málo kníh)
  - rozsekanie textu na kapitoly (korešpondujúce s mp3)
    - python skript
    - úpravy “na mieru” (odstránenie úvodu, označenia kapitol)

Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

- stav datasetu:
  - počet (nahrávka, prepis) - rádovo v desiatkach
  - trvanie nahrávky - rádovo v desiatkach minút
- požiadavky modelu:
  - trvanie nahrávky – jednotky – desiatky sekúnd
  - počet – čo najviac

Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

- stav datasetu:
  - počet (nahrávka, prepis) - rádovo v desiatkách
  - trvanie nahrávky - rádovo v desiatkách minút
- požiadavky modelu:
  - trvanie nahrávky – jednotky – desiatky sekúnd
  - počet – čo najviac
- riešenie
  - rozsekať kapitoly na menšie jednotky (vety)

Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

Forced alignment

- zarovnanie ortografických prepisov a zvukových záznamov
- väčšina nástrojov založená na ASR (automatic speech recognition)
- Princíp:
  - rozpoznanie obsahu nahrávky (STT)
  - namapovanie na originálny text

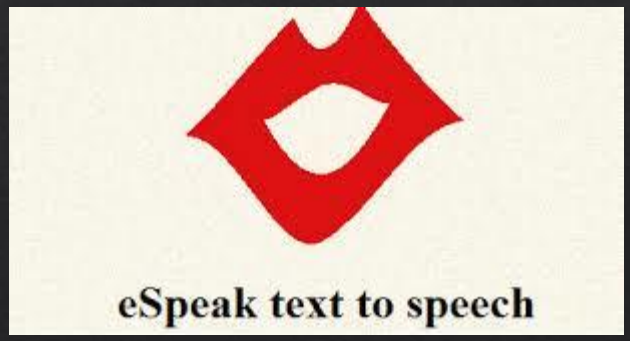
```
1                                                  => [00:00:00.000, 00:00:02.640]
From fairest creatures we desire increase,         => [00:00:02.640, 00:00:05.880]
That thereby beauty's rose might never die,        => [00:00:05.880, 00:00:09.240]
But as the riper should by time decease,           => [00:00:09.240, 00:00:11.920]
His tender heir might bear his memory:             => [00:00:11.920, 00:00:15.280]
But thou contracted to thine own bright eyes,      => [00:00:15.280, 00:00:18.800]
Feed'st thy light's flame with self-substantial fuel, => [00:00:18.800, 00:00:22.760]
Making a famine where abundance lies,              => [00:00:22.760, 00:00:25.680]
Thy self thy foe, to thy sweet self too cruel:     => [00:00:25.680, 00:00:31.240]
Thou that art now the world's fresh ornament,      => [00:00:31.240, 00:00:34.400]
And only herald to the gaudy spring,               => [00:00:34.400, 00:00:36.920]
Within thine own bud buriest thy content,          => [00:00:36.920, 00:00:40.640]
And tender churl mak'st waste in niggarding:       => [00:00:40.640, 00:00:43.640]
Pity the world, or else this glutton be,           => [00:00:43.640, 00:00:48.080]
To eat the world's due, by the grave and thee.     => [00:00:48.080, 00:00:53.240]
```

Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

Forced alignment - Aeneas

- produkt firmy ReadBeyond
  - Audio e-books (synchronizácia audioknihy a eknihy epub)
- Python/C knižnica
- odlišný prístup založený na spracovávaní signálu
- DTW (dynamic time wraping)
- TTS (text to speech)

- Princíp:
  - syntéza audia z textu (TTS)
    - eSpeak (aj SK)
  - hľadanie zhody audia a časti mp3



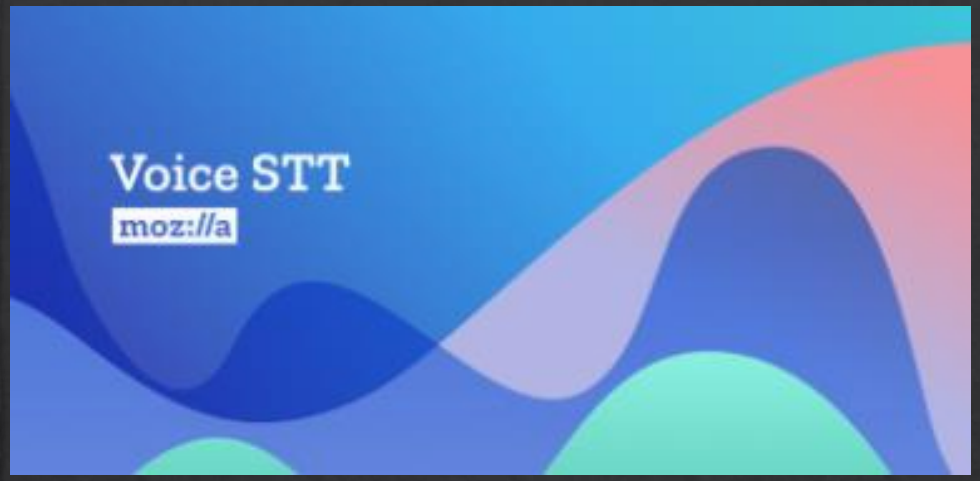
Vytvorenie tréningovej sady pre úlohu pomocou dostupných zdrojov

Aktuálny stav:

- 11 598 audiosegmentov dĺžky ~ 10s
- približne 30 hodín

Ďalšie možnosti:

- Spracovanie podcastov
  - potrebné presné prepisy
  - Newsletter
- Slovenský národný korpus https://korpus.sk/
  - bezplatné využitie na vedecko-výskumné ciele
  - k 22.07.2020 približne 1,65 mld tokenov
  - hovorený korpus ~ 700 hodín



Identifikovanie open source implementácie modelov (min. 2) rozpoznávania reči (STT)

MOZILLA

- Voice STT DeepSpeech
  - Tensorflow
  - end to end rozpoznávania reči
    - konverzia dataframe-ov na fonémy
    - generovanie textu z postupnosti foném
    - kombinovanie s jazykovým modelom
- Common voice
  - Iniciatíva zbierajúca dáta
- TTS

Ďalšie ciele

- Analyzovať existujúce prístupy STT
  - Získať prehľad v už existujúcich prístupoch STT konverzie
  - Porovnať dané prístupy
- Navrhnúť stratégiu tréningu modelov a ladenia hyperparametrov
- Základné porovnanie modelov na základe vybraných metrík resp. popis hlavných problémov

Najbližšie kroky

- Naštudovať aktuálne prístupy spracovávania prirodzeného jazyka s použitím hlbokého učenia
- Preskúmať architektúru open source implementácie STT algoritmu z dielne Mozilly
- Nájsť ďalšiu implementáciu STT algoritmu
- Získať väčší dataset
- Trénovať model
- Identifikovať hyperparametre modelov

Ďakujem za pozornosť

Otázky?

Literatúra

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf)
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning (http://www.deeplearningbook.org)
- Jacob Eisenstein. Natural Language Processing
- http://web.stanford.edu/class/cs224n/
- https://github.com/espnet/interspeech2019-tutorial
- https://github.com/mozilla/DeepSpeech open source

Literatúra

- https://medium.com/@techfirst/forced-alignment-how-to-match-audio-with-a-transcript-via-machine-learning-dd19da8c0f04
- https://medium.com/@klintcho/creating-an-open-speech-recognition-dataset-for-almost-any-language-c532fb2bc0cf
- https://github.com/readbeyond/aeneas
- https://www.readbeyond.it/aeneas/docs/
- https://usabilitygeek.com/automatic-speech-recognition-asr-software-an-introduction/
- https://github.com/readbeyond/aeneas/blob/master/wiki/HOWITWORKS.md
- https://medium.com/@techfirst/forced-alignment-how-to-match-audio-with-a-transcript-via-machine-learning-dd19da8c0f04
- http://espeak.sourceforge.net/