

Biblioteki cyfrowe - projekt

Jerzy Dembski



Części projektu

1. Akwizycja i przygotowanie obiektów cyfrowych
2. Indeksacja zawartości obiektów, generacja metadanych
3. Badanie jakości obiektów
4. Publikacja obiektów w bibliotece cyfrowej (dLibra)



1. Akwizycja i przygotowanie obiektów cyfrowych

Typy obiektów dla poszczególnych terminów zajęć:

- Czwartek 15:15 - 18:00 - nagrania (akwizycja tekstów czytanych).
- Czwartek 18:15 - 21:00 - nagrania (akwizycja tekstów czytanych).

Nagrywanie czytanego przez siebie tekstu. Cechy nagrań:

- styl własny – naturalny (podobny do codziennych wypowiedzi),
 - jakość - co najmniej 22 kHz, 16 bitów, mono,
 - zapis w pliku wav,
 - tekst wypowiedzi zostanie umieszczony na stronie przedmiotu do 14.03.2019,
 - należy wykonać 10 nagrań w podobnych warunkach (tym samym urządzeniem o tych samych ustawieniach).
-
-

1. Akwizycja i przygotowanie obiektów cyfrowych

Realizacja akwizycji danych:

- Dokonanie nagrań/ skanów.
- Obróbka (np. usunięcie dłuższych przerw/wycięcie nieistotnego fragmentu obrazu).

Urządzenia do cyfryzacji:

- Mikrofon, kamera, urządzenie mobilne.



2. Indeksacja zawartości obiektów, generacja metadanych

Indeksacja tekstów czytanych – oprogramowanie:

- Program własny dostępny na stronie – obecnie jest w fazie dopasowywania do wymagań w roku 2018/19
- <https://www.speechmatics.com>

- <https://github.com/Kyubyong/deepvoice3> (Deep Voice)
- <https://github.com/sotelo/parrot> (Char2wav)
- <https://github.com/facebookresearch/loop> (Voice Loop)
- http://mowa.clarin-pl.eu/korpusy/clarin_emu.zip – przykładowy korpus mowy polskiej

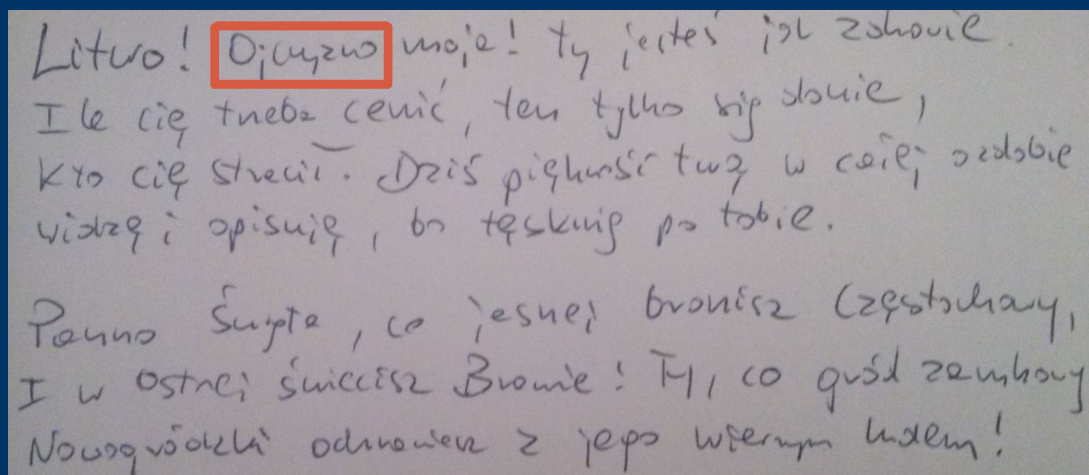
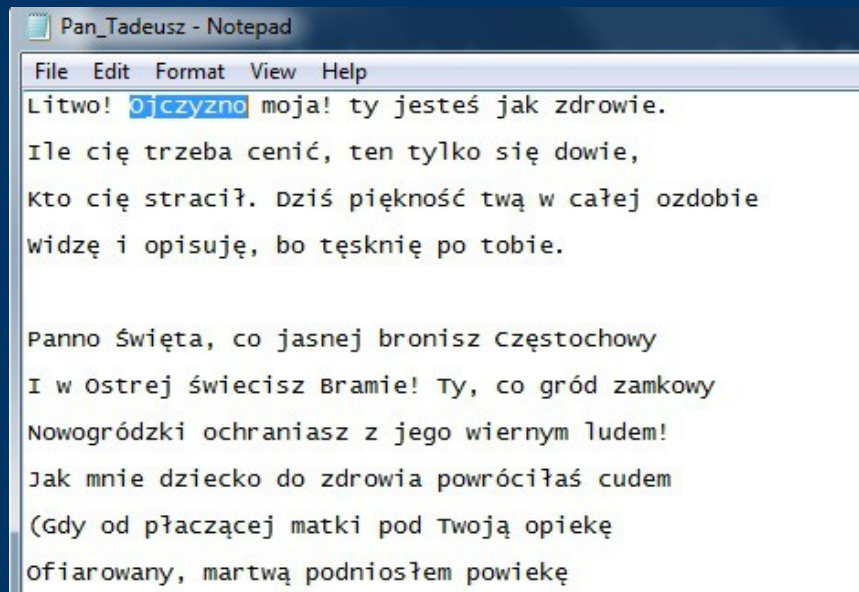
- system rozpoznawania mowy dostępny pod adresem:
<https://github.com/mozilla/DeepSpeech> (Deep Speech)
Uwaga: to wymaga dostępu do naszego serwera wydziałowego ze względu na czas uczenia.

- (<https://github.com/palles77/julius>, oryginał na <https://github.com/julius-speech/julius>, także jako pakiet Linuksa) i modele dla języka polskiego dostępne pod adresem:
<https://sourceforge.net/projects/juliusmodels/files/>
- <https://github.com/danijel/ClarinStudioKaldi> – system rozpoznawania mowy Kaldi dla języka polskiego.

- systemu segmentacji mowy
<https://github.com/julius-speech/segmentation-kit>

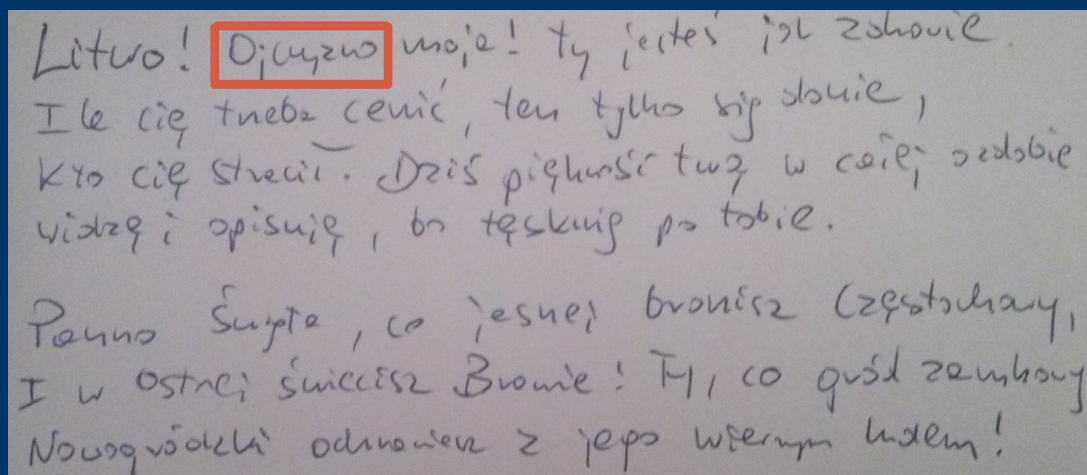
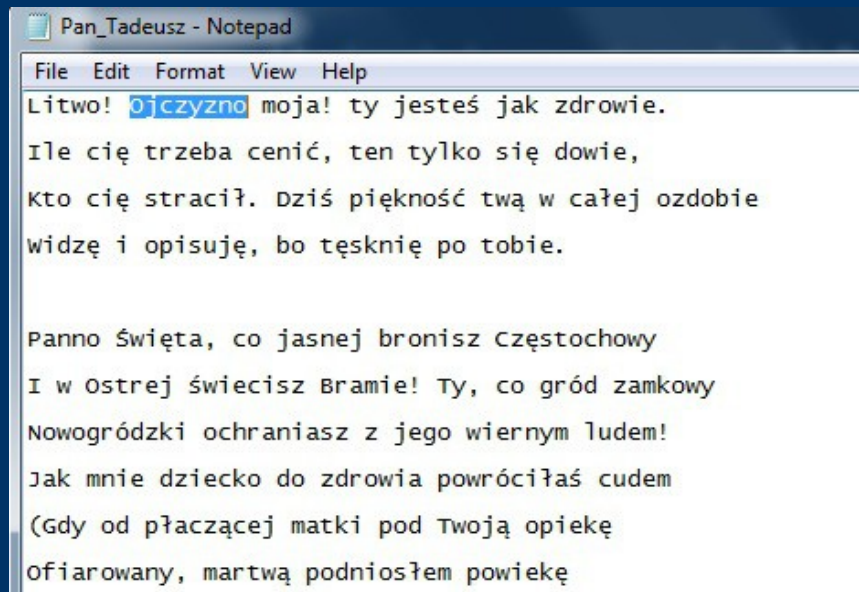
2. Indeksacja zawartości obiektów, generacja metadanych (2017/18)

Indeksacja tekstów pisanych



2. Indeksacja zawartości obiektów, generacja metadanych (2017/18)

Indeksacja tekstów pisanych



```
<skan>
  <numer>1</numer>
  <napis>
    <x>156</x>
    <y>46</y>
    <wys>23</wys>
    <szer>57</szer>
    <słowo>Ojczyzno</słowo>
  </napis>
</skan>
```

2. Indeksacja zawartości obiektów, generacja metadanych (2017/18)

Metadane – informacje wymagane:

- Autor nagrania/rękopisu
 - Osoba dokonująca indeksacji
 - Data nagrana/powstania rękopisu
 - Dane nagrania: częstotliwość próbkowania, liczba bitów, ...
/ dane urządzenia skanującego: rozdzielczość, balans bieli, ...
 - Dane urządzenia nagrywającego/skanującego
 - Subiektywna ocena jakości
-
-

2. Indeksacja zawartości obiektów, generacja metadanych (2017/18)

Projekt pliku z metadanymi:

- Format XML.
- Format Tekstowy/JSON



3. *Badanie jakości obiektów* (2017/18)

Procedura standardowa:

- Dla danego zbioru nagrań/rękopisów losowane są trzy osoby do przeprowadzenia testów dla dwóch wariantów oceny: kontekstowej i bezkontekstowej (szczegóły w instrukcji do części 3).
 - Dla każdej osoby testującej losowany jest fragment tekstu złożony z co najmniej 1000 słów.
 - Osoba testująca (nie może nią być autor nagrania/rękopisu) określa procentową zawartość wypowiedzi/napisów dla niej niezrozumiałych/nieczytelnych oraz takich, które stają się zrozumiałe/ czytelne dopiero po dłuższym zastanowieniu (szczegóły oceny jakości znajdują się w instrukcji do 3 części projektu).
 - Wyniki powinny zostać uśrednione i umieszczane w metadanych
-
-

4. Publikacja obiektów w bibliotece cyfrowej (dLibra) (2017/18)

- Stworzenie strony prezentującej dokumenty cyfrowe wraz z ich opisem, indeksacją oraz oceną jakości w formie przystępnej dla czytelnika bez przygotowania informatycznego.
 - Zamieszczenie strony wraz z danymi w bibliotece cyfrowej dLibra zgodnie z instrukcją.
-
-