

Lekturo-skurczator - etap 1

Dominik Doberski 133207, Michał Kukieła 132265

Tworzenie zbioru uczącego

Do utworzenia zbioru uczącego posłużyła nam lektura “Nad Niemnem” Elizy Orzeszkowej, którą pobraliśmy w formacie .epub ze strony

<https://wolnelektury.pl/katalog/lektura/nad-niemnem/>

Plik w formacie .epub przekonwertowaliśmy na plik .txt za pomocą biblioteki *epub_conversion*

<https://pypi.org/project/epub-conversion/>

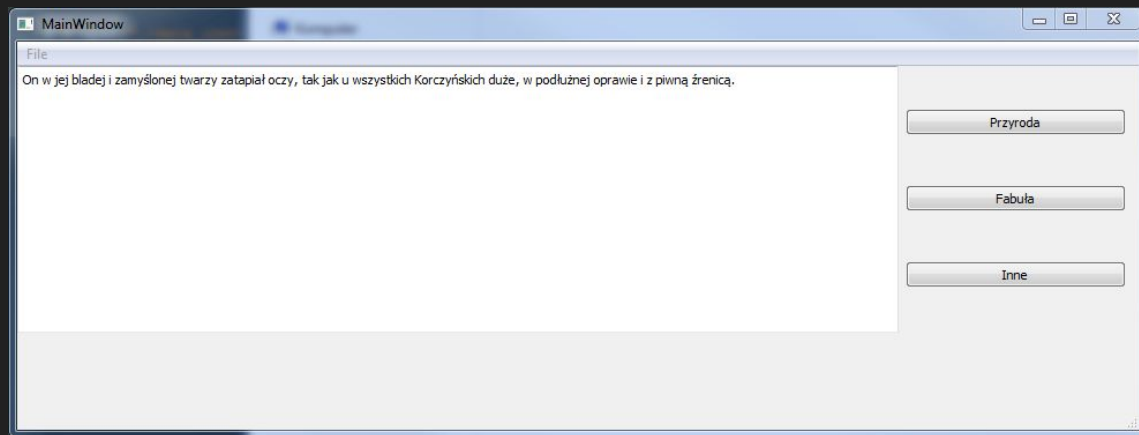
Następnie dokonaliśmy podziału całej książki na pojedyncze zdania. Te zdania to nasze przykłady uczące, którym musieliśmy przypisać etykiety (klasy) zgodnie z obraną przez nas konwencją: opis przyrody, fabuła, inne.

Tworzenie zbioru uczącego

- Opis przyrody - nazwa klasy mówi sama za siebie. Przypisywaliśmy tę klasę zdaniom, w którym poza opisem przyrody nie występowały treści istotne z punktu widzenia fabuły (różne akcje, zdarzenia, działania postaci), lub było ich niewiele i sprawiały wrażenie nieistotnych.
- Inne - głównie opisy nie będące opisami przyrody, treści nieistotne z punktu widzenia fabuły.
- Fabuła - wszelkie akcje, zdarzenia, działania postaci, fakty, informacje - wszystko, co nie należy do żadnej z pozostałych dwóch klas (większa część książki).

Tagowanie zbioru uczącego

Stworzyliśmy program pozwalający nam przypisywać każdemu zdaniu ze zbioru uczącego jedną z 3 klas.



Tagowanie zbioru uczącego

Uzyskaliśmy około 1200 otagowanych zdań (można oczywiście jeszcze powiększyć ten zbiór), w tym:

782 - Fabuła

329 - Inne

81 - Przyroda

Repozytorium projektu

https://github.com/Dombearx/TP_project