



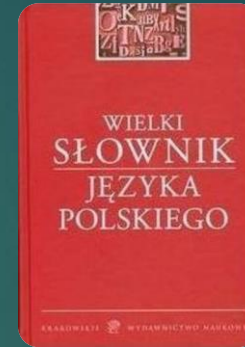
Jak napisać spell-
corrector?

Dane

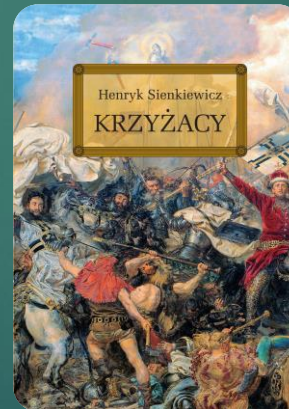
Big.txt

NARODOWY KORPUS
JĘZYKA POLSKIEGO

Podkorpus milionowy
Ręcznie anotowany milionowy podkorpus NJKP,



Słownik języka polskiego



Krzyżacy - Sienkiewicz Henryk

Wyciągnięcie danych z plików xml

NARODOWY KORPUS JĘZYKA POLSKIEGO

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude" xmlns="http://www.tei-c.org/ns/1.0">
  <xi:include href="NKJP_1M_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text xml:id="txt_text" xml:lang="pl">
      <body xml:id="txt_body">
        <div xml:id="txt_1-div" decls="#h_1-bibl">
          <ab n="p911in935of:PNW:030-2-000000002" xml:id="txt_1.1-ab">Gdy jej rogiaty łeb ukazał się w chaszczech, Coto odetchnął i puścił się jej śladem, na nic nie zważył, że jej śladem nie kwapiła, boby jej nie sprostał.</ab>
          <ab n="p912in936of:PNW:030-2-000000002" xml:id="txt_1.2-ab">Szczęściem, że była syta, więc się zbytnio do obroku nie kwapiła, boby jej nie sprostał.</ab>
        </div>
        <div xml:id="txt_2-div" decls="#h_2-bibl">
          <ab n="p658in682of:PNW:030-2-000000002" xml:id="txt_2.1-ab">I zawrócił na rzekę.</ab>
          <ab n="p659in683of:PNW:030-2-000000002" xml:id="txt_2.2-ab">Znowu toń była gładka, niebo błękitne, rozegrały się ptaki, zapachniało powietrze. Gdy przybili, Ku
          <ab n="p660in684of:PNW:030-2-000000002" xml:id="txt_2.3-ab">łatana Skóra zariżała do nich ze stajenki i z rękawa swego wytknął pyszczyk Kuba.</ab>
        </div>
        <div xml:id="txt_3-div" decls="#h_3-bibl">
          <ab n="p2327in2364of:PNW:030-2-000000002" xml:id="txt_3.1-ab">Wyciągnęli się jak struny i z żywiołowym rozmachem zaśpiewali:</ab>
          <ab n="p2328in2365of:PNW:030-2-000000002" xml:id="txt_3.2-ab">Jeszcze Polska nie zginęła...</ab>
          <ab n="p2329in2366of:PNW:030-2-000000002" xml:id="txt_3.3-ab">Rwała się pieśń w zdławieniu wrażenia potężnego, w łzach, co znają tylko miłujący i mężni, i kwit
          <ab n="p2330in2367of:PNW:030-2-000000002" xml:id="txt_3.4-ab">Mogila wyrosła, złocąc się z daleka, a obecni pożegnali ją ostatnią obrzędową pieśnią "Anioł Pańs
        </div>
        <div xml:id="txt_4-div" decls="#h_4-bibl">
          <ab n="p2298in2335of:PNW:030-2-000000002" xml:id="txt_4.1-ab">Na długo przed zorzą wrócili z folwarku Odnowaków z deskami. Szczepański przyniósł narzędzia, i
          <ab n="p2299in2336of:PNW:030-2-000000002" xml:id="txt_4.2-ab">Na posłanie z ziół i kwiatów ułożono kości i pokryto je szczątkami sztandaru. Po żdzieble srebrnej
          <ab n="p2300in2337of:PNW:030-2-000000002" xml:id="txt_4.3-ab">Potem Rosomak podał mu strzelbę, a stary ją przyjął, jak święty spadek, i oddał Bartnikowi.</ab>
        </div>
        <div xml:id="txt_5-div" decls="#h_5-bibl">
          <ab n="p95in103of:PNW:030-2-000000002" xml:id="txt_5.1-ab">Nie było pastucha i jego bata, nie było psów ani łańcucha w dusznej oborze, ani zdradliwych rogów ze
        </div>
        <div xml:id="txt_6-div" decls="#h_6-bibl">
          <ab n="p2382in2419of:PNW:030-2-000000002" xml:id="txt_6.1-ab">Wracając z obchodu ku domowi, skreślił Rosomak na mogile Chorażego, i odśpiewał mu pieśni narodov
          <ab n="p2383in2420of:PNW:030-2-000000002" xml:id="txt_6.2-ab">Słońce gaśło. Żurawie nie grały, chłodne mgły wstawały z bagien, długie cienie kładły gąszcz. S
```

odkorpusMilionowy\030-2-000000002\text.xml extract_text.py data.txt

```
import re

import os

1 usage    Paweł Krzyściak
def extract_text_from_xml(file_path):
    with open(file_path, 'r', encoding='utf-8') as file:
        content = file.read()
        # Use regex to find all text within </ab> tags
        texts = re.findall(pattern=r'<(.*)></ab>', content)
        return texts

Paweł Krzyściak
def main():
    base_path = os.path.dirname(os.path.realpath(__file__))
    target_folder = os.path.join(base_path, 'PodkorpusMilionowy')
    output_file_path = os.path.join(base_path, 'data.txt')
    file_count = 0

    with open(output_file_path, 'w', encoding='utf-8') as output_file:
        for root, dirs, files in os.walk(target_folder):
            for dir_name in dirs:
                xml_file_path = os.path.join(root, dir_name, 'text.xml')
                if os.path.exists(xml_file_path):
                    file_count += 1
                    texts = extract_text_from_xml(xml_file_path)
                    for text in texts:
                        output_file.write(f"{text}\n")
                        print(f"{text}")

    print(f"Liczba plików text.xml: {file_count}")

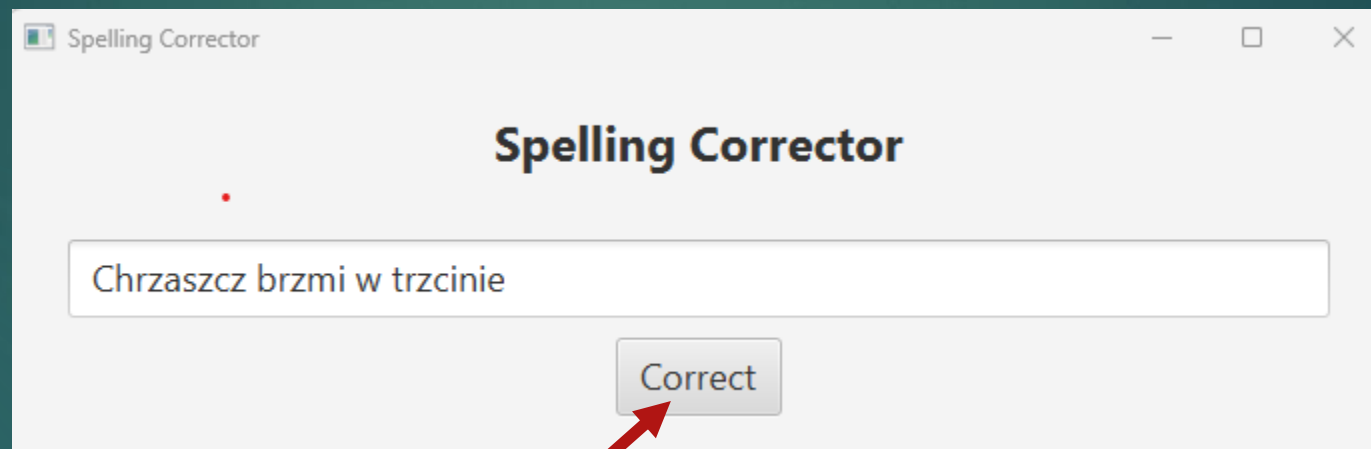
if __name__ == "__main__":
    main()
```

xml extract_text.py data.txt

file size (7,41 MB) exceeds the configured limit (2,56 MB). Code insight features are not available.

Zatrzasnął drzwi od mieszkania, dwa razy przekręcił klucz, nacisnął klamkę, by sprawdzić, czy dobrze zamknięte, zbiegł po schodach, minął furtkę, także ją za-
"Bohaterem powieści Paźniewskiego jest miasto, Krzemieniec."
"Jak za czasów Słowackiego funkcjonuje Liceum i płynie Ikwa. Krzemieniec powieściowy jest tamtym Krzemieńcem, ale jest także miastem wywołanym z osobistej pamięci
"Ale dzisiaj? Jaką dzisiaj odegra rolę poetyka Przybosa? Oczywiście, już sam fakt jej istnienia jest wartością. Nasza literatura, bogata w improwizacje i w akty
"Halina Auderska we wszystkich książkach każe swoim bohaterom szukać tożsamości. W "Babim lecie" ma odwagę uznać za najistotniejsze kryterium tożsamości poczucie
"Paźniewski w "Krótkich dniach" ofiarował Kresom nie mniej, niż z nich zaczerpnął. Zatrzymał potop. Zamówił kataklizm. Stworzył wizję oczekiwania, wizję spokoju
"Plama" Piętała, jedna spośród kilku najznakomitszych współczesnych powieści, także ze względu na jej zaklasyfikowanie wraz z całą twórczością tego pisarza do n
"To już nie są wątpliwości religijne, te wątpliwości pierwszego stopnia wtajemniczenia w sprawy świata, wątpliwości "Nieba w płomieniach" czy "Jana Barois".
"Tu nie chodzi o sensowność dogmatów czy ścisłość religijnych wyobrażeń, nie chodzi już o religię, o tajemnicę stworzenia, ale o normę etyczną. Kto ją ustanowi,
"tży padały na cremoński lakier. Szkoda, nie wolno było niszczyć przedmiotu, na który ojciec, biedaczysko, wydał całą schedę po Luizie... Otarła je lewą, umęczon
"Wszedł Adam. Zawiało wodą kwiatową Maréchal Niel. Starannie domykał drzwi za sobą."
"Trwała dalej w bezruchu."
"Pan dyrektor był nieobecny, a Róża stała pod piecem, czekając, Kiedy panna Aniela Bądska ukończy gamy i zechce łaskawie zaakompaniować jedynej uczennicy papy "M
"Nowe także były miesięczne wędrowniki na grób męża i sjęsty na cmentarnej ławeczce, ze wzrokiem żarliwie utkwionym w darninę. I tży po noca
"Prawie wesoła - dokończyła przy pomocy krokodyla zdejmowania chusty. Strzepnęła ją, złożyła, odniosła tamże, gdzie portrećik ojca."
"Telefon zadzwonił, upudrowała się gorączkowo, przybrała oschły wyraz twarzy - może Marta? - i podeszła do aparatu. Ledwie zdjęta słuchawkę, Sabina przyczoławała
"Róża milczała. Słuchała napomnień jak ptasiego szczebiotu, który do niczego nie obowiązuję, gdyż nic ludzkiego nie oznacza. Oczy powlekła biaława szklistość, co
"Uśmiechnęła się do córki ze swej błogiej dali."
"Radio gra "Close your eyes". Róża - wzburzona pominięciem przez córkę swojej wizyty, zirytowana brakiem szacunku tutaj, w tym zięciowskim domu, dla empirowego s
"Zerknęła w lustro, obciągnęła na siebie sweter w czerwone skrzydła..."
"Marta rzuciła telefon; stając na progu, stanęła twarzą w twarz z matką. Róża - wysoko osadzona - patrzyła przed siebie nad miarę otwartymi oczami. W tych oczach
"W okresie szkolnym nie inaczej przedstawiała się sprawa koleżanek. Raz do roku tylko bywały zapraszane. Róża wtedy występowała ze wspianiałym przyjęciem. Stół ug
"Ewa, Eveline - to było imię, które nadała jej Luiza, dla względów prestiżowych. Róża pamiętała dobrze ten dzień."
"Jesienią w niedzielę szły z ciotką na obrzędowy spacer do łaźnierek, placem Wareckim, ulicą Szpitalną, Bracką, Alejami... Dorożki i kabriolety turkotały po wyboi
"Wówczas Martę strach przejął. Jak to? Więc śpiew córki przestał być własną sprawą Róży? Więc pojawiły się jakieś inne sprawy? Prawda: codzienny, nieubłagany prz
"Marta ciskała nuty, zamykała się z trzaskiem u siebie."
"W ciągu następnych lekcji rozbijanie frazy dawało się powściągnąć, także styl i barwę uczuciową poddawała Róża; a jednak ostateczny rezultat pracy zawierał w so
"Róża opuściła ręce. Siadła - nogi drżały. Rzuciła smyczek..."
"Księżycowa orkiestra pod batutą Brahmsa grała dalej. Tylko na miejscu skrzypiec wystąpiła cisza - czarna jak zaskórna woda. Jeszcze tu i ówdzie błyskał refleks
"Gdy jej rogaty łeb ukazał się w chaszczach, Coto odetchnął i puścił się jej śladem, na nic nie zważając. Ale Matora też nie pilnowała się żadnego szlaku: rznąła
"Szczęściem, że była syta, więc się zbytnio do obroku nie kwapiła, boby jej nie sprostał."
"I zawrócił na rzekę."

Działanie



Kroki

- ▶ 0. Wczytywanie pliku Big.txt
- ▶ 1. Generowanie słów w odległości edycji pojedynczego znaku
np. Dla tszczyna = {szczyna, tzczyzna, tszczyn, tbzczyna (...)}
Elementy wygenerowane dla zdefiniowanego alfabetu :

```
public static final String ALPHABET = "abcdefghijklmnopqrstuvwxyząźćóęłż";
```

- ▶ 2. Filtracja znanych słów i znajdowanie najczęstszego znanego słowa w odległości pojedynczego znaku.

```
Optional<String> e1 = known(edits1(word)).max((a, b) -> dict.get(a) - dict.get(b));
```

- ▶ W przypadku gdy słowo będzie się zgadzać to zwraca je jeżeli nie to przechodzi do punktu 3

3. Znajdowanie najczęstszego znanego słowa w odległości 2 znaków.

```
Optional<String> e2 = known(edits1(word).flatMap(this::edits1)).max((a, b) -> dict.get(a) - dict.get(b));
```

Rezultat

Spelling Corrector

chrzaszcz brzmis w tzcinie

Correct

chrząszcz brzmi w trzcinie

Model

Do lepszej poprawy tekstu w tym interpunkcji dodano model **sdadas/byt5-text-correction**

Przeznaczony do prostej korekty tekstu. Ma na celu poprawę jakości tekstów pochodzących z sieci, często pozbawionych interpunkcji lub właściwej wielkości liter. Model został wytrenowany do wykonywania trzech typów korekcji:

- ▶ Przywracanie interpunkcji w zdaniach.
- ▶ Przywracanie wielkich liter w słowie.
- ▶ Przywracanie znaków diakrytycznych dla języków, które je zawierają.


```
import torch
import sys
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

1 usage
def correct_text(text):
    model_name = "sdadas/byt5-text-correction"
    model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
    tokenizer = AutoTokenizer.from_pretrained(model_name, model_max_length=512)

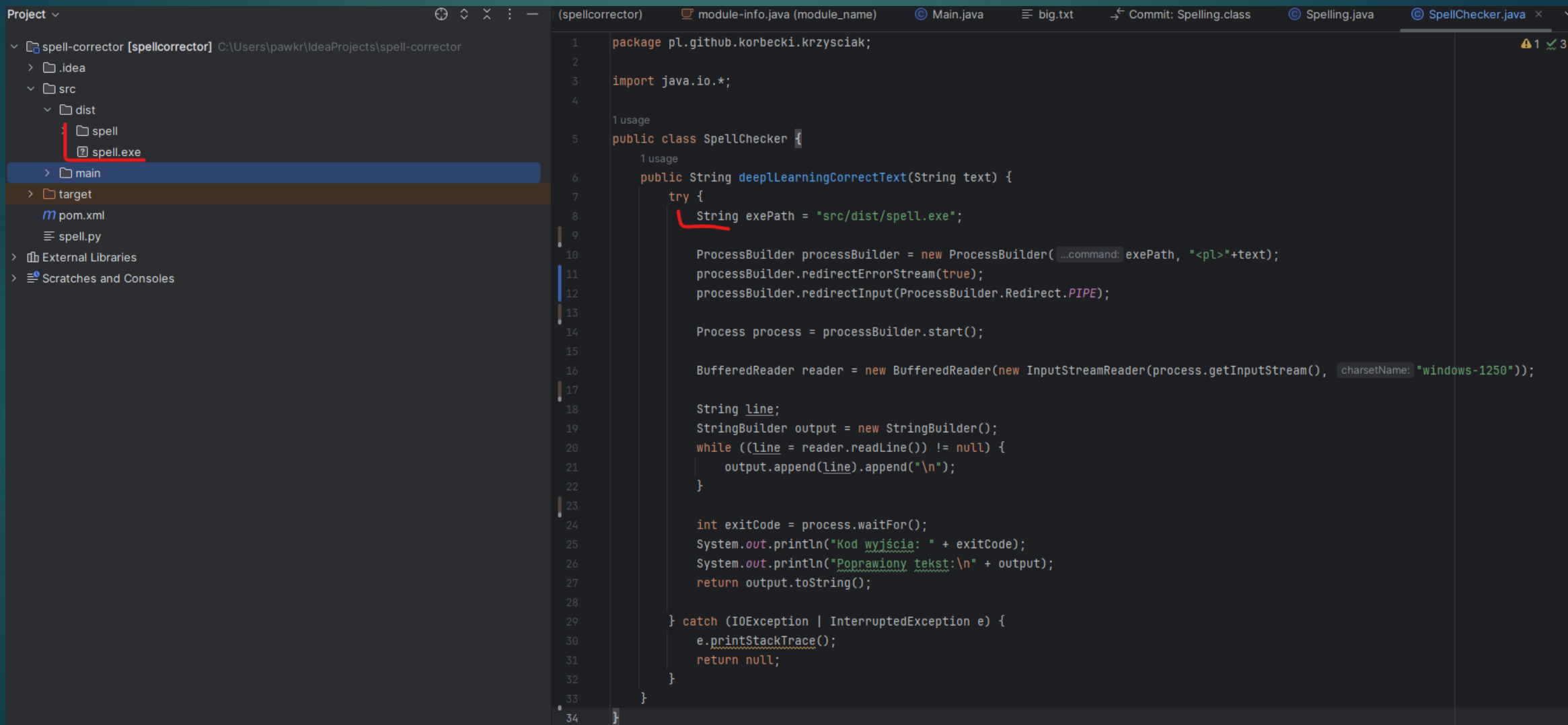
    # Tokenizuj tekst wejściowy
    inputs = tokenizer(text, return_tensors="pt", padding=True, truncation=True)

    # Wygeneruj poprawiony tekst
    with torch.no_grad():
        outputs = model.generate(**inputs, max_new_tokens=len(text) + 10)

    # Dekoduj wygenerowany tekst
    corrected_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return corrected_text

💡
if __name__ == "__main__":
    text = sys.argv[1]
    corrected_text = correct_text(text)
    print(f"{corrected_text}")
```

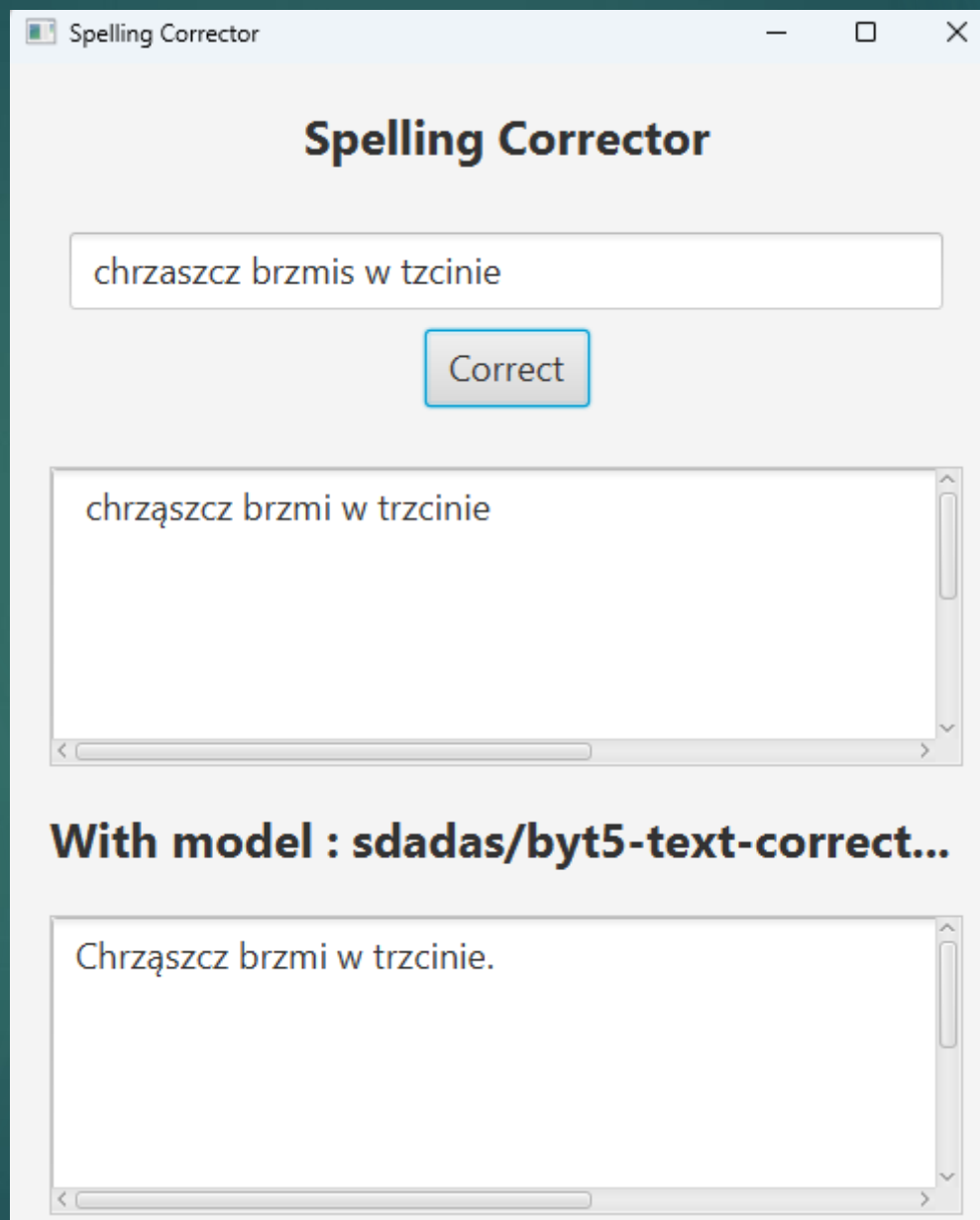
Plik wykonywalny modelu



The screenshot shows an IDE with a project named 'spell-corrector'. The left sidebar displays the project structure, including folders like '.idea', 'src', 'dist', 'spell', 'main', and 'target'. The 'spell.exe' file in the 'spell' folder is highlighted with a red box. The main editor displays the 'SpellChecker.java' file, which contains the following code:

```
1 package pl.github.korbecki.krzysciak;
2
3 import java.io.*;
4
5 1 usage
6 public class SpellChecker {
7     1 usage
8     public String deepLearningCorrectText(String text) {
9         try {
10             String exePath = "src/dist/spell.exe";
11
12             ProcessBuilder processBuilder = new ProcessBuilder( ...command: exePath, "<pl>" + text);
13             processBuilder.redirectErrorStream(true);
14             processBuilder.redirectInput(ProcessBuilder.Redirect.PIPE);
15
16             Process process = processBuilder.start();
17
18             BufferedReader reader = new BufferedReader(new InputStreamReader(process.getInputStream(), charsetName: "windows-1250"));
19
20             String line;
21             StringBuilder output = new StringBuilder();
22             while ((line = reader.readLine()) != null) {
23                 output.append(line).append("\n");
24             }
25
26             int exitCode = process.waitFor();
27             System.out.println("Kod wyjścia: " + exitCode);
28             System.out.println("Poprawiony tekst:\n" + output);
29             return output.toString();
30         } catch (IOException | InterruptedException e) {
31             e.printStackTrace();
32             return null;
33         }
34     }
```

Rezultat



The screenshot shows a window titled "Spelling Corrector" with a standard Windows title bar (minimize, maximize, close buttons). The window contains the following elements:

- Title:** "Spelling Corrector" in bold black font.
- Input Field:** A text box containing the sentence "chrzasczcz brzmis w tzcinnie".
- Correct Button:** A button labeled "Correct" with a blue border.
- Output Field (Before):** A text box containing the sentence "chrząszcz brzmi w trzcinie".
- Model Information:** Text below the output field: "With model : sdadas/byt5-text-correct...".
- Output Field (After):** A text box containing the sentence "Chrząszcz brzmi w trzcinie.".

The application demonstrates the correction of misspellings in a Polish sentence using a specific model.

Źródła

- ▶ © Narodowy Korpus Języka Polskiego 2008-2012
<https://nkjp.pl/index.php?page=14&lang=0>
- ▶ How to Write a Spelling Corrector (norvig.com)
- ▶ <https://huggingface.co/sdadas/byt5-text-correction>