Statystyczna analiza tekstu serii powieści 'Harry Potter'

May 30, 2020

Szymon Sroka 141312

1 Skąd pochodzi mój zbiór danych?

Zbiór danych, który poniżej będę analizował, wygenerowałem przy pomocy Pythona.

Najpierw załadowałem do programu pliki tekstowe z zawartością poszczególnych części "Harry'ego Pottera", podzieliłem tekst na zdania i słowa, a następnie dokonałem analizy takich cech jak ilość słów i ilość znaków w poszczególnych książkach, wydźwięk i obiektywność zdań (przy pomocy pythonowego NLTK) czy częstość występowania słów kluczowych dla tej serii powieści.

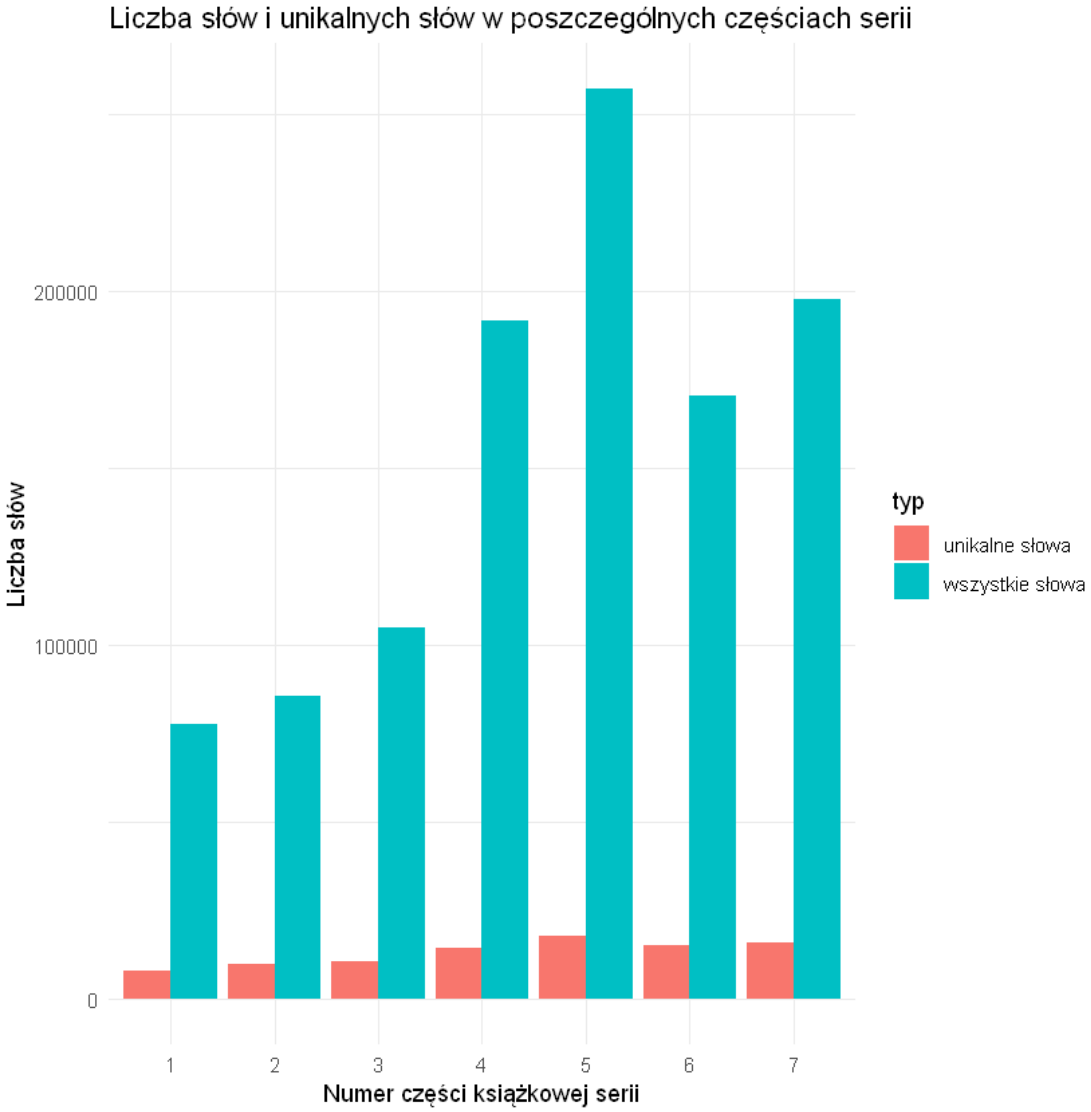
Obliczyłem także indeks czytelności Flescha oraz przeanalizowałem wzajemne położenie słów - na przykład czy słowo 'Hermione' pojawiało się statystycznie bliżej słowa 'Harry' niż 'Ron'. Uzyskane dane przeniosłem do arkusza excelowskiego.

Plik daneliczbowe.csv zawiera podstawowe informacje dotyczące liczby wystąpień słów i znakow, z podziałem na serie, obliczone indeksy czytelności i informacje pochodzące z analizy przy użyciu NLTK. Pozostałe pliki zawierają: najczęściej używane przymiotniki w 1. i 7 części serii (czestotl_przym.csv) i liczbę ich powtórzeń, listę najczęściej używanych używanych słów w całej serii i liczbę ich powtórzeń (czestotl_calosc.csv), liczbę samogłosek w każdym zdaniu serii (samogloski.csv) oraz infrmacje o liczbie liter w zdaniach, ich wydźwięku i obiektywności (dane2.csv).

Do analizy użyłem tekstu w języku angielskim, aby móc użyć wspomnianych wyżej bibliotek przetwarzania języka naturalnego, które w większości przypadków są dostępne tylko w języku angielskim.

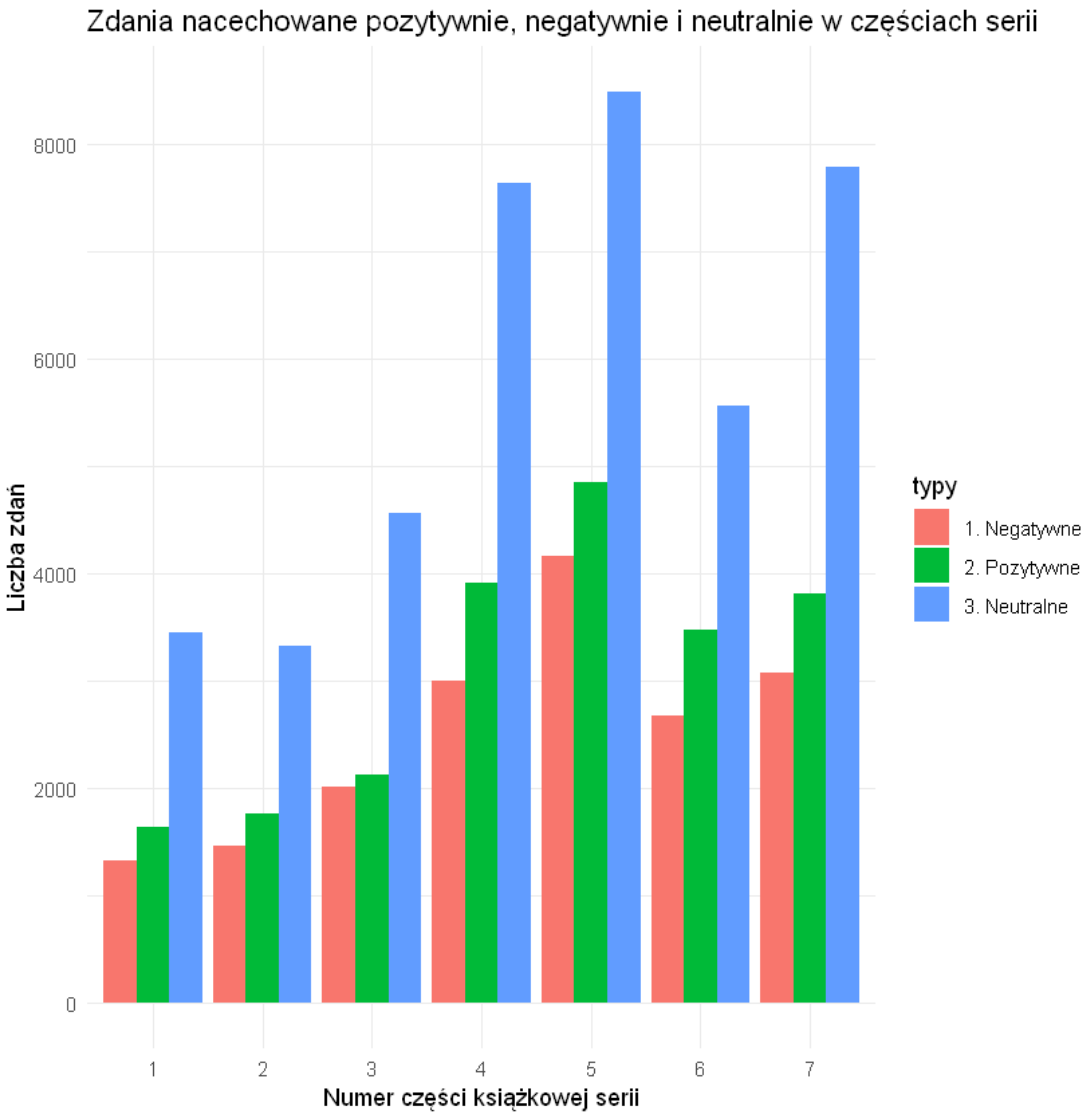
	wykrzyknienia	zn_zapytania	kropki	slowa	znaki	unikalne_slowa	harry
1	474	754	6136	77628	440684	7996	1326
2	535	704	6725	85675	487210	9899	1646
3	1012	1025	9445	104996	610035	10487	2004
4	1461	1746	12458	191632	1099604	14277	3171
5	1567	2595	19115	257306	1494678	17840	4109
6	1066	1760	12072	170679	978749	15233	2795
7	1540	2124	14067	197948	1127105	15990	3146
ALL	7655	10708	80018	1085863	6238087	38806	18197

2 Wykresy, chmury słowne, szeregi rozdzielcze i ich analizy

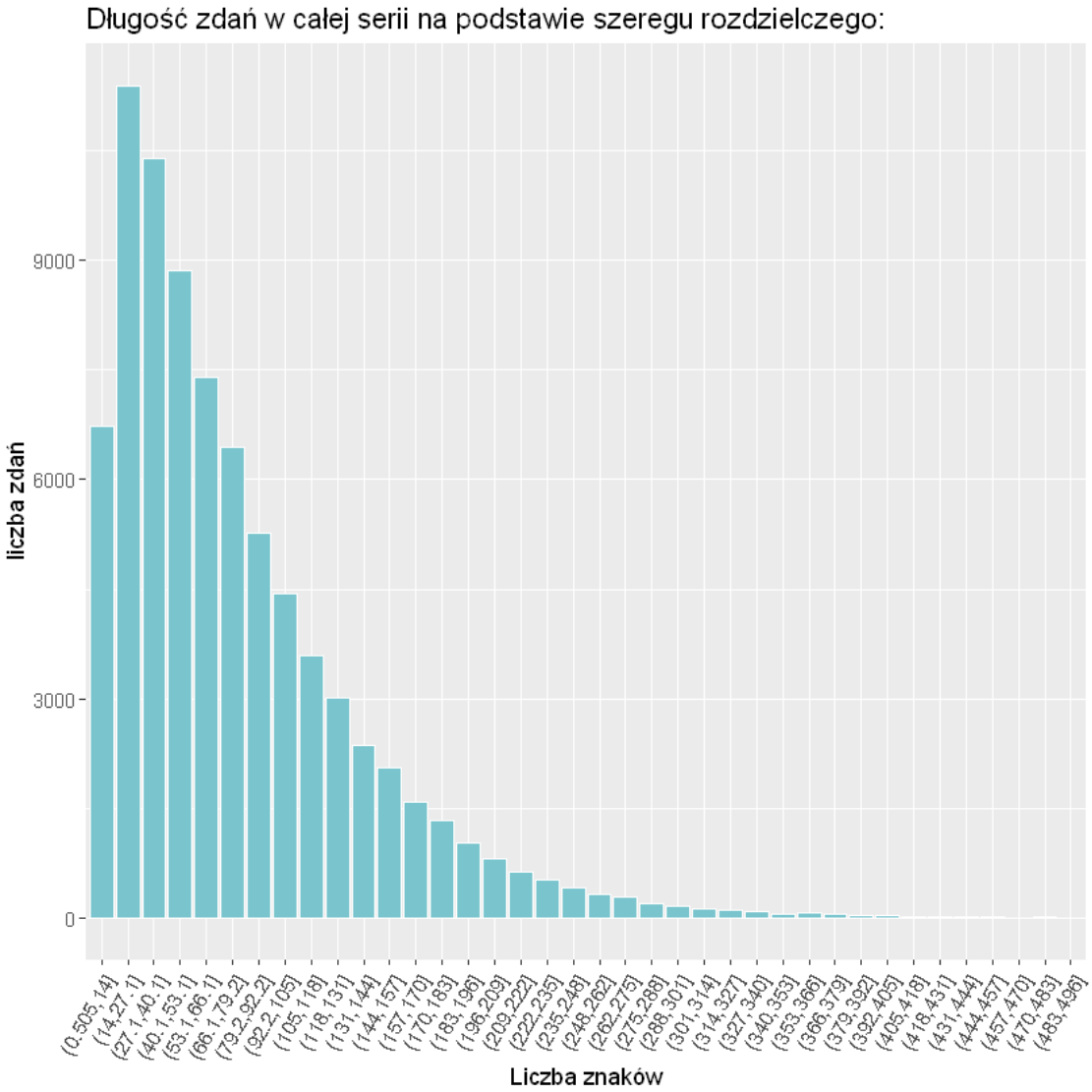


Dane statystyczne liczby **wszystkich i unikalnych słów** w z podziałem na serie:

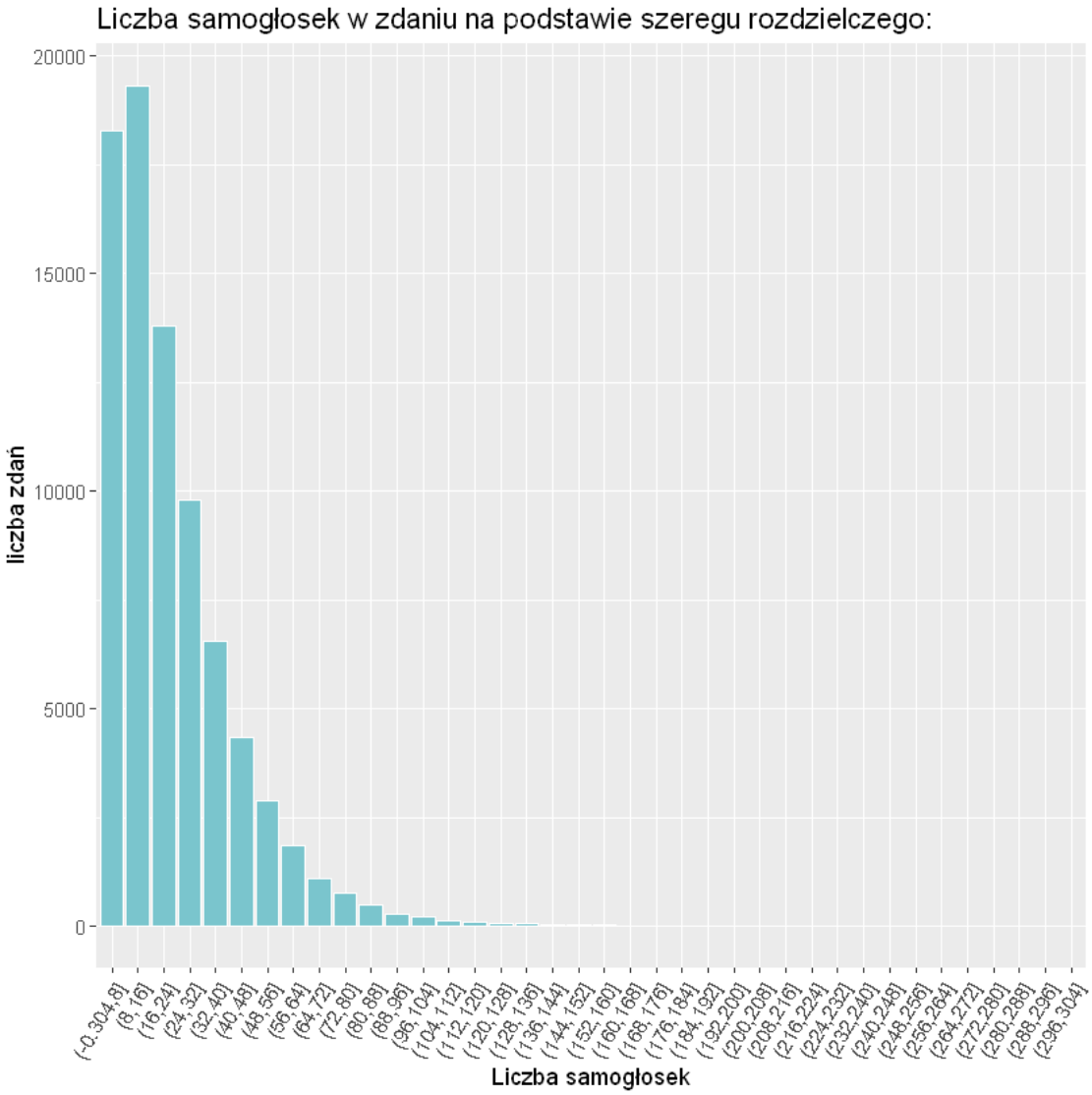
Nazwa_badanej_wielkości	Wszystkie_słowa	Unikalne_słowa
Średnia arytmetyczna	155123.4285714	13103.1428571
Wariancja	4531909697.9523811	13313272.4761905
Odchylenie standardowe	67319.4600242	3648.7357367
Rozstęp	1008235.0000000	30810.0000000
Mediana	170679.0000000	14277.0000000
Skośność	0.1271606	-0.1246549
Kurtoza	-1.7489422	-1.8490797



Poniżej prezentuję szeregi rozdzielcze dla długości zdania (liczba znaków włącznie ze znakami interpunkcyjnymi i spacjami) oraz liczby samogłosek w zdaniach. Zdania dłuższe niż 500 znaków stanowiły zaledwie 0.0009 wszystkich zdań, dlatego dla przejrzystości obliczeń i rysowania odrzuciłem je jako błąd, który mógł wystąpić podczas podziału tekstu na zdania w Pythonie.

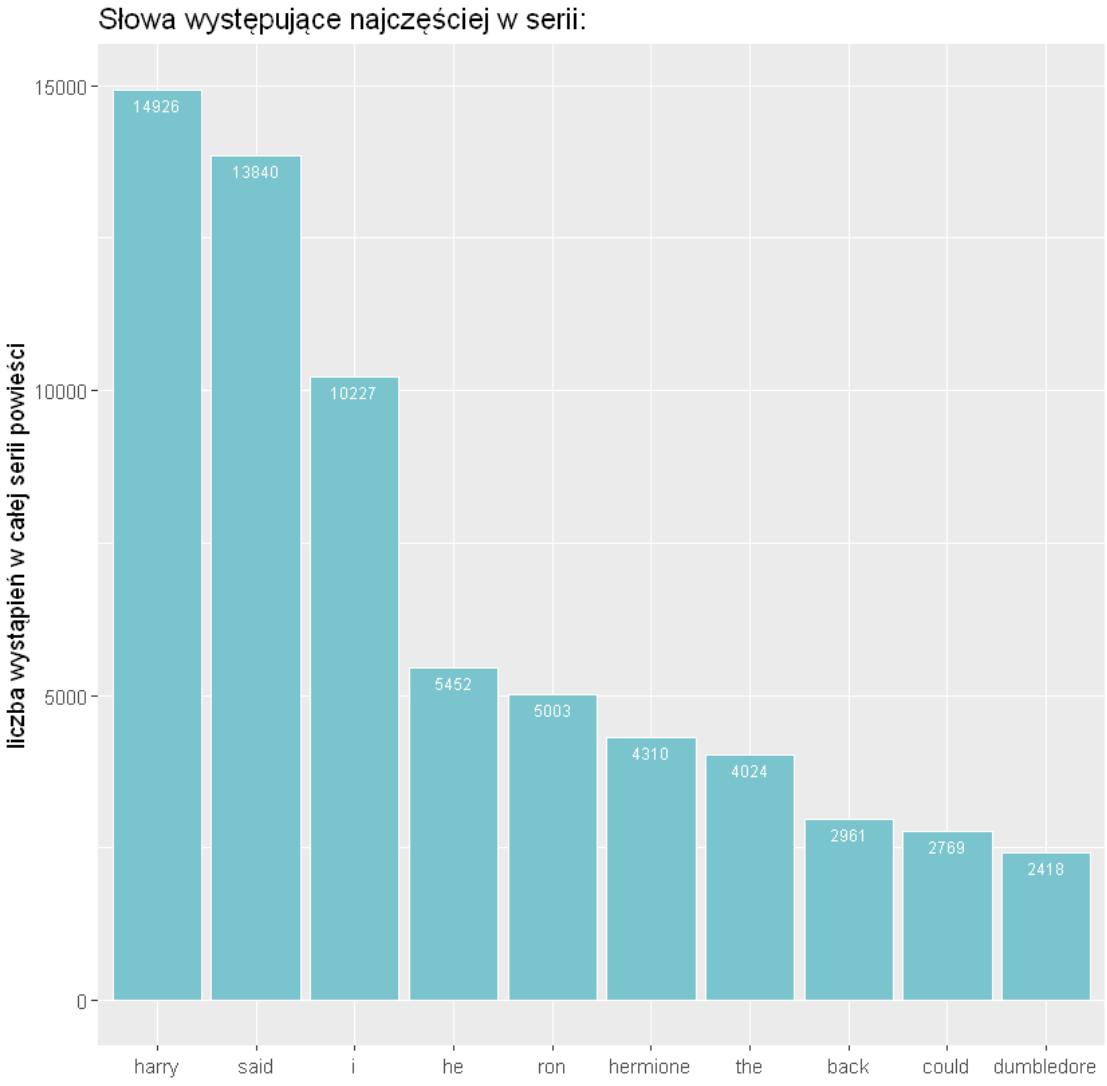


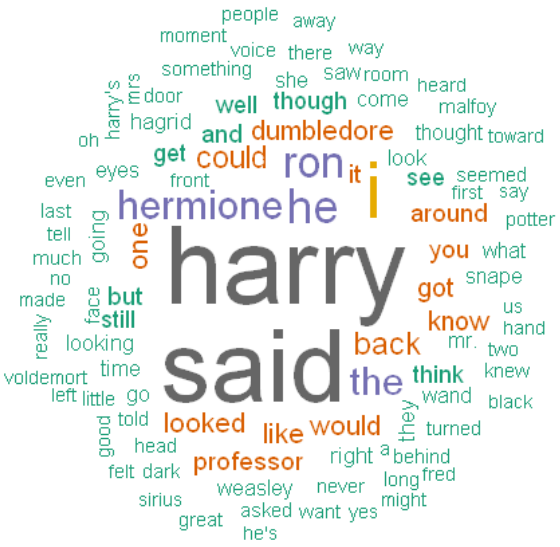
Nazwa	Wartość.dla.długości.zdań
Średnia arytmetyczna	76.06903
Wariancja	4261.83528
Odchylenie standardowe	65.28273
Rozstęp	1031.00000
Mediana	58.00000



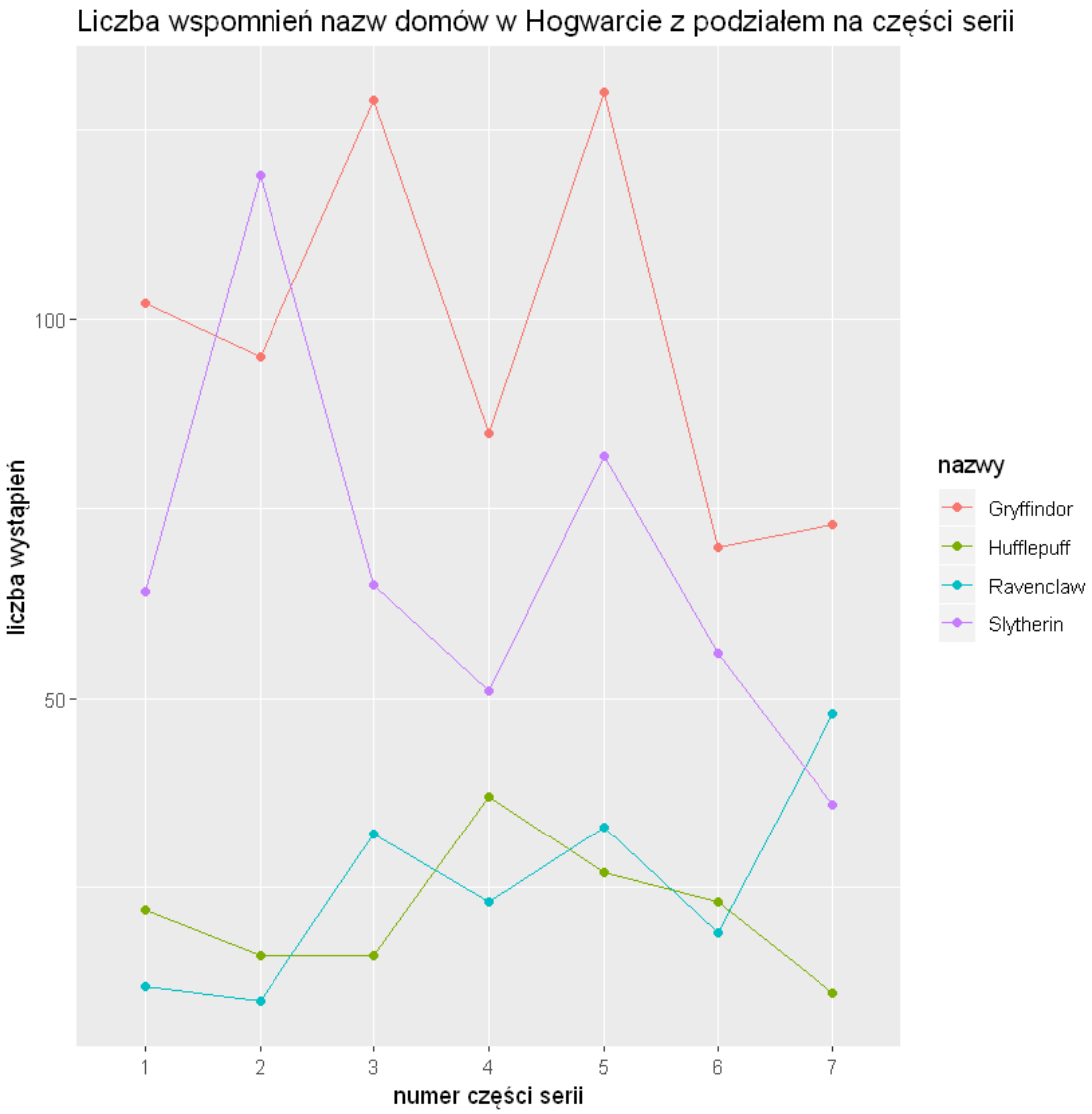
Nazwa	Wartość.dla.liczby.samogłosek
Średnia arytmetyczna	23.17921
Wariancja	394.98697
Odchylenie standardowe	19.87428
Rozstęp	304.00000
Mediana	18.00000

Poniżej przedstawione jest 10 najczęściej występujących wyrazów w całej serii, a w'chmurze słownej' przedstawione jest 100 najczęściej występujących słów z wielkością słowa proporcjonalną do liczby wystąpień.





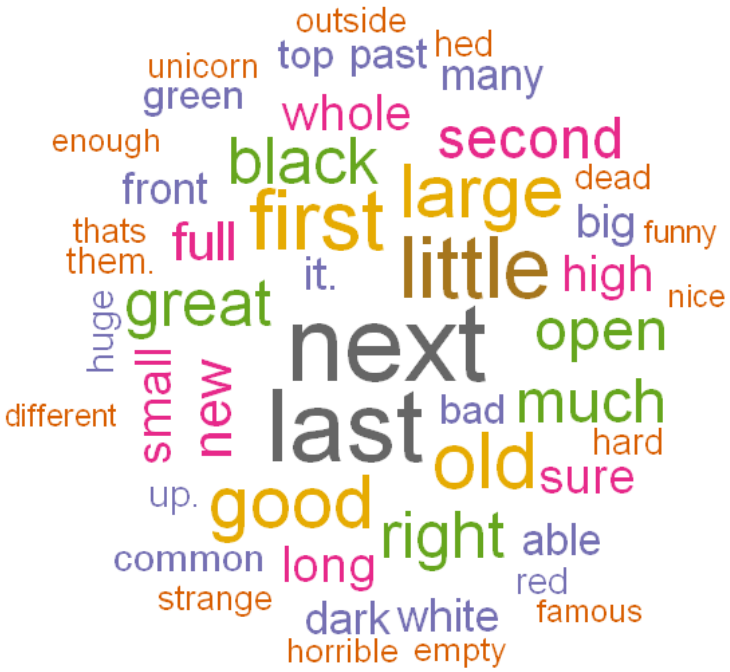
Na uwagę zasługuje nieco zaskakujący fakt, że słowo *Ron* pojawiło się w książce o prawie 700 razy więcej niż *Hermione*. Poniżej zamieszczam wykres prezentujący liczbę wspomnień nazw domów w Hogwarcie z podziałem na części serii:



Spośród mniej oczywistych wniosków płynących z analizy powyższego wykresu można wymienić relatywnie dużą ilość powtórzeń słowa 'Slytherin' w drugiej części serii. Jest to spowodowane faktem, że centrum akcji tej serii znajduje się Komnata Tajemnic, która mogła być otwarta jedynie przez Dziedzica Slytherina, stąd bardzo często pojawiał się on w treści powieści. Można również zauważyć, że słowo 'Ravenclav' w ostatniej częsci serii pojawiało się znacznie częściej niż w poprzednich - ma to związek z diademem Roweny Ravenclaw, który był jednym z horkruksów, poszukiwanych przez głównych bohaterów w ostatniej części serii.

Autorka w jednym z wywiadów zdradziła, że jej celem było, aby każda kolejna powieść serii była coraz bardziej mroczna, złożona, i dojrzała. Na podstawie porównania najczęściej używanych przymiotników w pierwszej i ostatniej serii sprawdzę, czy istotnie tak jest.

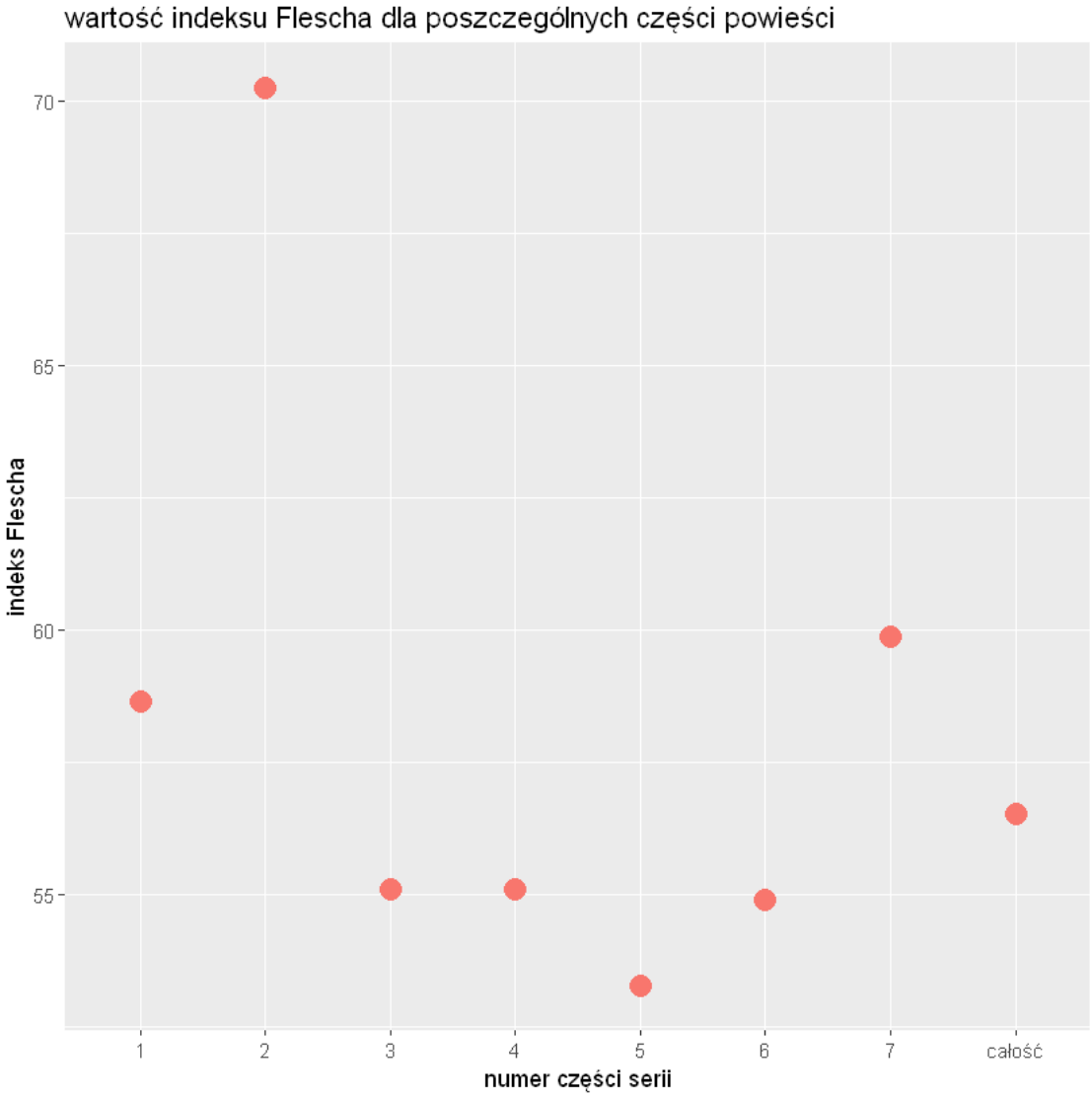
Chmura słów poniżej prezentuje **najczęściej używane przymiotniki w pierwszej części serii.** Rozmiar słowa odpowiada częstości jego występowania w tekście części.



Chmura słów poniżej prezentuje **najczęściej używane przymiotniki w ostatniej, siódmej części serii.**

W pierwszej części istotnie dominują słowa przymiotniki pozytywnie nacechowane: good, right, great, new, funny, top, right, white; natomiast w drugiej części przeważają słowa takie jak: little, last, old, black, small, dark, dead, cold, hard, silent. Stąd można wywnioskować, że rzeczywiście pierwsza część była bajkowa, pogodna, a wraz z rozwojem akcji siódma część stała się tą najbardziej mroczną.

Na koniec przedstawiam wyliczony **wskaźnik czytelności Flescha** dla każdej części serii. Indeks ten określa stopień trudności zrozumienia danego tekstu w języku angielskim. Oblicza się go na podstawie wzoru: 206.835 - 1.015*(liczba_słów/liczba_zdań)-84.6*(liczba_sylab/liczba_słów).



Zakres punktów oraz odpowiadające mu objaśnienie (źródło -> wikipedia.org):

100.00-90.00	Very easy to read. Easily understood by an average 11-year-old student.
90.0-80.0	Easy to read. Conversational English for consumers.
80.0-70.0	Fairly easy to read.
70.0-60.0	Plain English. Easily understood by 13- to 15-year-old students.
60.0-50.0	Fairly difficult to read.
50.0-30.0	Difficult to read.
30.0-0.0	Very difficult to read. Best understood by university graduates.

Im watość indeksu jest niższa, tym tekst jest trudniejszy do zrozumienia.

Wynik w zakresie od 50 do 60 oznacza tekst dość trudny do zrozumienia - na poziomie 12-14 klasy w amerykańskim systemie nauczania i to w jego zakresie mieści się większość części serii powieści oraz wskaźnik dotyczący tekstu całej powieści.

Część 1. i 7. znajduje się blisko wyższego zakresu, od 60 do 70 punktów. Oznacza on, że tekst jest na poziomie 8 i 9 klasy, czyli prostego języka, zrozumiałego przez 13- czy 15-latków.

Druga część powieści znalazła się w w przedziale od 70 do 80 punktów, który oznacza, że tekst jest na poziomie 7. klasy szkoły podstawowej: dość łatwy do zrozumienia.

Z wykresu wynika, że najtrudniejszą w odbiorze jest część piąta ("Harry Potter and the Order of Phoenix"), a najłatwiejszą - druga ("Harry Potter and the chamber of secrets").

3 Testy statystyczne:

1) Chcę sprawdzić (na poziomie istotności = 0.01), **czy zdania nacechowane pozytywnie stanową 1/5 liczby wszyskich zdań w 1. części Harry'ego Pottera, czy może więcej** (liczba zdań nacechowanych pozytywnie ma w przybliżeniu rozkład normalny). W tym celu wylosowałem próbkę n=700 zdań. 176 z nich okazało się nacechowane pozytywnie.

Przyjęcty poziom istotności i wybrany test: alfa=0.01, test Z
Hipoteza zerowa: p=0.2
Hipoteza alternatywna: p>0.2
Obszar krytyczny: (2.326, +inf)
Wartość statystyki testowej: (176-20%*700)/(pierwiastek(700* 20/100 * 80/100))= 3.402

Decyzja: odrzucamy H0 na rzecz H1. Zprzeprowadzonego testu wynika, że zdania nacechowane pozytywnie stanowią więcej niż 20% liczby wszystkich zdań w 1. części Harry'ego Pottera. Nie popełniliśmy błędu ani pierwszego, ani drugiego rodzaju: w rzeczywistości zdania nacechowane pozytywnie stanowią około 27% wszystkich zdań w powieści.

2) Chcę sprawdzić (na poziomie istotności = 0.01), **czy średnia długość słowa w powieści "Harry Potter" jest równa długości słowa "Potter" (6), czy może mniejsza.** W tym celu wylosowałem próbkę n=650 słów i otrzymałem średnią liczbę słów w próbie równą 4.32 przy odchyleniu standardowym równym 2.2.

Przyjęcty poziom istotności i wybrany test: alfa=0.01, test Z
Hipoteza zerowa: mi=6
Hipoteza alternatywna: mi<6
Obszar krytyczny: (-inf, -2.326)
Wartość statystyki testowej: (4.32-6)*pierwiastek(650)/(2.2)= -19.469

Decyzja: odrzucamy H0 na rzecz H1. Z testu wynika, że średnia długość słowa w serii 'Harry Potter' jest mniejsza niż 6 znaków. Nie popełniliśmy błędu ani pierwszego, ani drugiego rodzaju: w rzeczywistości średnia długość słowa wynoi 4.561 znaków przy odchyleniu standardowym = 2.316.

3) Ron i Hermiona to najbliżsi przyjaciele głównego bohatera powieści. Słowo 'Hermione' pojawiło się w maksymalnej odległości 10 słów od słowa 'Harry' w 5% wszystkich wystąpień słowa 'Harry'. Chchę sprawdzić, **czy słowo 'Ron' pojawia się w powieści tak samo często, jak słowo 'Hermiona', czy może częściej.** W tym celu wylosowałem 200 wystąpień słowa 'Harry': słowo 'Ron' pojawiło się w jego 20-słownym otoczeniu w 21 przypadkach. Badana przeze mnie cecha ma w przybliżeniu rozkład normalny.

Przyjęcty poziom istotności i wybrany test: alfa=0.01, test Z
Hipoteza zerowa: p=0.05
Hipoteza alternatywna: p>0.05
Obszar krytyczny: (2.326, +inf)
Wartość statystyki testowej: (21-5%*200)/(pierwiastek(200* 5/100 * 95/100))= 3.5

Decyzja: znów odrzucamy H0 na rzecz H1. Z przeprowadzonego testu wynika, że słowo 'Ron' pojawia się częściej w otoczeniu 'Harry' niż 'Hermione'. Nie popełniliśmy błędu ani pierwszego, ani drugiego rodzaju: w rzeczywistości słowo 'Ron' pojawia się w 20-słownym otoczeniu słowa 'Harry' w około 7.5% wystąpień 'Harry'.

4 Podsumowanie:

- Wybrałem taki temat, ponieważ chciałem przetestować wiedzę zdobytą podczas zajęć z SiAD w kontekście praktycznym i jednocześnie zgłębić nieco temat przetwarzania języka naturalnego - ten aspekt informatyki bardzo mnie zaciekawił, a wcześniej nie miałem z nim styczności. Jestem także fanem Harrego Pottera, co sprawiło, że praca, której efekty widać powyżej, była ciekawym projektem, a nie jedynie uczelnianym obowiązkiem.
- Wykonując projekt nauczyłem się, jak za pomocą narzędzi statystycznych wyodrębnić z literatury te aspekty, których nie sposób poznać porzez zwykłe czytanie powieści czy nawet jej analizę literacką, takie jak porównywanie liczb wystąpień różnych wyrazów, badanie ich ładunku emocjonalnego na dużą skalę czy określanie wzajemnego położenia wyrazów w kontekście całego tekstu.
- Przygotowanie tekstu do analizy wymagało ode mnie wykonania wielu działań (takich jak zastąpienie wielokrotnych spacji pojedynczymi, usunięcie dywizów, ujednolicenie cudzysłowów), stąd zdaję sobie sprawę, że uzyskane przeze mnie dane mogą być przybliżone, a nie dokładne. Ponadto tekst, na którym pracowałem, został otrzymany w wyniku optycznego rozpoznawania tekstu, co naraża go na powstawanie literówek lub niepożądanych znaków. Nie znalazłem ich wiele, a te, które się pojawiły, strałem się wyeliminować.
- Temat statystycznej analizy tekstu jest bardzo szeroki i pozostawia bardzo dużą dowolność w wyborze danych do analizy. Nawet w samym "Harrym Potterze" pozostało jeszcze bardzo wiele wątków, które można by przeanalizować.
- Projekt zachęcił mnie to stosowania narzędzi statystycznych w realnych zagadnieniach i do dokładniejszego poznania narzędzi służących analizie języka naturalnego.