Arkadiusz Kuźma,   
Maciej Wróbel  
****   
  
  
**Projekt**

**Wykład monograficzny**

Rzeszów, rok 2024

Spis treści

[1. Wstęp 3](#_Toc168851040)

[2. Kluczowe prawa 4](#_Toc168851041)

[3. Aplikacja 6](#_Toc168851042)

[4. Przeprowadzone eksperymenty 8](#_Toc168851043)

[5. Eksperymenty 9](#_Toc168851044)

[6. Wnioski 18](#_Toc168851045)

1. Wstęp

Wiele zjawisk we współczesnym świecie zachodzi nierównomiernie. Oznacza to, że niektóre elementy pojawiają się znacznie częściej niż inne. Można to zaobserwować w różnych dziedzinach życia, od natury, przez ekonomię, aż po społeczeństwo. Na przykład w przyrodzie, niektóre gatunki zwierząt czy roślin są bardziej powszechne, podczas gdy inne występują sporadycznie. W ekonomii, niektóre firmy dominują na rynku, posiadając znaczną część udziału w rynku, podczas gdy inne, mniejsze przedsiębiorstwa, walczą o przetrwanie i starają się znaleźć swoje miejsce. Podobnie w społeczeństwie, niektóre opinie i poglądy są szeroko rozpowszechnione i akceptowane, podczas gdy inne są mniej popularne i czasami marginalizowane.

Jednakże, nie oznacza to, że elementy rzadziej występujące są bez znaczenia. Wręcz przeciwnie, te mniej powszechne komponenty mają istotny wpływ na kształtowanie   
i funkcjonowanie różnych zjawisk. W ekosystemach, rzadkie gatunki mogą pełnić kluczowe role w utrzymaniu równowagi biologicznej, działając jako wskaźniki zdrowia ekosystemu lub zapewniając niezbędne usługi ekosystemowe, takie jak zapylanie czy kontrola populacji szkodników. W gospodarce, małe i średnie przedsiębiorstwa mogą wprowadzać innowacje   
i tworzyć miejsca pracy, które są kluczowe dla zdrowia ekonomicznego, często adaptując się szybciej do zmian rynkowych i technologicznych niż ich więksi konkurenci. W społeczeństwie, mniej popularne opinie mogą wywoływać istotne zmiany społeczne i kulturowe, rozpoczynając dyskusje i reformy, które napędzają postęp oraz pogłębiają zrozumienie różnorodnych perspektyw.

Zarówno elementy częste, jak i te rzadkie, współdziałają ze sobą, tworząc kompleksowy i zróżnicowany system. Te różnorodne składniki, choć mogą wydawać się nieistotne na pierwszy rzut oka, dopełniają całość, zapewniając dynamikę i elastyczność   
w odpowiedzi na zmieniające się warunki. Różnorodność w przyrodzie, gospodarce   
i społeczeństwie jest fundamentem stabilności i innowacyjności, umożliwiając adaptację   
i rozwój w obliczu wyzwań i zmian. Wszystko to tworzy sieć wzajemnych powiązań, gdzie każdy komponent, bez względu na swoją częstość występowania ma znaczenie. W ten sposób, rozumienie i docenienie roli zarówno powszechnych, jak i rzadkich elementów, jest kluczowe dla holistycznego pojmowania i zarządzania złożonymi systemami współczesnego świata.

1. Kluczowe prawa

Naukowcy od dawna badają nierównomierny rozkład różnych zjawisk, co doprowadziło do sformułowania kluczowych praw, które mają istotne znaczenie po dziś dzień. Badania te obejmują analizę częstotliwości występowania zdarzeń w przyrodzie, ekonomii, a także   
w codziennym życiu społeczeństw. Odkrycia te nie tylko pomagają w zrozumieniu mechanizmów rządzących naszym światem, ale również w przewidywaniu przyszłych trendów i reagowaniu na zmieniające się warunki. Dzięki nim, możemy lepiej zarządzać zasobami, planować rozwój oraz tworzyć bardziej efektywne strategie działania, co ma fundamentalne znaczenie dla postępu i stabilności w dynamicznie zmieniającym się środowisku.

Pierwszym z tych kluczowych odkryć jest Prawo Pareto, które opisuje zjawisko nierównomiernego rozkładu wpływu różnych czynników na dane zjawisko. Vilfredo Pareto zauważył, że często około 80% efektów wynika z działania zaledwie 20% przyczyn, co implikuje nierównomierny rozkład wpływu poszczególnych czynników na końcowy rezultat. Oznacza to, że większość skutków pochodzi z niewielkiej liczby przyczyn, podczas gdy większość przyczyn ma minimalny wpływ na końcowy wynik. To prawo obserwuje się w różnych dziedzinach, od ekonomii, przez biznes, po życie społeczne. Na przykład w biznesie, około 20% klientów generuje aż 80% przychodów firmy, podczas gdy w sferze ekonomicznej, 20% najbogatszych osób kontroluje 80% majątku narodowego. To istotne dla zarządzania zasobami i alokacji środków, ponieważ skoncentrowanie się na tych kluczowych 20% może przynieść większe korzyści niż równomierne rozłożenie uwagi na całość.

Kolejnym ważnym twierdzeniem jest Prawo Benforda, które wyjaśnia nierównomierny rozkład pierwszych cyfr w zestawach danych liczbowych. Według tego prawa, pierwsze cyfry liczb w danych zbiorach nie są równomiernie rozłożone, lecz podlegają określonej rozkładowi. W praktyce oznacza to, że cyfra 1 często pojawia się na pierwszej pozycji znacznie częściej niż pozostałe cyfry, takie jak 2, 3 czy 4. Ponadto, im większa cyfra, tym rzadziej występuje na pierwszej pozycji. Na przykład, cyfry od 5 do 9 są mniej częste na pierwszej pozycji niż cyfry od 1 do 4. To zjawisko jest istotne, ponieważ odchylenia od tego rozkładu mogą wskazywać na potencjalne nieprawidłowości w danych liczbowych, co czyni analizę pierwszej cyfry użytecznym narzędziem do wykrywania anomali w danych.

Oprócz już wspomnianych zjawisk, warto także odnieść się do prawa Zipfa, które bada częstość występowania słów w tekście. To twierdzenie podkreśla istotę elementów występujących rzadziej. Według tego prawa, częstość występowania słowa w tekście maleje wraz z jego pozycją w rankingu popularności słów w tym tekście. Innymi słowy, im wyższa pozycja w rankingu, tym mniejsza częstość występowania słowa.

Prawo Zipfa, pierwotnie skoncentrowane na językach naturalnych, analizuje nierównomierny rozkład częstości występowania słów. Podczas tworzenia listy słów według ich częstotliwości, zazwyczaj na szczycie znajdują się te najczęściej używane, jak na przykład w języku angielskim słowo "the". Mimo że są one powszechne, zajmują niższe pozycje   
w rankingu z uwagi na ich częstotliwość. Z drugiej strony, mniej powszechne słowa, takie jak "delicious" czy "mesmerizing", mogą zajmować wyższe pozycje, gdyż są rzadsze. Prawo Zipfa opisuje tę zależność jako odwrotnie proporcjonalną do częstotliwości i rangi. Na przykład, w korpusie tekstu w języku angielskim, "the" może stanowić nawet 7% wszystkich słów, "of" około 3,5%, a "a" około 1,75%. Pierwsze 135 słów mogą obejmować nawet połowę całego tekstu.

Opisane zjawiska, takie jak Prawo Pareto, Prawo Benforda oraz Prawo Zipfa, mają znaczące zastosowanie w praktyce. Prawo Pareto jest powszechnie stosowane w ekonomii, biznesie oraz analizie społecznej do identyfikacji kluczowych obszarów skupienia uwagi. Dzięki niemu można efektywniej alokować zasoby oraz planować strategie działania, koncentrując się na tych obszarach, które generują największe rezultaty.

Prawo Benforda, jest niezwykle przydatne w audytach finansowych, analizach ekonomicznych oraz badaniach podatkowych. Dzięki niemu można wykrywać nieprawidłowości w danych liczbowych oraz analizować ich autentyczność, co ma istotne znaczenie dla wiarygodności finansowej, planowania podatkowego oraz monitorowania danych statystycznych.

Prawo Zipfa jest szczególnie istotne w dziedzinie językoznawstwa i lingwistyki, gdzie stosuje się je do badania struktury języka oraz analizy tekstu. Pierwsze badania oparte na tym prawie były przeprowadzone na języku angielskim, jednakże z czasem odkrycia te zostały zastosowane także w innych językach i dziedzinach nauki. Prawo Zipfa ma także zastosowanie w dziedzinie informatyki, szczególnie w analizie danych tekstowych, takich jak dokumenty, artykuły czy strony internetowe. Wykorzystuje się je również w modelowaniu dystrybucji słów w różnych zbiorach tekstowych oraz do oceny różnorodności językowej w tekstach literackich, naukowych czy publicystycznych.

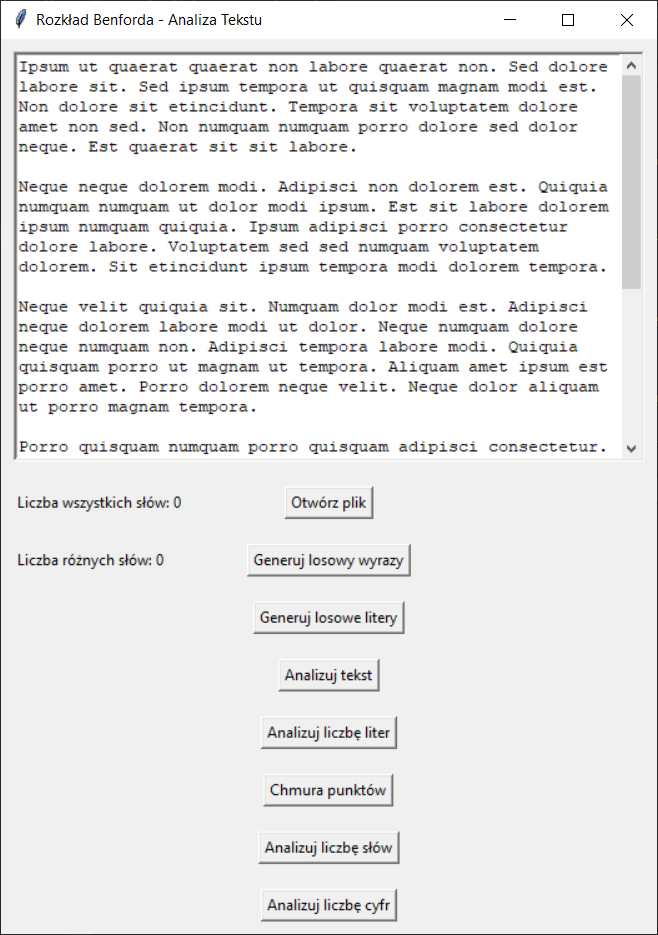
1. Aplikacja

W ramach projektu powstała interaktywna aplikacja okienkowa, która umożliwia analizę struktury tekstu w sposób wygodny i intuicyjny. Jedną z zalet tego programu jest fakt, że użytkownik może dodawać swoje dane bezpośrednio z plików tekstowych. Oprócz tego istnieje możliwość ręcznego wprowadzenia tekstu przez użytkownika To pozwala na elastyczne dostosowanie analizy do konkretnych potrzeb oraz na bieżące aktualizowanie danych, co zwiększa użyteczność i efektywność aplikacji. Dodatkowo, interfejs aplikacji został zaprojektowany w taki sposób, aby umożliwić łatwe manipulowanie danymi oraz szybki dostęp do różnych funkcji analizy tekstu.

W aplikacji można zobaczyć, ile słów i liter jest przetwarzanych i branych do analizy. Dodatkowo, użytkownik ma możliwość sprawdzenia innych statystyk. W aplikacji wyświetla się też liczba unikalnych słów i liter, co pozwala użytkownikowi na szybkie zrozumienie różnorodności i bogactwa językowego tekstu oraz na lepszą analizę jego zawartości. Aby przedstawić wyniki analizy danych dostarczonych przez użytkownika, można wyświetlić histogramy, które pokazują częstotliwość występowania poszczególnych wyrazów lub liter. Dzięki tej formie prezentacji użytkownik może szybko i łatwo zobaczyć najczęściej występujące elementy w tekście oraz analizować ich częstotliwość w sposób bardziej czytelny i przystępny. Ponadto, analiza różnych fragmentów tekstu pozwala na wykrycie powtarzających się wzorców i trendów, co umożliwia głębsze zrozumienie jego struktury oraz identyfikację kluczowych elementów.

Analizując nierównomierny rozkład niektórych zjawisk, możliwe jest przeprowadzenie analizy częstotliwości występowania poszczególnych cyfr. Aplikacja umożliwi także szczegółowe zobrazowanie, ile razy poszczególne cyfry pojawiają się w badanym tekście. Dzięki temu użytkownik będzie mógł łatwo zidentyfikować wzorce występowania cyfr   
w tekście oraz lepiej zrozumieć charakter nierównomiernego rozkładu tych zjawisk.

W aplikacji będzie opcja generowania losowych danych. Można będzie tworzyć losowe słowa na podstawie tekstu lorem ipsum lub losować litery na określonych pozycjach. Dzięki temu użytkownik będzie mógł wygenerować różnorodne sekwencje znaków. Takie funkcje pozwolą na porównanie sensownych fraz z losowo wygenerowanymi danymi, co umożliwi lepsze zrozumienie sposobu, w jaki dane są przetwarzane i interpretowane.



Rys. 3.1. Okno aplikacji do analizowania danych tekstowych i liczbowych

1. Przeprowadzone eksperymenty

Jak widać na podstawie nierównomiernego rozkładu zjawisk, można badać wiele związków, niekoniecznie sięgając po złożone modele sztucznej inteligencji czy analizy danych. Dzięki prostym narzędziom analitycznym opartym na opisanych wcześniej prawach, możliwe jest wyciąganie zaskakujących wniosków oraz zrozumienie głębszych struktur zachodzących w badanych obszarach. Ten sposób podejścia pozwala na bardziej intuicyjne i skuteczne badanie rzeczywistości, co ma istotne znaczenie zarówno w dziedzinie naukowej, jak   
i praktycznej.

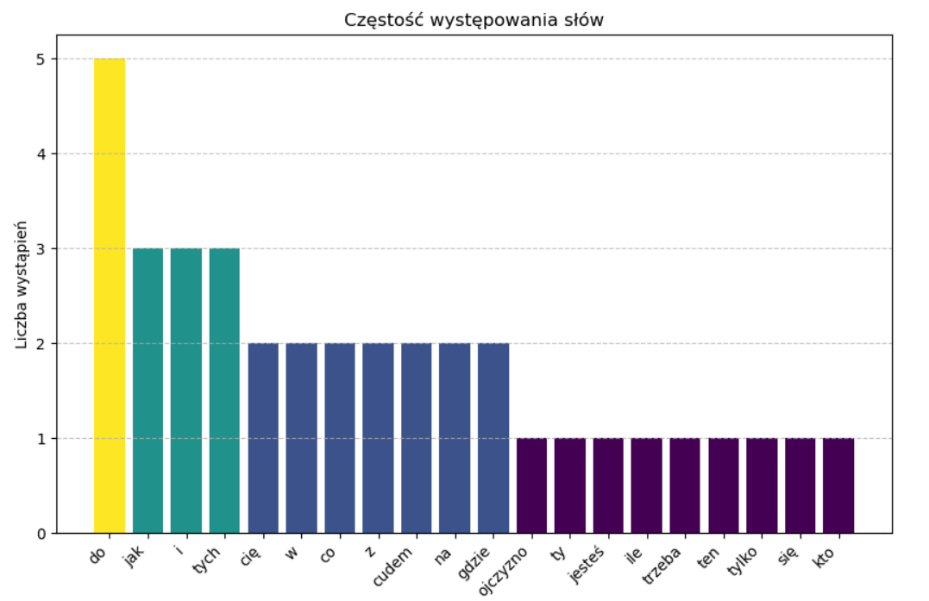
W ramach projektu przeprowadzona zostanie kompleksowa analiza zarówno rzeczywistego tekstu, jak i tekstu wygenerowanego przez program komputerowy. Badania obejmą porównanie liczby słów i liter. Ta analiza pozwoli na dogłębne zrozumienie różnic między oboma tekstami. Dzięki temu możliwe będzie ustalenie, czy na podstawie samych wykresów częstości występowania poszczególnych liter i słów można jednoznacznie odróżnić rzeczywisty tekst od tekstu wygenerowanego przez program komputerowy.

Dodatkowo, w ramach projektu przewidziano analizę zarówno długich, jak i krótkich fragmentów tekstu. Badania te umożliwią kompleksowe zrozumienie struktury danych. Wnioski z analizy różnych długości fragmentów tekstu pozwolą zrozumieć wpływ długości tekstu na badane zjawiska oraz określić, gdzie Prawo Pareto, Prawo Benforda i Prawo Zipfa są najbardziej widoczne.

Na końcu zostanie sprawdzone czy nierównomierny rozkład może także występować   
w danych liczbowych. Planowane jest przeprowadzenie kompleksowej analizy, która obejmie sprawdzenie zgodności z Prawem Benforda. Wnioski przeprowadzone z tego eksperymentu pozwolą na lepsze zrozumienie charakterystyki rozkładu danych liczbowych i ewentualnego występowania nierównomierności. Analiza zgodności z Prawem Benforda umożliwi ocenę, czy badane zbiory danych odpowiadają oczekiwanemu rozkładowi.

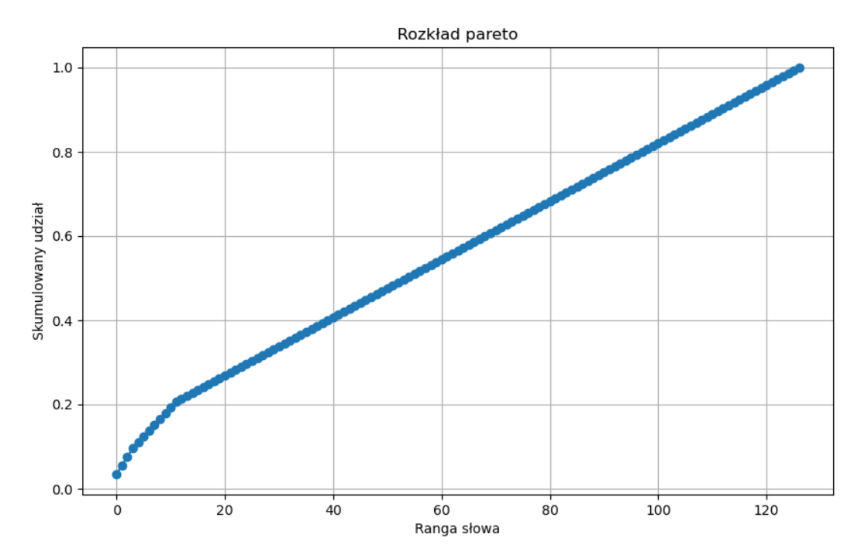
1. Eksperymenty

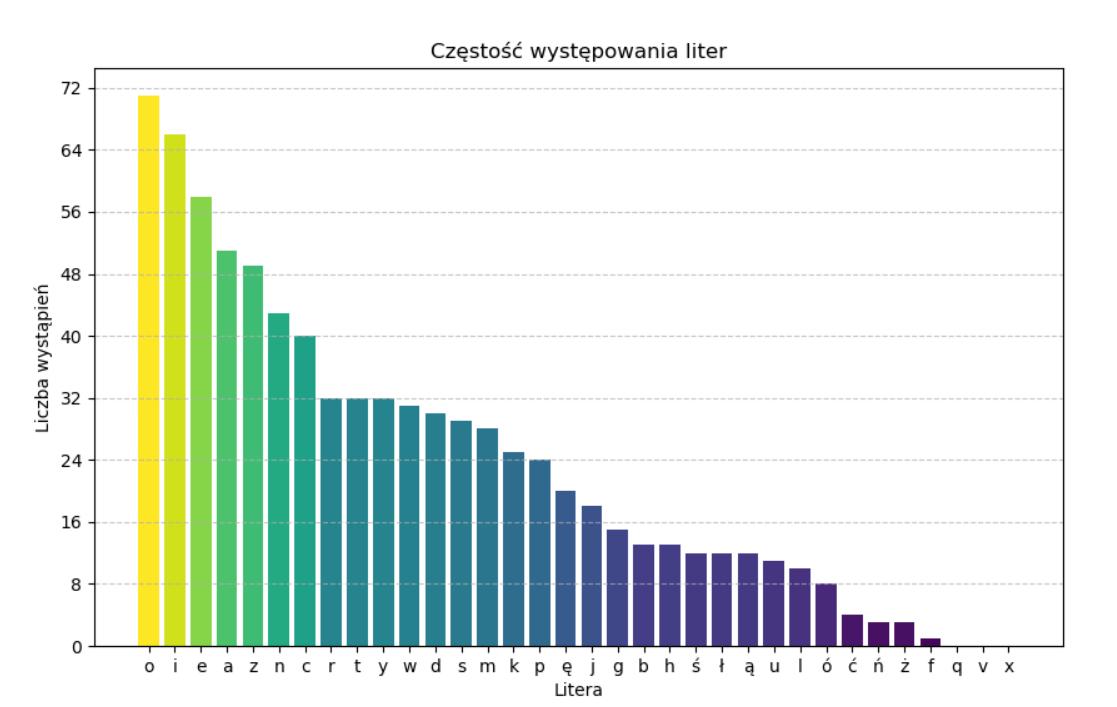
Na początku dokonano analizy częstości występowania liter i znaków w inwokacji "Pana Tadeusza" Adama Mickiewicza. Przeanalizowano krótki fragment, który zawierał 145 słów w sumie, spośród których było 127 różnorodnych. Badania wykazały, że zarówno   
w przypadku wyrazów, jak i liter, można wyodrębnić pewne dominujące elementy.

W analizowanym fragmencie najczęściej występowało słowo „do”, aż 5 raz. Dodatkowo, trzy różne słowa wystąpiły trzykrotnie, a siedem słów pojawiało się dwukrotnie. Te obserwacje wskazują na to, że w analizowanym fragmencie jest niewiele słów, które powtarzają się często. Można z tego wnioskować, że analiza ta potwierdza prawo Zipfa, które mówi o nierównomiernym rozkładzie częstości występowania słów w tekście, ponieważ liczba wystąpień poszczególnych słów maleje wraz z ich rangą w rankingu popularności.

Rys. 5.1. Histogram przedstawiający częstotliwość występowania słów w inwokacji „Pan Tadeusza”

Rys. 5.2. Chmura punktów wygenerowana z inwokacji „Pana Tadeusza”

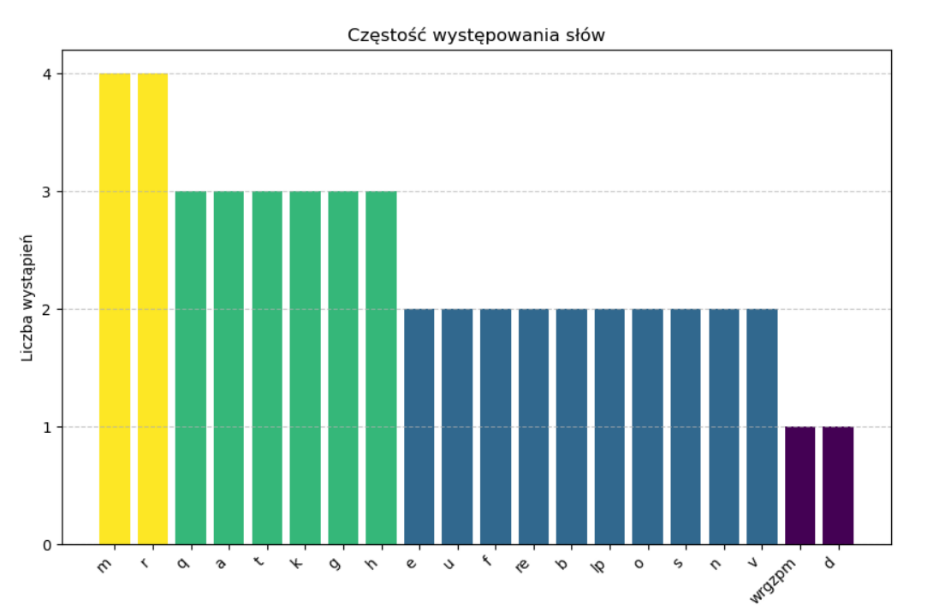
Rys. 5.3. Rozkład Pareto słów występujących w inwokacji „Pana Tadeusza"

W analizie liczby wystąpień liter w tekście można zobaczyć, że kilka liter dominuje nad resztą. Jednak liczba wystąpień każdej kolejnej litery diametralnie maleje w rankingu. Najczęściej występującą literą jest "o". Pojawiającą się on aż 72 razy. Ósma litera w rankingu w tekście występuje już tylko 32 razy. To wyraźnie pokazuje, że liczba wystąpień liter maleje wraz z ich rangą w rankingu. Ta obserwacja potwierdza zasadą nierównomiernego rozkładu, którą opisuje Prawo Zipfa. Twierdzenie to zakłada, że drugi element powinien występować około dwukrotnie rzadziej niż pierwszy, trzeci trzykrotnie rzadziej niż pierwszy, czwarty czterokrotnie rzadziej, i tak dalej. W badanym rozkładzie można zauważyć te proporcje, które odpowiadają zasadzie nierównomiernego rozkładu opisanego w Prawie Zipfa, gdzie liczba wystąpień elementów maleje zgodnie z potęgą ich rangi.

Rys. 5.4. Histogram przedstawiający częstotliwość występowania liter w inwokacji „Pan Tadeusza”

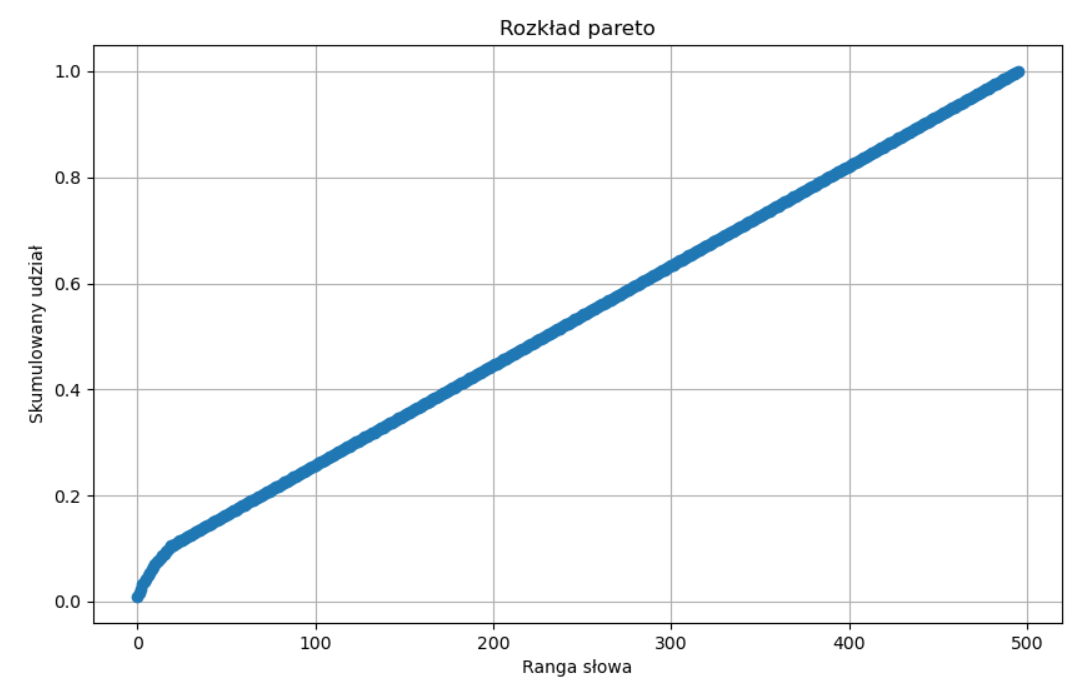
Następnie przeprowadzono badanie dystrybucji liter i słów, wybierając losowo generowane słowa za pomocą programu. Warto zaznaczyć, że w tym przypadku na każdej pozycji losowano literę. Ten eksperyment pozwolił na lepsze zrozumienie zależności między nierównomiernym rozkładem a strukturą tekstu.

Choć w tym przypadku można wyróżnić elementy, które pojawiają się najczęściej, różnica w ich częstotliwości występowania jest znacznie mniejsza niż w przypadku standardowego tekstu. Porównując częstość występowania kolejnych liter i znaków, można stwierdzić, że nie maleją one potęgowo, jak miało to miejsce wcześniej.

Warto również zauważyć, że najczęściej występującymi wyrazami są te jednowyrazowe. Ponadto, w tej analizie występuje więcej składników, które pojawiają się więcej niż jeden raz. W pierwszej analizie było tylko 11 takich składników, natomiast w tym badaniu aż 18 wyrazów występowało więcej niż raz.

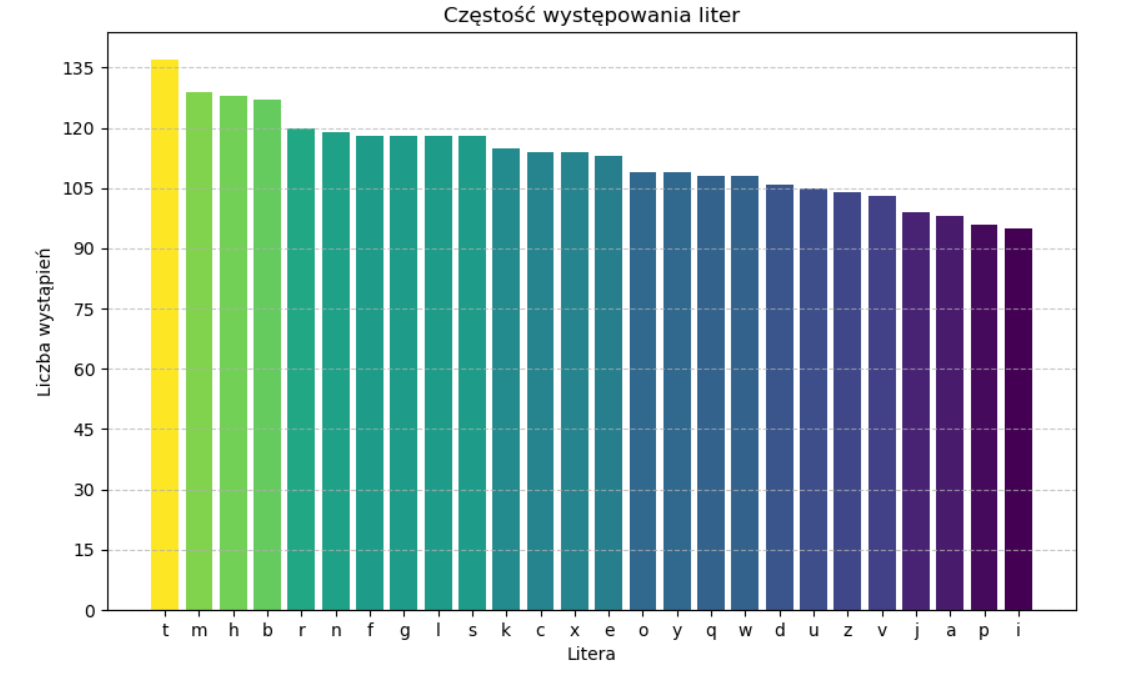
Rys. 5.5. Histogram przedstawiający częstotliwość występowania słów w wygenerowanym tekście

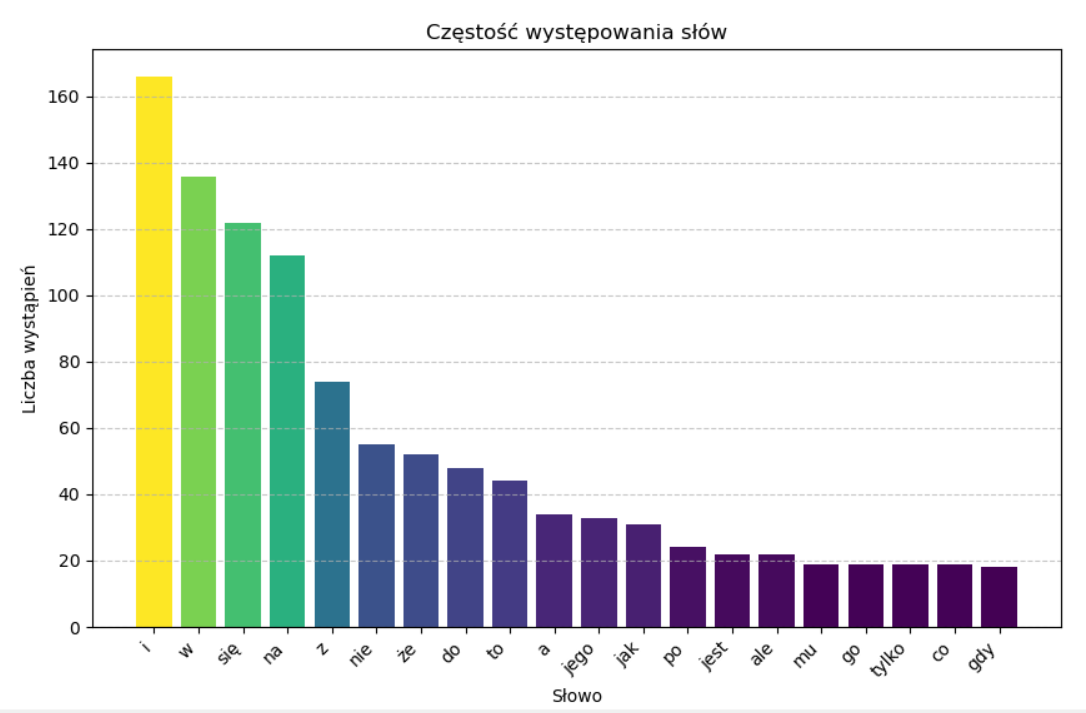
Rys. 5.6. Chmura punktów wygenerowana z wygenerowanego tekstu

  
Rys. 5.7. Rozkład Pareto słów występujących w wygenerowanym tekście

W przypadku losowego generowania liter, ostatnia litera w rankingu, czyli ta najrzadziej występująca, także pojawia się bardzo często. Litera "i", która występuje najrzadziej spośród wszystkich liter, pojawia się ponad 90 razy, co jest zauważalnym odstępstwem od wcześniejszych obserwacji, gdzie litery z ostatnich pozycji występowały rzadko. Ten fakt sugeruje, że losowy proces generowania liter może wprowadzać pewne nieregularności   
w dystrybucji występowania liter.

Przeprowadzone analizy sugerują, że istnieją różnice w porównaniu między zwykłym tekstem a tekstem wygenerowanym przez program. W przypadku tekstu wygenerowanego przez program, nie zawsze obserwuje się zgodność z prawem Zipfa. W takim tekście ilość kolejnych elementów nie maleje tak wyraźnie, a ostatnie składniki mogą występować bardzo często, co jest sprzeczne z oczekiwanym spadkiem częstości występowania zgodnie   
z założeniami prawa Zipfa. Ponadto, liczba elementów, które występują często, jest znacznie większa w tekście wygenerowanym przez program w porównaniu do zwykłego tekstu. To sugeruje, że proces generowania tekstu przez program może wprowadzać pewne nieregularności w częstotliwość występowania poszczególnych słów lub liter.

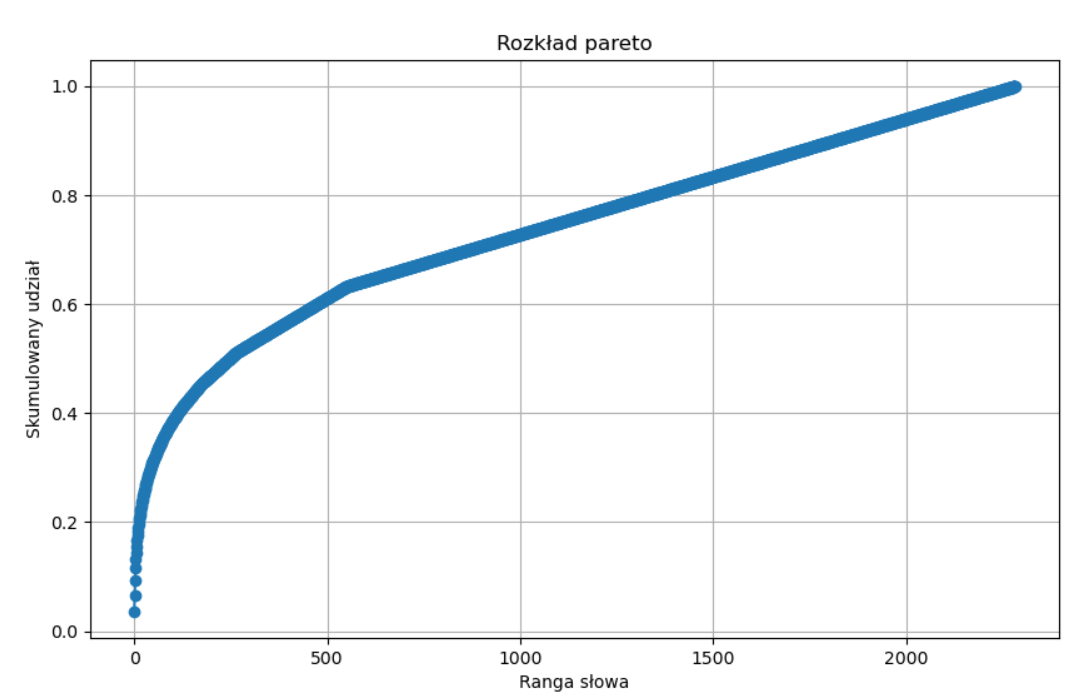
Rys. 5.8. Histogram przedstawiający częstotliwość występowania liter w wygenerowanym tekście 

Następnie przeprowadzono analizę trochę dłuższego tekstu niż w poprzednich przypadkach. Do analizy wybrano tekst krótkiej noweli "Latarnik". W tej książce można wyróżnić 4715 słów. Wraz z długością tekstu zauważalnie wzrasta częstotliwość występowania poszczególnych słów. Wśród dwudziestu najczęściej występujących słów, każde występuje ponad 15 razy. Oznacza to, że niektóre słowa pojawiają się bardzo często, co może mieć istotne znaczenie dla struktury tekstu i jego treści.

Rys. 5.9. Histogram przedstawiający częstotliwość występowania słów w „Latarniku”

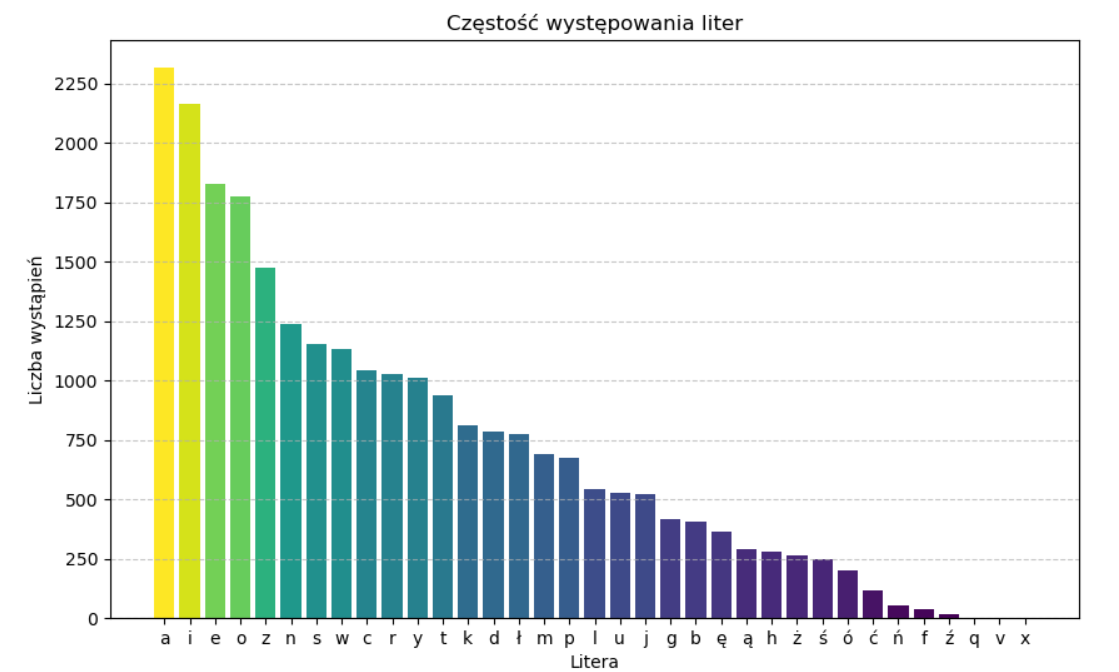


Rys. 5.10. Chmura punktów wygenerowana z „Latarnika”

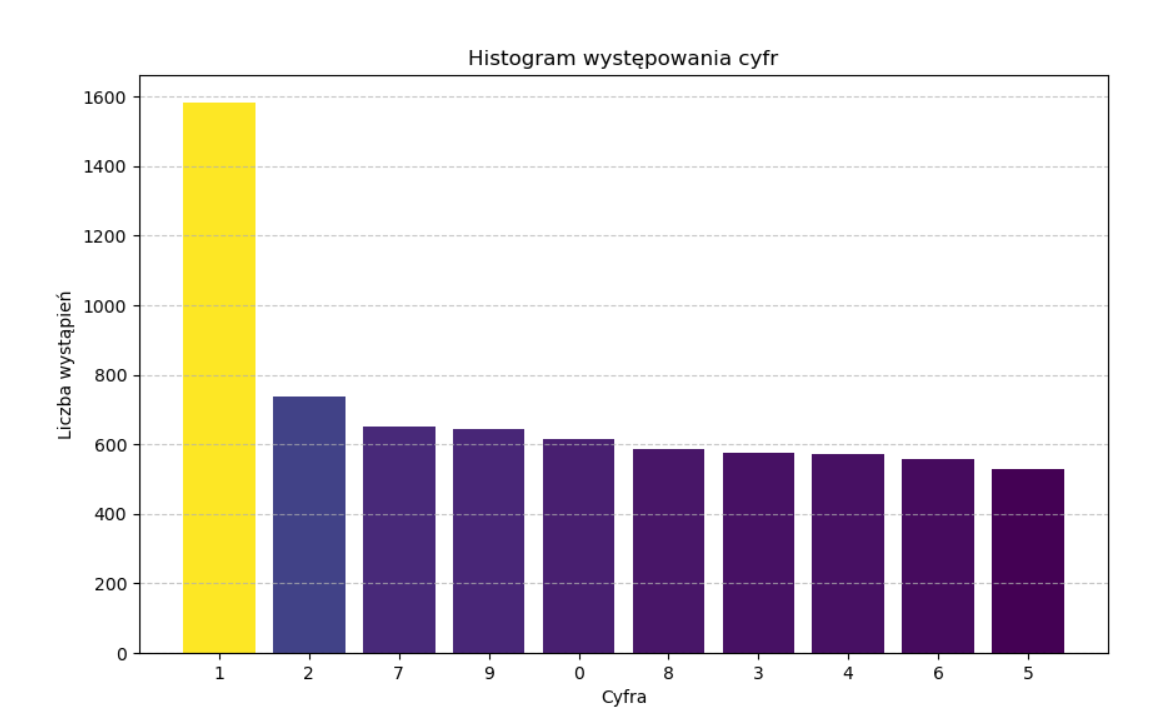
Wykres rozkładu Pareto, opisujący częstość występowania pojedynczych słów, w końcu nie przypomina linii prostej. W poprzednich przypadkach taki kształt wynikał z małej liczby słów, które występowały często, oraz dużej liczby słów, które występowały tylko raz.   
W przypadku tego tekstu istnieje duża liczba słów, które występują więcej niż jeden raz, co powoduje, że wykres nie ma charakteru prostej linii.

Rys. 5.11. Rozkład Pareto słów występujących w „Latarniku”

Wykres częstości liter wydaje się być bardzo podobny do tego, który powstał podczas pierwszego eksperymentu, gdzie analizowano tekst inwokacji "Pana Tadeusza". Co więcej, zauważalne jest, że litery, które często występowały w pierwszym eksperymencie, są także powszechne w tekście "Latarnik", natomiast litery, które występowały rzadziej, pojawiają się również sporadycznie w analizowanym tekście.



Rys. 5.12. Histogram przedstawiający częstotliwość występowania liter w „Latarniku”

Na końcu przeprowadzono analizę, mającą na celu zbadanie, czy w danych liczbowych można zauważyć zjawisko opisane przez prawo Benforda. W ramach tego eksperymentu wykorzystano ceny zamknięcia akcji Google, ze zbioru danych znajdujących się pod linkiem: https://www.kaggle.com/datasets/shreenidhihipparagi/google-stock-prediction/data. Analiza histogramu występowania poszczególnych cyfr w tych danych ujawniła nierównomierny rozkład, gdzie szczególnie liczby 1 i 2 występowały najczęściej. To zjawisko potwierdza tendencję przewidywaną przez prawo Benforda, które mówi o dominacji cyfr początkowych w danych liczbowych o naturalnym pochodzeniu.

Rys. 5.13. Histogram przedstawiający częstotliwość występowania cyfr w cenach zamknięcia cen akcji Google

1. Wnioski

W danych tekstowych i liczbowych widoczna jest zasada nierównomiernego rozkładu, objawiająca się w zróżnicowanych wzorcach występowania poszczególnych elementów. Analiza danych tekstowych wyraźnie pokazuje, że najczęściej występujące słowa znacząco przewyższają liczbę wystąpień pozostałych, co zgadza się z zasadą nierównomiernego rozkładu opisaną przez prawo Zipfa. Natomiast w przypadku danych liczbowych obserwuje się tendencję do występowania cyfr początkowych, co potwierdza przewidywania prawa Benforda. Te reguły są kluczowe dla zrozumienia struktury danych i ich charakterystycznych cech, co odgrywa fundamentalną rolę w analizie i interpretacji danych. Warto podkreślić, że nawet rzadko występujące dane liczbowe lub cyfry mogą zawierać istotne informacje dla analizy, co podkreśla znaczenie uwzględnienia wszystkich elementów danych w procesie analizy.