Radosław Mikołajczyk

Grupa: ZIISN2-2411

Numer albumu: 235530

Kraków, 11.06.2025

PRZETWARZANIE JĘZYKA NATURALNEGO

Analiza wyników przetwarzania tekstów literackich

Prowadzący ćwiczenia: dr Katarzyna Wójcik
Prowadzący wykłady: prof. dr hab. Paweł Lula

Kierunek: Informatyka stosowana

Specjalność: Systemy inteligentne

Studia niestacjonarne drugiego stopnia

Semestr letni 2024/2025

Charakterystyka zbioru danych

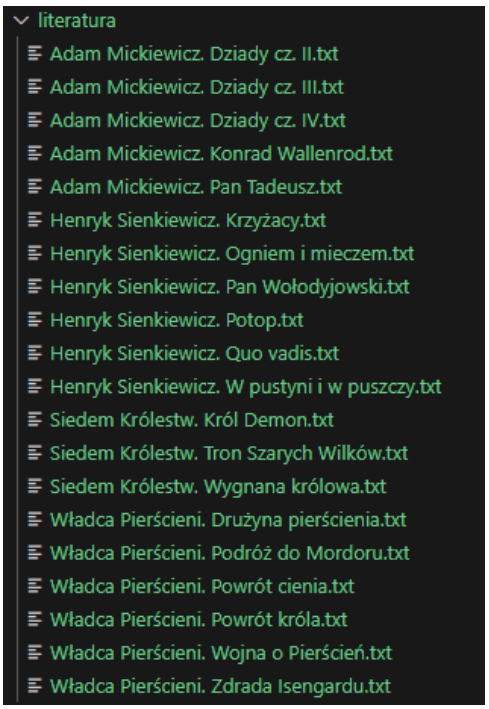
Zbiór danych wykorzystany w niniejszym projekcie składa się z **20 dokumentów tekstowych w języku polskim**, przygotowanych do analizy komputerowej. Każdy dokument reprezentuje jedną książkę i został zapisany w formacie .txt. Zawartość zbioru została celowo dobrana tak, aby objąć **cztery zróżnicowane tematycznie obszary literackie**:

1. **Twórczość Adama Mickiewicza**
   Ta kategoria zawiera 5 utworów, reprezentujących romantyzm polski. Teksty te cechują się podniosłym tonem, obecnością mowy stylizowanej, refleksją filozoficzno-patriotyczną oraz silną symboliką.

2. **Twórczość Henryka Sienkiewicza**
   W tej kategorii znajdują się 6 powieści historycznych i przygodowych autorstwa Henryka Sienkiewicza. Styl tych tekstów jest narracyjny, z wyraźną strukturą fabularną i licznymi odniesieniami do historii Polski i chrześcijaństwa.

3. **Trylogia Władcy Pierścieni – J.R.R. Tolkien**
   Ta grupa obejmuje 6 tekstów opartych na dziełach Tolkiena. To klasyczna literatura fantasy, z rozbudowanym światem przedstawionym, archaizowanym językiem i epicką strukturą opowieści.

4. **Seria „Siedem Królestw" – Cinda Williams Chima**
   Czwarta kategoria to literatura młodzieżowa fantasy, zawierająca 3 książki z serii *Siedem Królestw*. Styl tych powieści jest bardziej współczesny, dynamiczny i przystępny, z wyraźnym podziałem na bohaterów i antagonistów oraz obecnością magii i wątków dorastania.



Rysunek 1. Przygotowane dokumenty tekstowe w języku polskim.

Wstępne przetwarzanie tekstów

Wstępne przetwarzanie tekstów miało na celu przygotowanie danych do dalszej analizy językowej. Proces ten został przeprowadzony w kilku krokach, wykorzystując narzędzia takie jak Morfeusz2 oraz biblioteki NLTK i standardowe moduły Pythona.

1. **Wczytanie danych**
   Dokumenty tekstowe zostały pobrane z korpusu i zapisane jako słownik, gdzie kluczem była nazwa pliku, a wartością jego treść.

2. **Ujednolicenie tekstu i usuwanie znaków specjalnych**
   Wszystkie teksty zostały zamienione na małe litery, aby uniknąć rozróżniania form różniących się jedynie wielkością liter. Następnie usunięto znaki interpunkcyjne, które nie wnoszą wartości semantycznej do analizy.

3. **Lematyzacja**
   Do analizy morfologicznej wykorzystano bibliotekę Morfeusz2. Dla każdego słowa wybierano jego podstawową formę (lemat), a następnie eliminowano elementy niebędące słowami (np. liczby, znaki).

4. **Usuwanie słów stopowych**
   Po tokenizacji (podziale tekstu na słowa), usunięto typowe polskie słowa stopowe na podstawie wcześniej wczytanej listy.

Wyniki – chmury wyrazów

W ramach projektu wygenerowano chmury wyrazów dla wszystkich przetworzonych dokumentów, jednak w poniższej analizie szczegółowo omawiam po jednym wybranym przykładzie z każdej kategorii tekstów. Pozwoliło to zachować reprezentatywność, a jednocześnie skupić się na najciekawszych przypadkach.

Adam Mickiewicz. „Dziady cz. II”

W chmurze wyrazów dla *Dziadów cz. II* wyraźnie dominują słowa „guślarz”, „kaplica”, „duch”, „prosić” i „niebo”. Od razu zauważalne są motywy obrzędu i kontaktu ze światem zmarłych – centralne dla tego aktu dramatu. Obecność takich słów jak „dziewczyna”, „dziecko”, „widma”, „ziemia” czy „świeca” dodatkowo podkreśla atmosferę rytuału, moralnej refleksji i duchowej przemiany. Chmura trafnie oddaje sakralny i metafizyczny charakter tekstu oraz relację między żywymi a zmarłymi.

Rysunek 2. Chmura wyrazów dla pliku Adam Mickiewicz. Dziady cz. II.txt.

Henryk Sienkiewicz. „Krzyżacy”

W chmurze wyrazów dla powieści *Krzyżacy* Henryka Sienkiewicza wyraźnie dominują imiona „danusia”, „zbyszko”, „jurand”, wskazujące na głównych bohaterów utworu. Obecność słów takich jak „krzyżak”, „zakon”, „bitwa”, „spychowo” i „rycerz” podkreśla historyczno-wojenny charakter fabuły oraz tło konfliktu z zakonem krzyżackim. Często pojawiają się też słowa emocjonalne i związane z relacjami, np. „córka”, „dziewczyna”, „ojciec” czy „uwolnić”, co odzwierciedla dramatyczny wątek porwania i walki o honor. Chmura wyrazów trafnie oddaje zarówno tematykę miłosno-przygodową, jak i historyczno-patriotyczną powieści.

Rysunek 3. Chmura wyrazów dla pliku Henryk Sienkiewicz. Krzyżacy.txt.

„Siedem Królestw. Tron Szarych Wilków”

W chmurze wyrazów dla powieści *Siedem Królestw. Tron Szarych Wilków* dominują imiona „rebeka” i „han”, które wskazują na głównych bohaterów fabuły. Silnie wybijają się także słowa „tron”, „śmierć”, „stracić” i „poświęcić”, sugerujące obecność wątków walki o władzę, ryzyka oraz osobistych strat. Pojęcia takie jak „kochać”, „przyjaciółka”, „żyła”, „duch” czy „przekonanie” podkreślają emocjonalny charakter historii i wewnętrzne konflikty postaci. Chmura dobrze oddaje połączenie elementów fantasy, intryg dworskich i relacji międzyludzkich, które są typowe dla tej serii młodzieżowej.

Rysunek 4. Chmura wyrazów dla pliku Siedem Królestw. Tron Szarych Wilków.txt.

Władca Pierścieni. „Powrót króla”

W chmurze wyrazów dla powieści *Władca Pierścieni. Powrót króla* wyraźnie wyróżniają się imiona „sam” i „frodo” – bohaterowie kluczowej misji związanej z Pierścieniem. Często pojawiają się także słowa „pierścień”, „zostawać”, „udawać”, „sauron” i „armia”, co wskazuje na tematykę poświęcenia, walki dobra ze złem i wielkiej bitwy. Obecność wyrazów takich jak „aragorn”, „gandalf”, „minas tirith” czy „powracać” dodatkowo podkreśla epicki charakter utworu oraz finałowy moment zjednoczenia sił przeciwko ciemności. Chmura dobrze oddaje emocjonalny i heroiczny ton ostatniej części trylogii.

Rysunek 5. Chmura wyrazów dla pliku Władca Pierścieni. Powrót króla.txt.

Podsumowanie

Analiza chmur wyrazów wygenerowanych dla wybranych tekstów literackich dobrze oddaje dominujące motywy, postacie i stylistykę każdej z książek.

W *Dziadach cz. II* Adama Mickiewicza dominują słowa związane z obrzędem i duchowością, takie jak „guślarz", „kaplica" i „duch", co podkreśla mistyczny i moralny charakter utworu.
W *Krzyżakach* Henryka Sienkiewicza wybijają się imiona „danusia", „zbyszko", „jurand" oraz terminy „krzyżak", „zakon", „bitwa", odzwierciedlając wątki historyczno-przygodowe i narodowe.
W *Tronie Szarych Wilków* z serii *Siedem Królestw* dominują bohaterowie „rebeka" i „alister" oraz emocjonalne hasła jak „śmierć", „kochać", „przekonanie", co wskazuje na młodzieżową literaturę fantasy z silnym ładunkiem relacyjnym.
Z kolei *Powrót króla* J.R.R. Tolkiena prezentuje bohaterów „sam", „frodo", „aragorn" i motywy takie jak „pierścień", „zostawać", „sauron" – oddając heroiczny klimat i walkę o ostateczne zwycięstwo dobra.

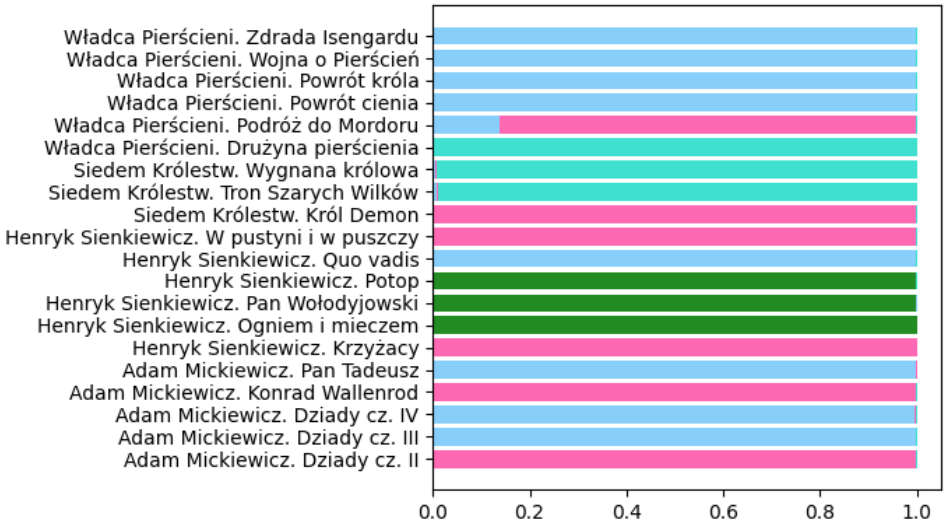
Chmury wyrazów okazały się trafnym wizualnym narzędziem do identyfikacji kluczowych treści, stylów i tematów obecnych w każdej kategorii literackiej analizowanego zbioru.

Modelowanie tematów

W projekcie zastosowano trzy różne modele tematyczne: **LDA (Latent Dirichlet Allocation)**, **NMF z funkcją straty Frobeniusa** oraz **NMF z dywergencją Kullbacka-Leiblera (KL)**. Celem było automatyczne wykrycie ukrytych tematów w zbiorze dokumentów literackich z czterech różnych kategorii.

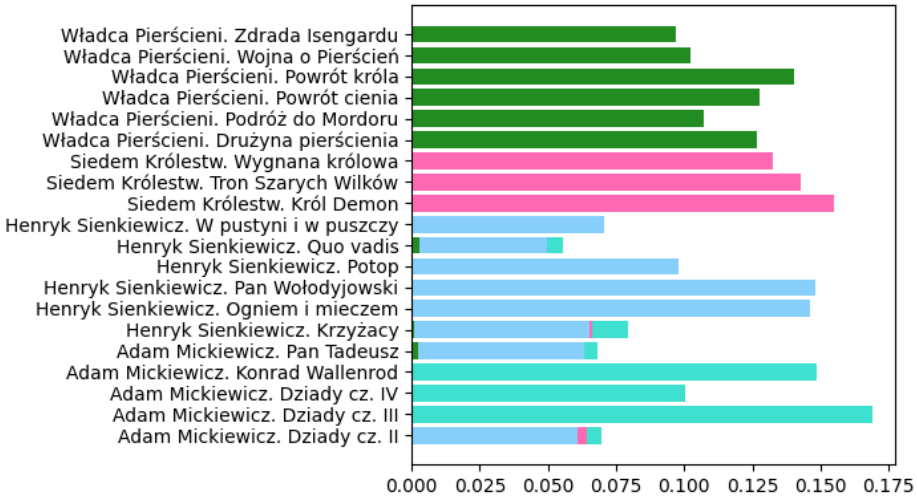
Porównanie modeli:

- **LDA**:
    - **Władca Pierścieni** został dość spójnie sklasyfikowany – niemal wszystkie tomy znalazły się w jednej grupie (jasnoniebieski), co świadczy o podobieństwie stylu i tematyki.
    - **Siedem Królestw** zostało częściowo rozdzielone – np. *Król Demon* przypisano do innego tematu niż pozostałe tomy.
    - **Twórczość Sienkiewicza** została podzielona – tylko kilka utworów znalazło się w tej samej grupie.
    - **Twórczość Mickiewicza** została podzielona – tylko kilka utworów znalazło się w tej samej grupie.



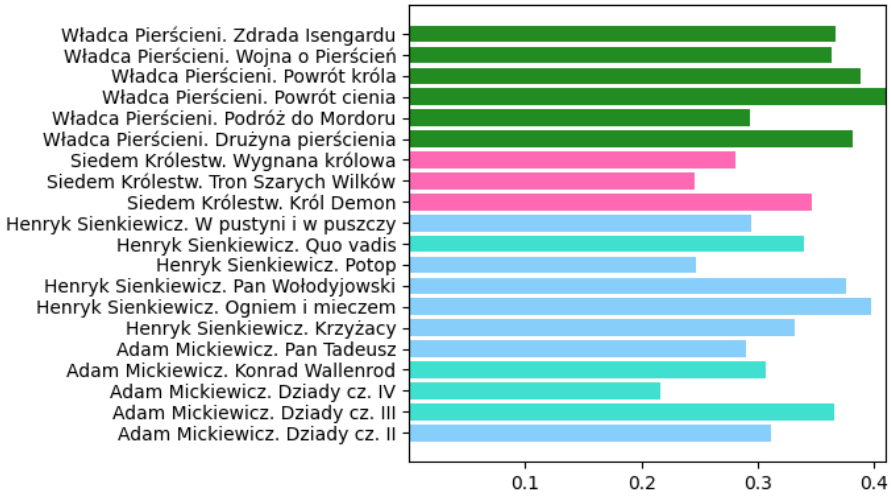
Rysunek 6. Tematy w modelu LDA.

- **NMF (Frobenius)**:
    - **Władca Pierścieni** został bardzo dobrze wyodrębniony – wszystkie tomy mają ten sam kolor (ciemnozielony), co wskazuje na spójność stylu i tematyki.
    - **Siedem Królestw** (różowy) także tworzy wyraźną, spójną grupę – to znak, że model dobrze rozpoznał wspólny język i motywy literatury młodzieżowej fantasy.
    - **Twórczość Sienkiewicza** (jasnoniebieski) została częściowo dobrze przypisana, ale kilka dokumentów (np. *Quo vadis* i *Krzyżacy*) otrzymało mieszaną tematykę.
    - **Mickiewicz** (turkusowy) w większości został poprawnie przypisany, choć np. *Pan Tadeusz* i *Dziady cz. II* pokazują udział innych tematów.



Rysunek 7. Tematy w modelu NMF (Frobenius).

- **NMF (KL-divergence)**:
    - **Władca Pierścieni** został bardzo wyraźnie i spójnie przypisany do jednej grupy (ciemnozielony), co świadczy o wysokiej spójności językowej i fabularnej tej serii.
    - **Siedem Królestw** (różowy) również zostało poprawnie oddzielone – wszystkie książki tej serii znalazły się w jednej, dobrze odseparowanej grupie.
    - **Henryk Sienkiewicz** (jasnoniebieski) tworzy spójną kategorię, model dobrze rozpoznał jego styl – historyczny, narracyjny i realistyczny.
    - **Adam Mickiewicz** (turkusowy) również został poprawnie wyodrębniony, mimo nieco większego zróżnicowania między utworami (np. *Pan Tadeusz* vs *Dziady*).



Rysunek 8. Tematy w modelu NMF (KL-divergence).

Podsumowanie:

Model LDA uchwycił ogólne podobieństwa językowe i tematyczne, szczególnie dobrze dla prozy fantasy (*Władca Pierścieni*), jednak miał trudności z rozróżnieniem bardziej subtelnych różnic stylu – np. między literaturą klasyczną. W porównaniu z modelami NMF (Frobenius lub KL) ten podział jest mniej jednoznaczny i bardziej „rozmyty", co jest charakterystyczne dla probabilistycznego podejścia LDA, które zakłada, że dokument może mieszać tematy. Model NMF z Frobeniusem lepiej niż LDA rozdziela dokumenty na spójne grupy tematyczne, szczególnie widoczne przy seriach takich jak *Władca Pierścieni* czy *Siedem Królestw*. Jednocześnie zachowuje pewną elastyczność w przypadku tekstów klasycznych, gdzie style bywają mniej jednoznaczne. Można go uznać za bardziej precyzyjny niż LDA, ale mniej wyraźny niż NMF z dywergencją KL, jeśli chodzi o maksymalną separację grup. Model NMF z KL osiągnął najlepsze wyniki pod względem jednoznaczności przypisania tematów. Każda kategoria literacka została wyraźnie rozdzielona, a dokumenty w obrębie grup mają wysoki udział tematu dominującego.

Analiza skupień

Zastosowano dwie metody grupowania dokumentów na podstawie ich podobieństwa:

- **Cosine similarity + ward linkage**
- **Euclidean distance + complete linkage**

Na podstawie wektorów TF-IDF wygenerowano dendrogramy przedstawiające hierarchiczną strukturę podobieństwa między dokumentami.

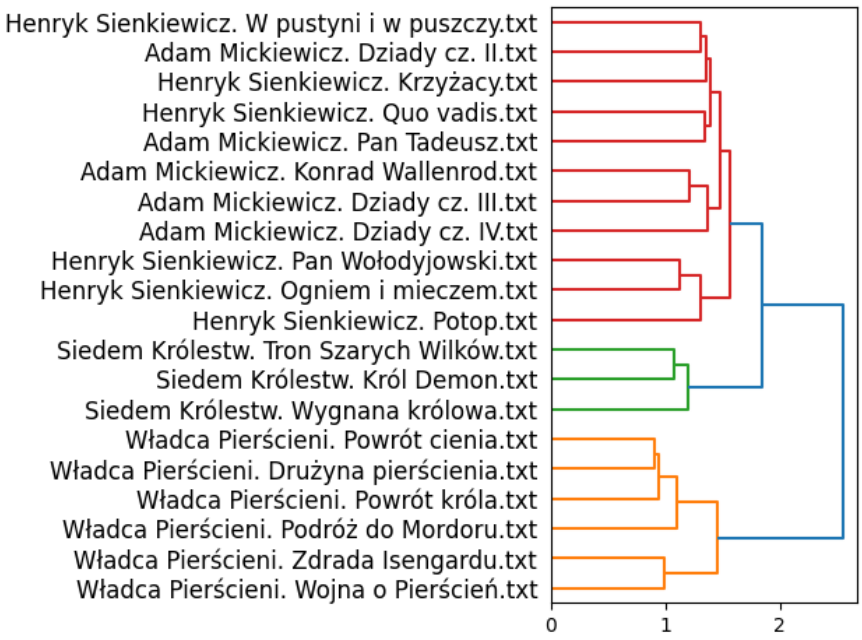
Dendrogram – metoda ward linkage

Pierwszy dendrogram został utworzony na podstawie metody **ward linkage**:

- **Władca Pierścieni** (dolny blok) tworzy bardzo zwartą i dobrze odseparowaną grupę, co potwierdza jego spójność stylu i słownictwa.
- **Siedem Królestw** również tworzy osobny, jednolity klaster – teksty tej serii są językowo zbliżone i wyraźnie odróżnialne od pozostałych.
- **Sienkiewicz i Mickiewicz** zostały umieszczone we wspólnym nadrzędnym klastrze, choć wewnątrz widoczne są podgrupy odpowiadające konkretnym autorom i stylom:
    - Mickiewicz: *Dziady*, *Pan Tadeusz*, *Konrad Wallenrod* – styl poetycki, klasyczny.
    - Sienkiewicz: *Krzyżacy*, *Potop*, *Quo vadis* – styl narracyjny, historyczny.

Dendrogram potwierdza skuteczność grupowania dokumentów według stylu i tematyki. Szczególnie dobrze zostały odseparowane teksty fantasy (Tolkien, Chima), co pokazuje dużą różnicę w słownictwie i strukturze w porównaniu z literaturą klasyczną. Metoda Ward linkage okazała się użyteczna dla **hierarchicznego podziału dokumentów**, zachowując zarówno

lokalne podobieństwa (np. między książkami jednej serii), jak i globalne (np. między autorami podobnego stylu).



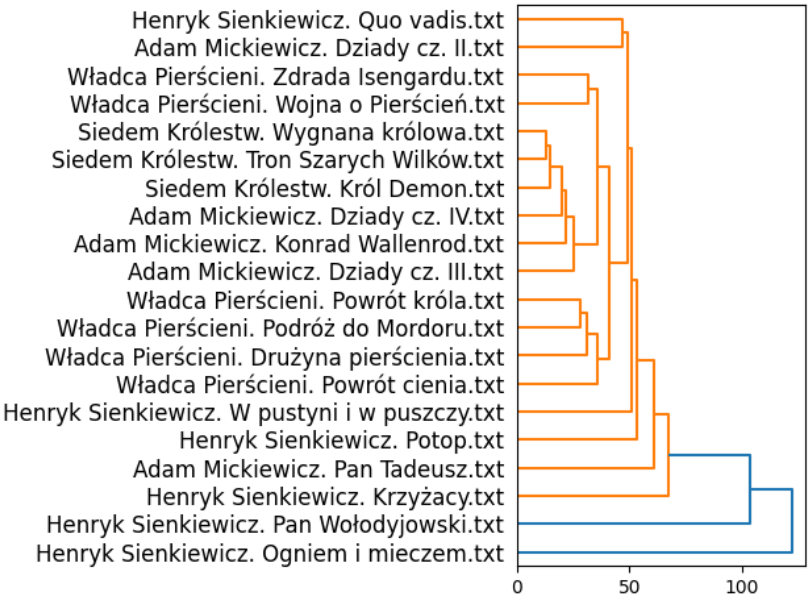
Rysunek 9. Wynik klasteryzacji metodą ward linkage.

Dendrogram – metoda complete linkage

W drugim podejściu użyto tej macierzy odległości euklidesowych i zastosowano metodę **complete linkage**:

- **Najbardziej zwartą i odseparowaną grupę** (oznaczoną na niebiesko) stanowią niektóre dzieła **Henryka Sienkiewicza** – *Krzyżacy*, *Pan Wołodyjowski*, *Ogniem i mieczem*. Zostały zgrupowane razem jako silnie podobne stylistycznie i leksykalnie.
- Pozostałe dokumenty zostały połączone w dużą strukturę (pomarańczową), co oznacza, że ich wzajemne różnice są większe, a metoda complete długo „wahała się", zanim je połączyła.
- **Władca Pierścieni**, **Siedem Królestw** i **Mickiewicz** (np. *Dziady*, *Konrad Wallenrod*) pojawiają się wymieszane – co sugeruje, że pod względem mierzonej odległości (w przestrzeni cech) nie różnią się od siebie na tyle, by utworzyć wyraźnie oddzielne klastry.
- Słabsze rozgraniczenie Tolkiena i fantasy młodzieżowego może wynikać z podobieństw w strukturze opowieści, słownictwie fabularnym lub częstotliwościach stylistycznych.

Complete linkage przy takim zbiorze i metryce wykazał największą spójność w obrębie tekstów Sienkiewicza, ale nie oddzielił wyraźnie pozostałych kategorii**.** W porównaniu z ward linkage**,** ten model słabiej różnicuje subtelne granice stylu – jest bardziej „ostrożny" i często odkłada decyzję o połączeniu klastrów aż do wyższych poziomów drzewa.



Rysunek 10. Wynik klasteryzacji metodą complete linkage.

Analiza n-gramów

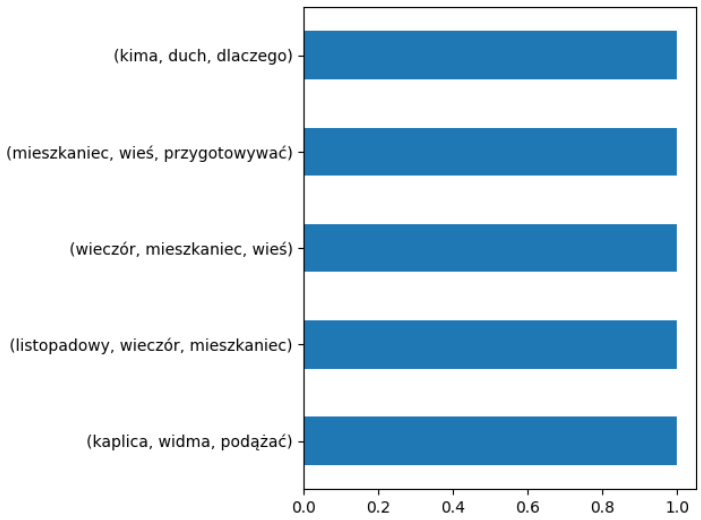
W celu lepszego zrozumienia struktury językowej dokumentów przeprowadzono analizę **n-gramów**, czyli ciągów kolejnych słów występujących w tekście. Skoncentrowano się na **trigramach (3-gramach)**, które pozwalają uchwycić najczęściej powtarzające się frazy, struktury składniowe oraz związki między bohaterami i wydarzeniami.

Dziady cz. II – Adam Mickiewicz

Najczęstsze trigramy to m.in.:

- *(kaplica, widma, podążać)*
- *(listopadowy, wieczór, mieszkaniec)*
- *(kima, duch, dlaczego)*

Frazy te silnie korespondują z rytualnym charakterem utworu – pojawiają się obrazy kaplicy, duchów i rozmów ze zmarłymi. Trigramy podkreślają klimat obrzędu dziadów oraz narrację o kontaktach z zaświatami.



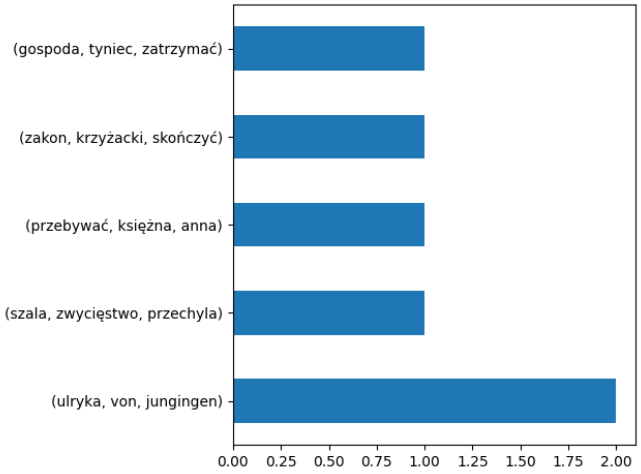
Rysunek 11. Najczęstsze trigramy w Dziady cz. II.

Krzyżacy – Henryk Sienkiewicz

Najczęściej występujące trigramy to:

- *(ulryka, von, jungingen)*
- *(szala, zwycięstwo, przechyla)*
- *(zakon, krzyżacki, skończyć)*

Wskazują one na obecność nazw historycznych oraz kontekst militarny. Frazy te odzwierciedlają główne motywy powieści – konflikt z zakonem, rycerskość oraz patriotyzm.



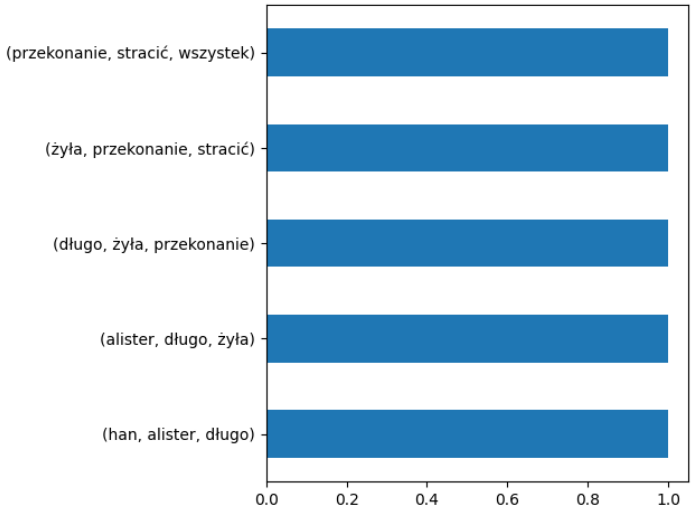
Rysunek 12. Najczęstsze trigramy w Krzyżacy.

Tron Szarych Wilków – Cinda Williams Chima

Najczęstsze trigramy:

- *(han, alister, długo)*
- *(żyła, przekonanie, stracić)*
- *(przekonanie, stracić, wszystek)*

Wskazują one na emocjonalny charakter fabuły, silnie skoncentrowanej na relacjach i dylematach moralnych. Imiona bohaterów pojawiają się w kontekście straty, przekonań i poświęcenia.



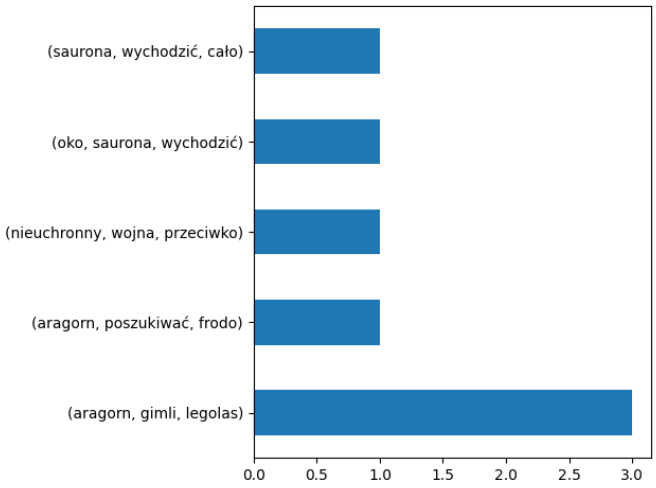
Rysunek 13. Najczęstsze trigramy w Tron Szarych Wilków.

Zdrada Isengardu – J.R.R. Tolkien

Najczęstsze trigramy:

- *(aragorn, gimli, legolas)*
- *(nieuchronny, wojna, przeciwko)*
- *(oko, saurona, wychodzić)*

Frazy te pokazują podział drużyny, przygotowania do walki oraz nieuchronność konfliktu. Podkreślona jest również rosnąca obecność Saurona – głównego zagrożenia.



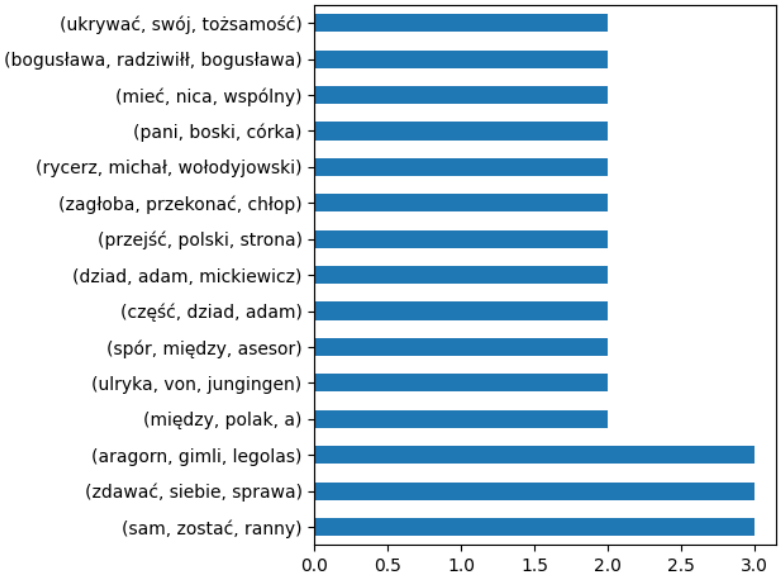
Rysunek 14. Najczęstsze trigramy w Zdrada Isengardu.

Zbiorcza analiza całego korpusu

Dla całego zbioru tekstów wygenerowano wspólny wykres pokazujący **15 najczęściej występujących trigramów**. Wśród nich pojawiły się:

- *(sam, zostać, ranny)*
- *(zdawać, siebie, sprawa)*
- *(aragorn, gimli, legolas)*
- *(ulryka, von, jungingen)*
- *(między, polak, a)*
- *(dziad, adam, mickiewicz)*

Analiza wskazuje na **dominację motywów patriotycznych, dramatycznych i fantastycznych**. Częste są imiona własne, motywy konfliktów, rozważań moralnych oraz elementy epickiej narracji.



Rysunek 15. Najczęstsze trigramy w całym korpusie.

Wnioski

Analiza trigramów pozwoliła uchwycić:

- charakterystyczne frazy dla danego autora lub cyklu literackiego,
- relacje między bohaterami i powtarzalne sekwencje działań,
- struktury emocjonalne i konflikty (np. wojna, poświęcenie, tożsamość),
- cechy językowe odzwierciedlające styl (np. retoryka Mickiewicza, narracja epicka Tolkiena).

Dzięki wizualizacji n-gramów możliwe było szybkie wychwycenie dominujących tematów i konstrukcji narracyjnych, co znacząco wspomaga eksplorację danych tekstowych.