UNIWERSYTET EKONOMICZNY W KATOWICACH

KIERUNEK Informatyka

Filip Misiak
147678

Analiza książek H.P Lovecraft

KATOWICE 2025

Spis treści

1. Wstęp

Celem niniejszej pracy jest przeprowadzenie kompleksowej analizy porównawczej trzech wybranych utworów literackich H.P. Lovecrafta: The Call of Cthulhu, The Dunwich Horror oraz The Shadow over Innsmouth, z wykorzystaniem narzędzi i metod przetwarzania języka naturalnego (Natural Language Processing – NLP). Przedmiotem analizy są pełne teksty dzieł, rozdzielone na rozdziały i akapity, co umożliwia badanie struktury i stylu pisarskiego autora zarówno na poziomie pojedynczych zdań i tokenów, jak i całych rozdziałów, akapitów i książek.

W pierwszym etapie pracy teksty zostały poddane wstępnemu przetwarzaniu, obejmującemu czyszczenie, normalizację, tokenizację oraz lematyzację przy użyciu biblioteki spaCy. W dalszej analizie przeprowadzono selekcję terminów, eliminując słowa nieistotne lub nieprzynoszące wartości informacyjnej, co umożliwiło skoncentrowanie się na jednostkach leksykalnych mających znaczenie semantyczne i stylistyczne. Następnie wykonano analizę częstości słów, co pozwoliło na stworzenie chmur słów zarówno dla terminów wspólnych dla dwóch dzieł, jak i dla słownictwa charakterystycznego dla poszczególnych utworów.

W kolejnych etapach zastosowano metody porównania podobieństwa między książkami i ich rozdziałami w oparciu o reprezentacje wektorowe tekstu. Wykorzystano dwa podejścia: klasyczne reprezentacje Tf-Idf oraz osadzenia semantyczne (word embeddings) generowane za pomocą modeli spaCy. Uzyskane reprezentacje posłużyły do przeprowadzenia grupowania rozdziałów metodą k-średnich (K-Means clustering), co pozwoliło na ocenę, czy poszczególne fragmenty tekstu mogą być skutecznie przypisane do dzieł, z których pochodzą.

Równolegle przeprowadzono analizę tematyczną (topic modeling) z zastosowaniem modelu LDA (Latent Dirichlet Allocation), umożliwiającą wyodrębnienie dominujących tematów w każdej z książek oraz przypisanie rozdziałów do utworów na podstawie ich tematycznej struktury.

W końcowej części projektu skupiono się na zadaniu klasyfikacji akapitów do odpowiadających im dzieł. W tym celu wykorzystano trzy klasyfikatory: Logistic Regression, MLPClassifier oraz SVM, testując różne reprezentacje tekstu: wagi binarne, logarytmiczne, Tf-Idf oraz osadzenia semantyczne. Wyniki klasyfikacji zostały porównane pod względem skuteczności, co umożliwiło identyfikację najlepszego modelu w kontekście przyjętego zadania.

Uzupełnieniem analizy była eksploracja nastroju poszczególnych rozdziałów, przeprowadzona z wykorzystaniem biblioteki TextBlob i Vader. Ocenie poddano biegunowość polaryzację tekstu w ramach każdego rozdziału. Wyniki zilustrowano wykresami ukazującymi zmienność nastroju w strukturze utworu. Dodatkowo, w celu głębszego wglądu w emocjonalny wydźwięk tekstów, zastosowano leksykon NRC Emotion Lexicon, który umożliwił zliczenie wystąpień słów związanych z ośmioma podstawowymi emocjami (takimi jak strach, gniew, radość czy smutek) i stworzenie profilu emocjonalnego dla każdej książki. Pozwoliło to na jakościowe i ilościowe porównanie emocjonalnej tonacji dzieł Lovecrafta.

2. Analiza termów

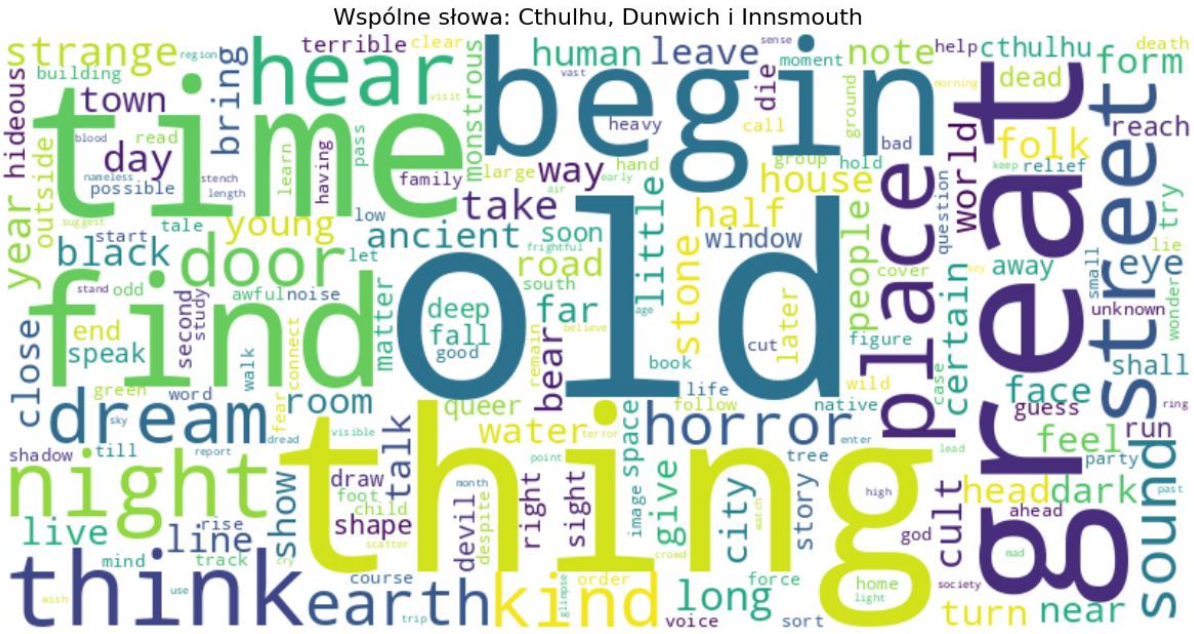
Do funkcji tworzącej termy dodano rozszerzony zestaw słów stop (ang. stopwords), zawarty w zmiennej my_stopwords, ponieważ zawarte tam wyrazy nie wnoszą istotnych informacji semantycznych i nie wpływają znacząco na odróżnianie tematów, stylu ani treści książek.

Słowa takie jak "one", "man", "come", "say", "know", "see", "look", "go", "like" czy "tell" są bardzo ogólne i uniwersalne — pojawiają się w niemal każdym rodzaju narracji i nie niosą unikalnych cech stylistycznych ani tematycznych. Ich częste występowanie mogłoby zaciemniać wyniki analizy, np. w chmurach słów, modelach tematycznych LDA, analizie TF-IDF czy klasyfikacji.

Z kolei słowa takie jak "state" (w tym przpadku nazwa stanu), "mean" czy "set" są wieloznaczne i często pełnią funkcje pomocnicze w zdaniach, nie wnosząc jednoznacznej wartości analitycznej. Wyrazy typu "ba", "obe", "aout" mogą być pozostałościami błędnego wczytywania lub przetwarzania tekstu (literówki, przekłamania OCR) i również nie mają wartości analitycznej możliwe jest również, że stanowią część języka stworzonego przez Lovecrafta, którym posługują się ludzie do porozumiewania się z istotami przez niego wykreowanymi.

Podsumowując, decyzja o dodaniu tych słów do listy stopwords została podjęta w celu oczyszczenia analizy ze zbędnych elementów językowych, które mogłyby zaburzyć dokładność interpretacji, analizy tematycznej i klasyfikacji. Umożliwia to wyeksponowanie bardziej znaczących terminów charakterystycznych dla poszczególnych książek i stylu autora.

3. Chmury słów



Wspólne słowa: Cthulhu, Dunwich i Innsmouth

Chmura przedstawia najczęściej powtarzające się słowa we wszystkich trzech opowiadaniach Lovecrafta. Pokazuje, co łączy „The Call of Cthulhu", „The Dunwich Horror" i „The Shadow over Innsmouth" na poziomie języka i motywów.

Narracja odkrywania tajemnicy:

Czasowniki takie jak **find**, **begin**, **hear**, **think**, **see**, **dream**, **talk**, **note**, **face** odzwierciedlają drogę bohatera w głąb tajemnicy. Narrator nie wie wszystkiego – jego wiedza budowana jest stopniowo, przez słuchanie, czytanie, śledzenie znaków.

Przestrzeń jako element grozy:

Słowa **place**, **house**, **street**, **room**, **stone**, **hill**, **city**, **outside** wskazują na znaczenie przestrzeni – często zamkniętej, klaustrofobicznej lub przesiąkniętej starożytnością. Lovecraft buduje lęk także poprzez scenerię.

Czas i starożytność:

Wyrazy **old**, **ancient**, **year**, **time**, **later**, **night**, **day** pokazują kontrast między teraźniejszością a przeszłością. Kluczowym motywem jest **powrót przeszłości**, która miała pozostać zapomniana.

Groza egzystencjalna:

Częste słowa to **horror**, **thing**, **fear**, **dark**, **dead**, **hideous**, **monstrous**, **dread**, **terror**, **unknown**. Opisy Lovecrafta są zdominowane przez **niewyrażalne zagrożenie**, które budzi nie tyle strach, co egzystencjalny lęk.

Mitologia i kulty:

Słowa **cult**, **god**, **form**, **figure**, **folk**, **voice**, **world**, **human** sugerują zderzenie człowieka z nieludzkimi siłami. Lovecraft kreuje świat, w którym ludzkość nie jest centrum wszechświata – a raczej ofiarą kosmicznych bytów.

Ciało, zmysły, przemiana:

Pojawiają się słowa jak **eye**, **face**, **head**, **voice**, **form**, **shape**, **shadow**, **sound**, **figure** – wskazujące na cielesność doświadczenia grozy.

Ruch i akcja:

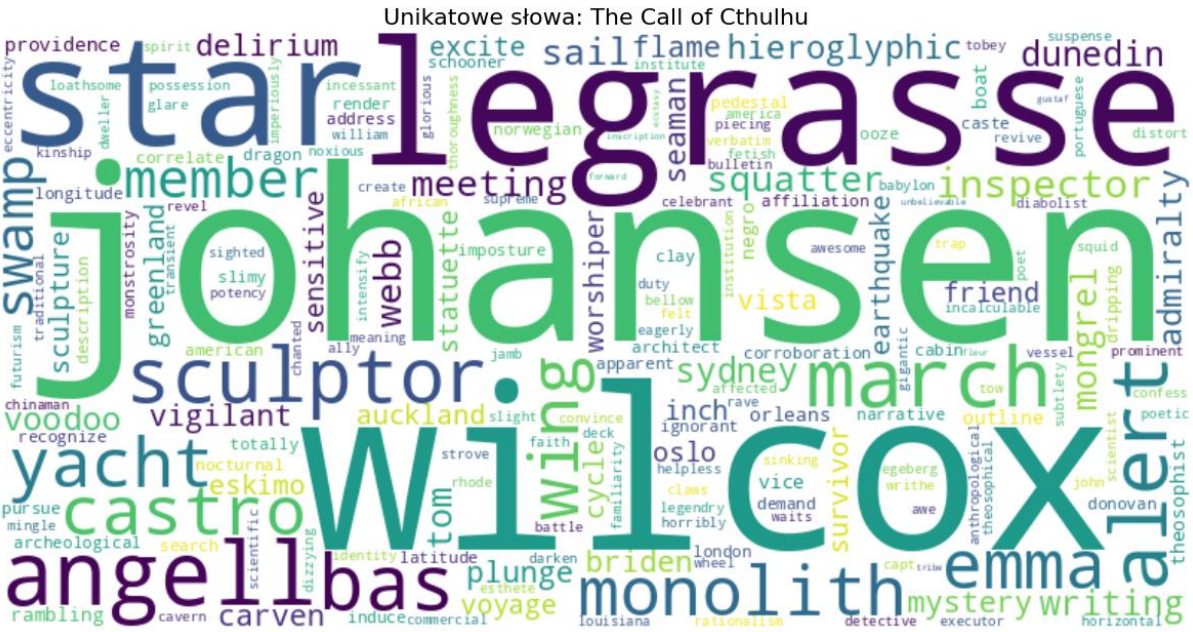
Czasowniki **leave**, **take**, **run**, **fall**, **turn**, **walk**, **start**, **pass**, **follow** wprowadzają dynamikę. Narracje Lovecrafta często mają formę podróży – bohater przemieszcza się fizycznie i metaforycznie w stronę wiedzy (i zagłady).

Znaczenie słowa i zapisu:

Słowa **note**, **story**, **book**, **word**, **report**, **tale** wskazują, że informacje o świecie przedstawionym bywa przekazywana przez teksty i opowieści.

Podsumowanie:

Wspólne słowa ukazują trzon stylu Lovecrafta: **ontologiczny lęk, starożytną mitologię, powolne odkrywanie prawdy oraz przenikanie rzeczywistości przez coś niewyrażalnego**. Choć każde z opowiadań ma odrębną fabułę, to łączy je język opisu nieznanego, formy narracyjne i pesymizm.



Chmura słów zawiera terminy, które pojawiają się wyłącznie w opowiadaniu **„The Call of Cthulhu”**, a nie występują w „The Dunwich Horror” ani „The Shadow over Innsmouth”.

Postacie i lokalizacje globalne:

Nazwiska takie jak **Legrasse**, **Wilcox**, **Johansen**, **Castro**, **Angell**, a także miejsca: **Auckland**, **Oslo**, **Greenland**, **Dunedin**, **Sydney**, **Swamp** – wskazują na szeroką skalę wydarzeń i zróżnicowaną geografię. Fabuła oparta jest na gromadzeniu relacji świadków z różnych części świata.

Motywy morskie i eksploracyjne:

Wyrazy **yacht**, **voyage**, **admiralty**, **seaman**, **iceberg**, **nautical**, **ship**, **sail**, **survivor** wskazują na obecność tematyki podróży morskiej i odkrycia zatopionego miasta. Ocean symbolizuje tutaj niepoznane – przestrzeń, z której wyłania się pradawna groza.

Kult i rytuał:

Pojawiają się słowa takie jak **voodoo**, **idol**, **cult**, **hieroglyphic**, **statue**, **statuette**, **carven**, **clay**, **sculptor**, **inscription**, **monolith** – wszystkie związane z materialnymi śladami działalności czcicieli Cthulhu. Elementy te sugerują istnienie zorganizowanego, globalnego kultu i jego rytualnych artefaktów.

Inność i kolonialna egzotyka:

Obecność słów jak **eskimo**, **negro**, **mongrel**, **chinaman**, **tribe**, **african**, **american** ukazuje sposób, w jaki Lovecraft przedstawiał „Innych" – jako nosicieli starożytnej wiedzy, często w sposób nacechowany kolonialnymi stereotypami.

Groza wewnętrzna i psychiczna:

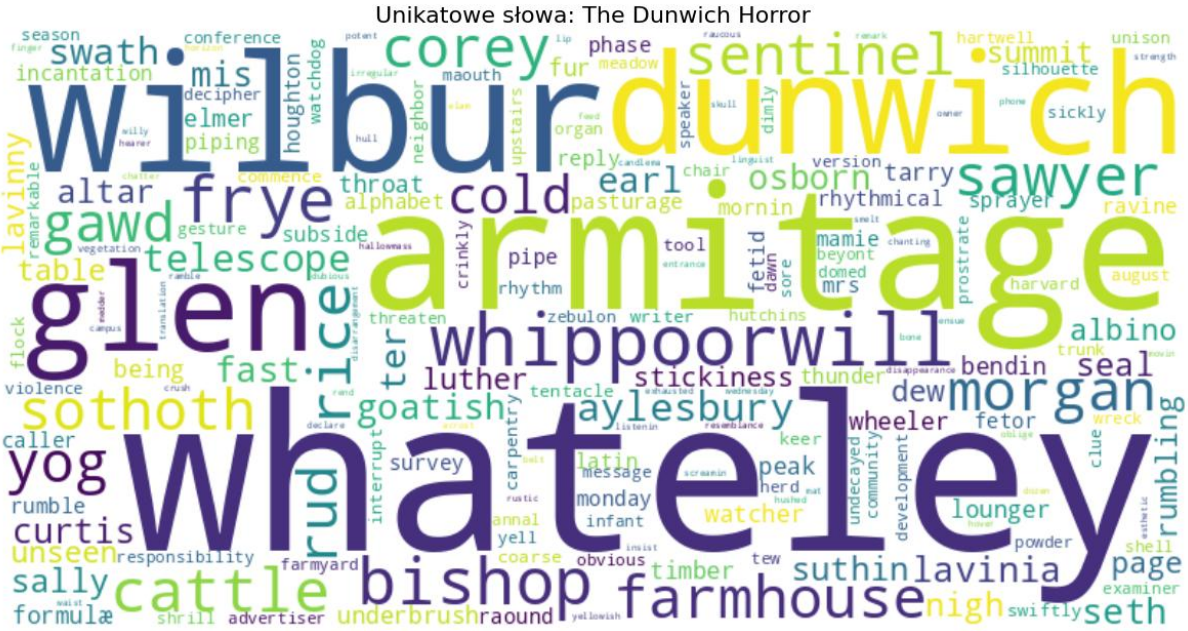
Słowa **delirium**, **obsession**, **fearfully**, **clammy**, **ooze**, **vice**, **horribly**, **diabolist**, **monstrosity** wskazują na horror o charakterze psychologicznym i egzystencjalnym. Bohaterowie często balansują na krawędzi obłędu wskutek zetknięcia z niepojętą prawdą.

Forma śledcza i dokumentalna:

Terminy takie jak **report**, **outline**, **narrative**, **writing**, **executor**, **address**, **bulletin**, **meeting** podkreślają quasi-dziennikarski charakter narracji. Cała opowieść ma strukturę rekonstrukcji zdarzeń na podstawie dokumentów i relacji.

Podsumowanie:

„The Call of Cthulhu" wyróżnia się globalną skalą, narracją opartą na dokumentach oraz centralnym motywem kultu pradawnego bóstwa, które mimo snu na dnie oceanu wciąż wpływa na ludzi. Unikatowe słowa w chmurze uwypuklają **rozległość geograficzną, elementy morskie, rytuały kultowe oraz psychologiczne napięcie**, które odróżniają to opowiadanie od pozostałych.



Chmura słów: „Unikatowe słowa: The Dunwich Horror”

Chmura słów przedstawia wyrazy, które pojawiają się wyłącznie w opowiadaniu **„The Dunwich Horror”** H.P. Lovecrafta, a nie występują w „The Shadow over Innsmouth” ani „The Call of Cthulhu”.

Postacie i nazwiska lokalne:

Największe słowa to m.in. **Whateley**, **Wilbur**, **Armitage**, **Lavinia**, **Bishop**, **Curtis**, **Frye**, **Morgan**, **Osborn** – to mieszkańcy Dunwich i osoby związane z głównym wątkiem fabularnym. Wskazuje to na silne osadzenie opowiadania w małej, odizolowanej społeczności wiejskiej.

Topografia i wiejskie otoczenie:

Słowa takie jak **Dunwich**, **farmhouse**, **glen**, **ravine**, **cattle**, **pasturage**, **underbrush**, **swath**, **farmyard** wskazują na rustykalne, górskie otoczenie akcji. Przestrzeń ta ma charakter zamknięty, pełen zarośli, wzgórz i odizolowanych gospodarstw.

Motywy okultystyczne i nadprzyrodzone:

Wyrazy **Yog**, **Sothoth**, **incantation**, **altar**, **formulae**, **interrupter**, **undergraduate**, **Harvard**, **telescope** wskazują na obecność zakazanej wiedzy, rytuałów i nauki jako narzędzi do walki z nadprzyrodzonym.

Atmosfera grozy i dezintegracji:

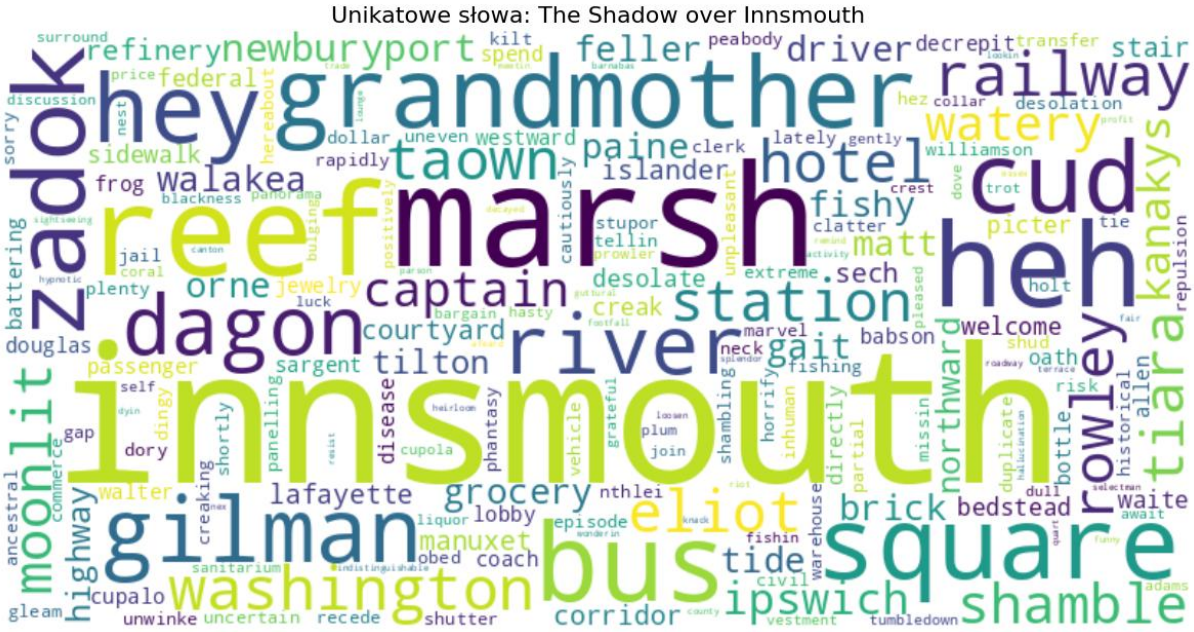
Pojawiają się słowa takie jak **goatish**, **thunder**, **stickiness**, **unseen**, **fetor**, **piping**, **prostrate**, **violence**, które budują nastrój tajemnicy, przemocy i cielesnej degeneracji.

Zwierzęta i natura jako zwiastuny zagrożenia:

Obecność słów **whippoorwill**, **cattle**, podkreśla rolę przyrody w narracji – zwierzęta wyczuwają zagrożenie, stają się jego ofiarami lub świadkami.

Podsumowanie:

„The Dunwich Horror" to opowieść silnie osadzona w górskim, wiejskim krajobrazie, pełna symboliki degeneracji, okultystycznych praktyk i zderzenia wiedzy akademickiej z niewypowiedzianym horrorem. Unikatowe słowa tej chmury podkreślają **lokalność, pokoleniowe tajemnice i nie-ludzką obecność**, która czai się w ukryciu.



Chmura słów: „Unikatowe słowa: The Shadow over Innsmouth"

Chmura słów zawiera terminy charakterystyczne tylko dla opowiadania **„The Shadow over Innsmouth"** i nieobecne w „The Dunwich Horror" oraz „The Call of Cthulhu".

Miejsce akcji i geografia miejska:

Dominują słowa **Innsmouth**, **square**, **station**, **railway**, **bus**, **hotel**, **refinery**, **grocery**, **courtyard**, które wskazują na zrujnowane, nadmorskie miasteczko. Miasto jest jednocześnie przestrzenią fizyczną i metaforyczną – zamkniętą, zdegenerowaną enklawą.

Postacie i lokalna społeczność:

Imiona i nazwiska jak **Zadok**, **Gilman**, **Howley**, **Grandmother**, **Hey**, **Feller**, **Eliot**, **Cud** budują obraz zamkniętej, podejrzanej społeczności z własną historią i dziwactwami. Postać Zadoka Allena jest szczególnie ważna jako nośnik ukrytej wiedzy.

Motywy morskie i rybo-ludzkie:

Pojawiają się słowa **reef**, **dagon**, **watery**, **river**, **moonlit**, **fishy**, **tide**, **frogs**, **shamble** –

wskazujące na obecność istot z głębin i ich wpływ na ludzi. Te elementy są centralne dla grozy opowiadania – stopniowej przemiany mieszkańców w hybrydy.

Atmosfera degeneracji i obcości:

Wyrazy takie jak **decrepit**, **desolate**, **disease**, **repulsion**, **brick**, **tumbledown** podkreślają zniszczenie, zaniedbanie i moralną pustkę miasta. To miejsce zapomniane przez świat i zawładnięte przez nieludzkie siły.

Podróż i odkrywanie tajemnicy:

Słowa **driver**, **bus**, **railway**, **station**, **highway**, **refinery**, **transfer**, **welcome** ukazują, że fabuła rozgrywa się jako podróż – dosłowna i metaforyczna – w głąb ukrytej prawdy o Innsmouth i jego mieszkańcach.

Podsumowanie:

„The Shadow over Innsmouth" to historia osadzona w zrujnowanym, odizolowanym mieście, gdzie morskie bóstwa i ich potomkowie dominują nad upadłą społecznością. Unikatowe słowa podkreślają **motyw degeneracji, izolacji oraz wypieranej wiedzy o przerażającym dziedzictwie**. W przeciwieństwie do bardziej rozproszonych i globalnych narracji Lovecrafta, „Innsmouth" jest zamkniętym mikrokosmosem grozy.

4. Grupowanie

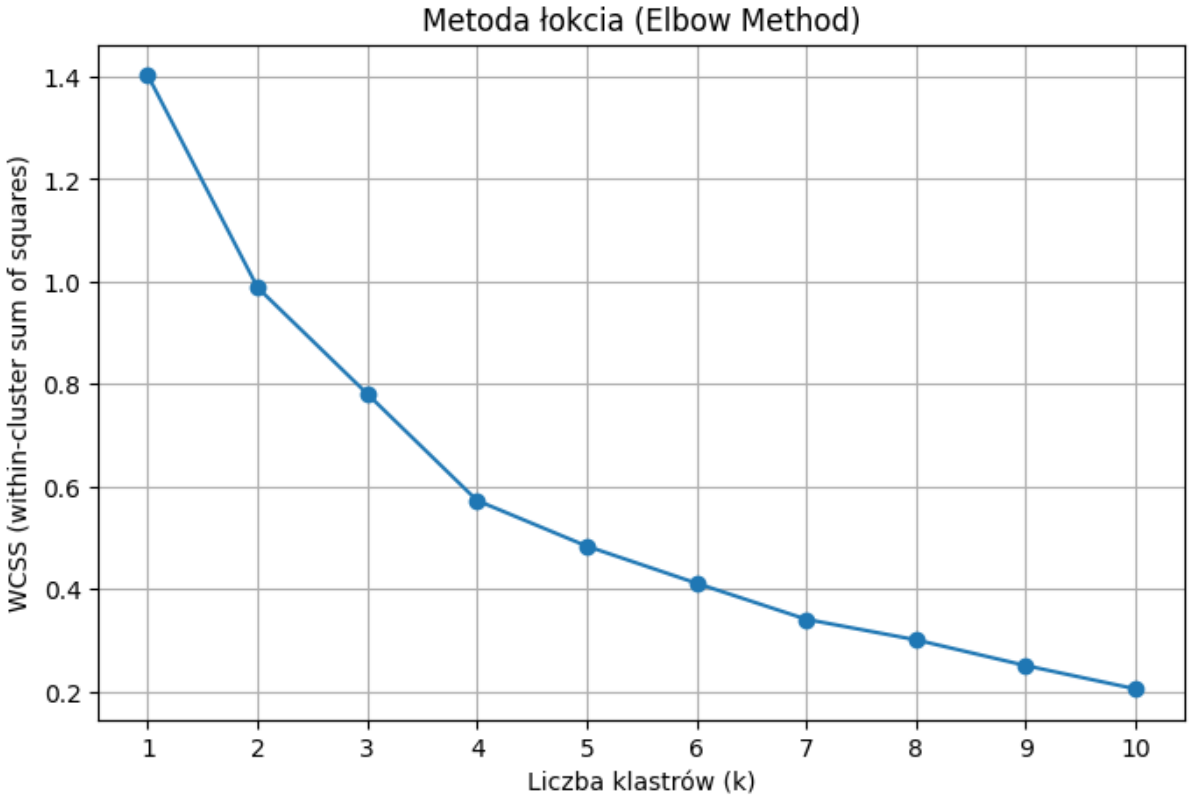
Celem była analiza struktury tematycznej korpusu tekstów H.P. Lovecrafta pod kątem sprawdzenia, czy istniejące w zbiorze naturalne grupy dokumentów odpowiadają przyjętym kategoriom, czyli opowiadaniom: The Call of Cthulhu, The Dunwich Horror i The Shadow over Innsmouth. W tym celu zastosowaliśmy dwie niezależne metody eksploracyjne: grupowanie (clustering) oraz modelowanie tematów (topic modelling).

Grupowanie (clustering)

W ramach grupowania wykorzystaliśmy algorytm KMeans, operując na wektorowych reprezentacjach tekstów (m.in. TfIdf). Każdy dokument został zredukowany do przestrzeni cech, a następnie przypisany do jednej z grup wyznaczonych przez algorytm.

Dla KMeans analizowaliśmy zarówno wartości silhouette score, jak i zgodność otrzymanych klastrów z rzeczywistymi etykietami (tj. przypisaniem dokumentów do jednego z trzech opowiadań). W przypadku dobrze dobranej liczby klastrów (np. k=3) zauważyliśmy częściowe dopasowanie do oryginalnych kategorii, co sugeruje, że poszczególne opowiadania zawierają odrębne, rozpoznawalne cechy językowe i tematyczne. Jednak pewna liczba dokumentów (zwłaszcza na granicach tematycznych) była klasyfikowana błędnie, co wskazuje na nakładanie się stylów i słownictwa Lovecrafta pomiędzy tekstami.

Analiza grupowania na wektorach dokumentów ze spacy



Na wykresie zgięcie łokcia w okolicach k = 3 lub 4.

--- Przypisanie dokumentów do klastrów ---

Dokument 0: klaster 0, książka: Cthulhu

Dokument 1: klaster 0, książka: Cthulhu

Dokument 2: klaster 0, książka: Cthulhu

Dokument 3: klaster 0, książka: Dunwich

Dokument 4: klaster 0, książka: Dunwich

Dokument 5: klaster 0, książka: Dunwich
Dokument 6: klaster 1, książka: Dunwich
Dokument 7: klaster 1, książka: Dunwich
Dokument 8: klaster 0, książka: Dunwich
Dokument 9: klaster 1, książka: Dunwich
Dokument 10: klaster 0, książka: Dunwich
Dokument 11: klaster 1, książka: Dunwich
Dokument 12: klaster 1, książka: Dunwich
Dokument 13: klaster 2, książka: Innsmouth
Dokument 14: klaster 0, książka: Innsmouth
Dokument 15: klaster 1, książka: Innsmouth
Dokument 16: klaster 2, książka: Innsmouth
Dokument 17: klaster 2, książka: Innsmouth

Silhouette Score: 0.2089

Analiza wyników klasteryzacji dokumentów przy użyciu algorytmu **KMeans** na podstawie wektorów uzyskanych z modelu **spaCy** wskazuje na pewien poziom struktury w danych, ale także istotne trudności w precyzyjnym oddzieleniu książek.
Klaster 0: Ten klaster zawiera aż 10 dokumentów, z czego:

- 3 pochodzą z książki **Cthulhu** (wszystkie rozdziały tej książki),
- 6 z książki **Dunwich**,
- 1 z książki **Innsmouth**.

Oznacza to, że dokumenty z **The Call of Cthulhu** zostały całkowicie zgrupowane razem, co sugeruje ich względną spójność stylistyczno-semantyczną według reprezentacji spaCy. Jednak duża część rozdziałów z **The Dunwich Horror** również trafiła do tego klastra – wskazuje to na znaczące podobieństwo między tymi dwoma książkami w ujęciu wektorowym.
Klaster 1: Zawiera 5 dokumentów:

- 4 z książki **Dunwich**,
- 1 z książki **Innsmouth**.

Ten klaster zbiera pozostałe rozdziały z **The Dunwich Horror**, które najprawdopodobniej różniły się od tych z klastra 0 (np. pod względem tematyki, tonu lub słownictwa). Obecność dokumentu z **Innsmouth** może świadczyć o pewnym pokrewieństwie semantycznym jednego z jej rozdziałów do tej właśnie grupy.

Klaster 2: Zawiera 3 dokumenty:

- wszystkie pochodzą z **The Shadow over Innsmouth**.

Ten wynik jest częściowo zgodny z oczekiwaniami – część rozdziałów z tej książki tworzy własny, osobny klaster, co świadczy o tym, że mogą one znacząco różnić się od pozostałych książek, prawdopodobnie ze względu na inne słownictwo, tematykę lub ton narracji.

Podsumowanie i wnioski:

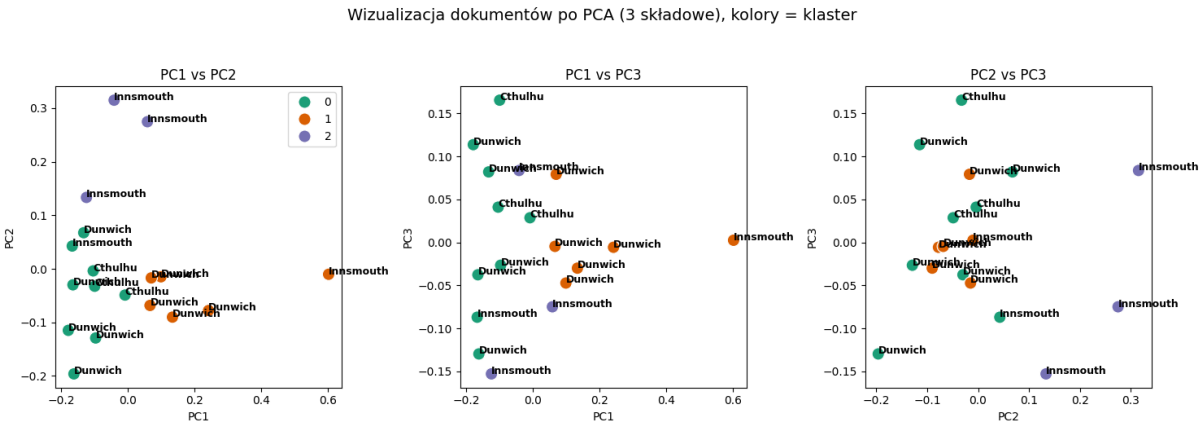
- Spójność wewnętrzna „Cthulhu”: Wszystkie rozdziały tej książki zostały przypisane do jednego klastra, co sugeruje ich wysoką spójność semantyczną.

- Rozproszenie „Dunwich”: Rozdziały z „The Dunwich Horror” są rozdzielone pomiędzy dwa klastry (0 i 1), co wskazuje na większą różnorodność tematyczno-stylistyczną tej książki – jej fragmenty są bardziej zróżnicowane lub pośrednie między pozostałymi dziełami.

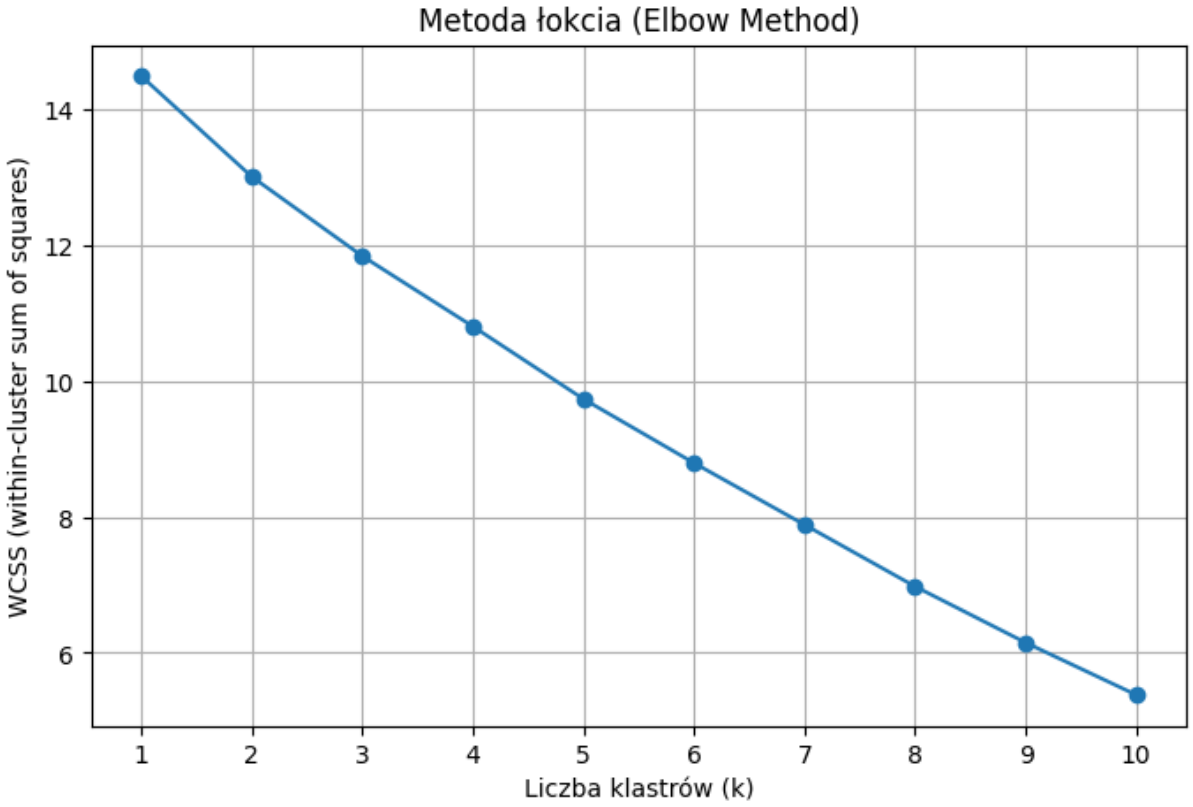
- Częściowa separacja „Innsmouth”: Książka ta została podzielona między trzy klastry (0, 1 i 2), ale większość jej rozdziałów znalazła się w klastrze 2, co sugeruje, że przynajmniej część książki jest dobrze odróżnialna od innych.

- Jakość klasteryzacji: KMeans w tym przypadku częściowo uchwycił strukturę danych – jeden klaster dobrze odpowiada jednej książce („Cthulhu”), inny dominuje w „Innsmouth”, ale „Dunwich” okazuje się trudniejszy do jednoznacznej segmentacji. To może wynikać z większego zróżnicowania rozdziałów tej książki lub pośredniego charakteru jej stylu względem pozostałych.

- Silhouette Score na poziomie 0.2089 wskazuje, że struktura klastrów jest dość słaba – co oznacza, że granice między klastrami nie są wyraźne, a dokumenty w niektórych przypadkach są bardziej podobne do dokumentów z innych klastrów niż do tych ze swojego własnego. Oznacza to, że wektory dokumentów wygenerowane przez model spaCy nie rozdzielają wystarczająco wyraźnie rozdziałów trzech książek Lovecrafta.



Analiza grupowania TFIDF



Na wykresie nie widać wyraźnego zgięcia łokcia — krzywa WCSS zmniejsza się dość równomiernie i liniowo w całym zakresie od k=1 do k=10. Brakuje punktu, w którym redukcja WCSS nagle zwalnia, co utrudnia wskazanie optymalnej liczby klastrów na podstawie metody łokcia.

--- Przypisanie dokumentów do klastrów ---
Dokument 0: klaster 0, książka: Cthulhu
Dokument 1: klaster 0, książka: Cthulhu
Dokument 2: klaster 0, książka: Cthulhu
Dokument 3: klaster 2, książka: Dunwich
Dokument 4: klaster 1, książka: Dunwich
Dokument 5: klaster 1, książka: Dunwich
Dokument 6: klaster 1, książka: Dunwich
Dokument 7: klaster 1, książka: Dunwich
Dokument 8: klaster 1, książka: Dunwich
Dokument 9: klaster 1, książka: Dunwich
Dokument 10: klaster 1, książka: Dunwich
Dokument 11: klaster 1, książka: Dunwich
Dokument 12: klaster 1, książka: Dunwich
Dokument 13: klaster 2, książka: Innsmouth
Dokument 14: klaster 2, książka: Innsmouth
Dokument 15: klaster 2, książka: Innsmouth
Dokument 16: klaster 2, książka: Innsmouth
Dokument 17: klaster 2, książka: Innsmouth

--- Najważniejsze termy dla klastrów ---
Klaster 0: dream, johansen, cult, wilcox, professor, uncle, legrasse, cthulhu, star, angell
Klaster 1: whateley, armitage, wilbur, hill, old, glen, dunwich, thing, whippoorwill, great
Klaster 2: innsmouth, street, marsh, old, thing, town, grandmother, cross, door, bus

Silhouette Score: 0.0506

Przypisania dokumentów do klastrów:

Dokumenty 0–2 (Cthulhu) zostały jednoznacznie przypisane do klastra 0, co sugeruje, że teksty z tej książki mają odrębny, rozpoznawalny styl lub słownictwo względem pozostałych.

Dokumenty 3–12 (Dunwich Horror) zostały w większości przypisane do klastra 1, z wyjątkiem dokumentu 3 (klaster 2). To oznacza względną spójność stylu Dunwich, ale obecność dokumentu 3 w klastrze 2 może świadczyć o fragmentarycznym podobieństwie do Innsmouth lub niejednorodności samej książki.

Dokumenty 13–17 (Shadow over Innsmouth) konsekwentnie trafiły do klastra 2, co wskazuje na ich wewnętrzną spójność i odmienność względem pozostałych tekstów.
Najważniejsze termy dla każdego klastra: Analiza słów kluczowych dla klastrów bardzo dobrze koresponduje z treścią książek:

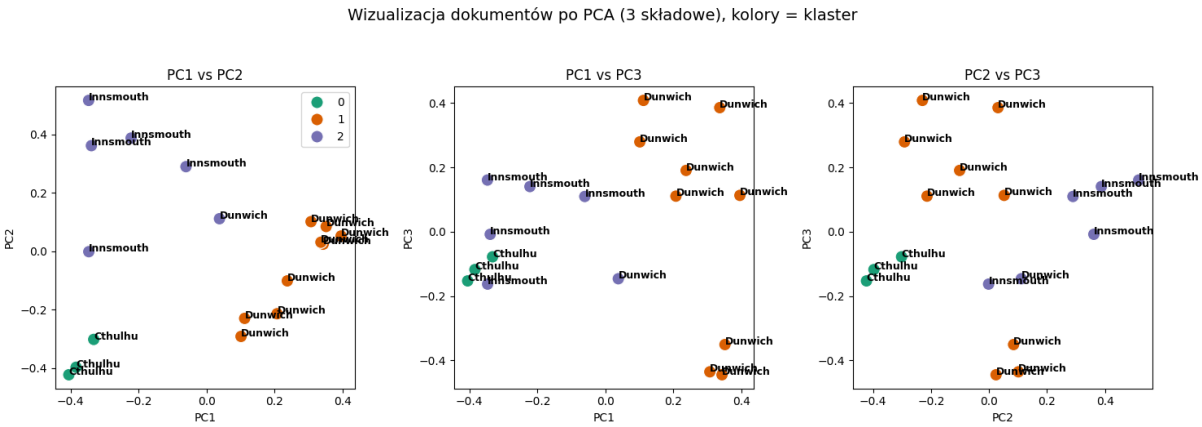
Klaster 0 (Cthulhu): zawiera wyraźnie charakterystyczne słowa jak cthulhu, legrasse, wilcox, johansen, professor, które są ściśle związane z fabułą i postaciami z „The Call of Cthulhu".

Klaster 1 (Dunwich): dominuje tu słownictwo typowe dla „The Dunwich Horror" – whateley, armitage, wilbur, dunwich, whippoorwill – co dobrze odzwierciedla tematykę tej opowieści.

Klaster 2 (Innsmouth): zawiera terminy takie jak innsmouth, marsh, grandmother, bus, street – jednoznacznie kojarzące się z miejscem akcji i narracją w „The Shadow over Innsmouth".

Wartość Silhouette Score: 0.0506

To bardzo niska wartość, wskazująca to na słabe wyodrębnienie klastrów. Punkty są słabo oddzielone od siebie, a wiele dokumentów znajduje się blisko granic klastrów lub w niejednoznacznym położeniu.



5. LDA Topic modeling

W drugiej części kroku zastosowaliśmy modelowanie tematów przy pomocy algorytmu LDA (Latent Dirichlet Allocation), aby zbadać, czy da się wyodrębnić dominujące motywy tematyczne w korpusie, które pokrywałyby się z przyjętymi kategoriami tekstów.

Każdy dokument został przekształcony do formy rozkładu tematów, a następnie analizowaliśmy, jak często dany temat występuje w poszczególnych tekstach. Dla wartości

liczby tematów zbliżonej do liczby opowiadań (np. 3–4) obserwowaliśmy, że tematy rzeczywiście grupują się wokół słów charakterystycznych dla danego opowiadania – np. w jednym temacie dominowały słowa takie jak "Cthulhu", "cult", "dream", w innym "Dunwich", "Whateley", "ritual", a w jeszcze innym "Innsmouth", "marsh", "deep".

Choć nie każdy dokument zawierał tylko jeden temat (co jest naturalne w LDA), dało się zauważyć wyraźne powiązania semantyczne między tematami a opowiadaniami, co sugeruje, że model LDA potrafi częściowo odzwierciedlić ukryte struktury tematyczne pokrywające się z rzeczywistym podziałem tekstów.

Tematy charakterystyczne dla książki 'The Call of Cthulhu':

Temat 7: dream, cult, professor, cthulhu, legrasse, city, star, uncle, wilcox, sea

Temat 2: uncle, manuscript, armitage, shall, wilbur, dream, young, case, wilcox, letter

Temat 1: innsmouth, street, door, marsh, town, cross, room, bus, road, window

Tematy charakterystyczne dla książki 'The Dunwich Horror':

Temat 3: glen, frye, wilbur, bishop, armitage, big, corey, boy, tree, rud

Temat 4: armitage, wilbur, yog, sothoth, whippoorwill, window, copy, monstrous, shall, goatish

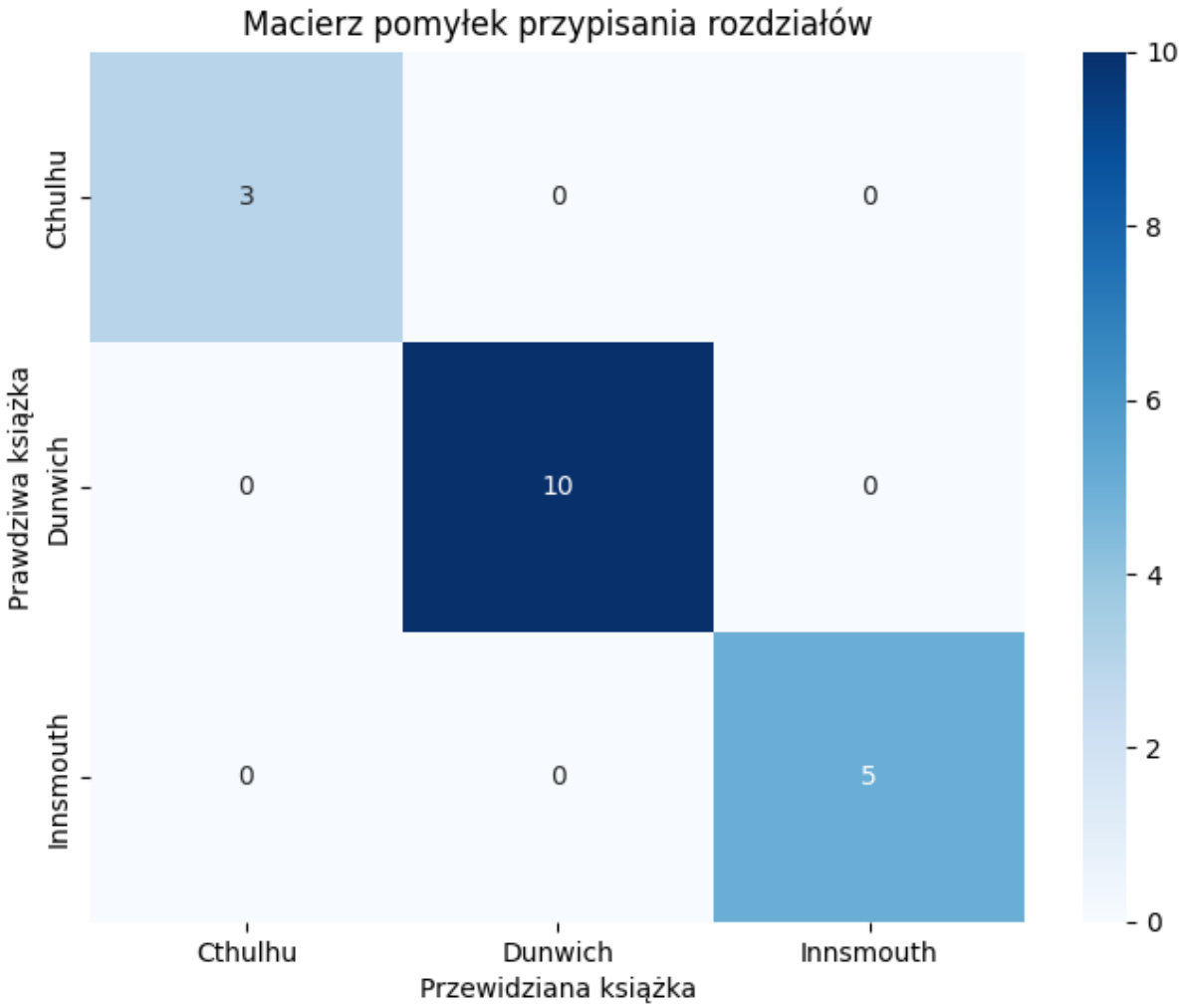
Temat 2: uncle, manuscript, armitage, shall, wilbur, dream, young, case, wilcox, letter

Tematy charakterystyczne dla książki 'The Shadow over Innsmouth':

Temat 1: innsmouth, street, door, marsh, town, cross, room, bus, road, window

Temat 0: folk, fer, get, water, fish, git, island, abaout, sea, daown

Temat 2: uncle, manuscript, armitage, shall, wilbur, dream, young, case, wilcox, letter



Dokładność klasyfikacji: 100.00%

```
Szczegółowe wyniki przypisania:
1. The Horror in Clay. (Cthulhu) => Cthulhu ✓ (Cthulu: 0.796, Dunwich: 0.312, Insmouth: 0.066)
2. The Tale of Inspector Legrasse. (Cthulhu) => Cthulhu ✓ (Cthulu: 0.977, Dunwich: 0.001, Insmouth: 0.000)
3. The Madness from the Sea. (Cthulhu) => Cthulhu ✓ (Cthulu: 0.977, Dunwich: 0.001, Insmouth: 0.000)
1 (Dunwich) => Dunwich ✓ (Cthulu: 0.001, Dunwich: 0.205, Insmouth: 0.001)
2 (Dunwich) => Dunwich ✓ (Cthulu: 0.001, Dunwich: 0.613, Insmouth: 0.001)
3 (Dunwich) => Dunwich ✓ (Cthulu: 0.215, Dunwich: 0.409, Insmouth: 0.087)
4 (Dunwich) => Dunwich ✓ (Cthulu: 0.001, Dunwich: 0.613, Insmouth: 0.001)
5 (Dunwich) => Dunwich ✓ (Cthulu: 0.001, Dunwich: 0.612, Insmouth: 0.001)
6 (Dunwich) => Dunwich ✓ (Cthulu: 0.001, Dunwich: 0.612, Insmouth: 0.001)
7 (Dunwich) => Dunwich ✓ (Cthulu: 0.000, Dunwich: 0.613, Insmouth: 0.000)
8 (Dunwich) => Dunwich ✓ (Cthulu: 0.215, Dunwich: 0.409, Insmouth: 0.087)
9 (Dunwich) => Dunwich ✓ (Cthulu: 0.000, Dunwich: 0.613, Insmouth: 0.000)
10 (Dunwich) => Dunwich ✓ (Cthulu: 0.000, Dunwich: 0.205, Insmouth: 0.198)
I (Innsmouth) => Innsmouth ✓ (Cthulu: 0.000, Dunwich: 0.001, Insmouth: 0.976)
II (Innsmouth) => Innsmouth ✓ (Cthulu: 0.000, Dunwich: 0.001, Insmouth: 0.976)
III (Innsmouth) => Innsmouth ✓ (Cthulu: 0.000, Dunwich: 0.198, Insmouth: 0.451)
IV (Innsmouth) => Innsmouth ✓ (Cthulu: 0.000, Dunwich: 0.001, Insmouth: 0.976)
V (Innsmouth) => Innsmouth ✓ (Cthulu: 0.100, Dunwich: 0.190, Insmouth: 0.906)
```

Tematy charakterystyczne dla każdej książki

- The Call of Cthulhu: Tematy mają silny związek z kluczowymi motywami tej powieści: „dream", „cult", „professor", „cthulhu", „legrasse", „city", „star". Widać tu odniesienia do snów, kultów, postaci (np. profesor Legrasse), oraz oczywiście samego Cthulhu. Drugi temat zawiera słowa takie jak „uncle", „manuscript", „armitage", co może wskazywać na wątki związane z dokumentami, postaciami i tajemnicami. Trzeci temat skupia się na, „street", „door" – tutaj widoczna jest pewna przestrzenna lub lokalna narracja.

- The Dunwich Horror: Tutaj tematy zawierają imiona i miejsca związane z tą historią: „glen", „frye", „wilbur", „bishop", „armitage", „yog", „sothoth". Słowa takie jak „monstrous", „goatish" oraz „whippoorwill" dobrze oddają klimat grozy i mitologii. Temat 2 z „uncle", „manuscript" i „armitage" pojawia się też w Cthulhu i Innsmouth, co sugeruje pewien wspólny kontekst lub motywy w tekstach.

- The Shadow over Innsmouth: Tematy skupiają się na nazwie miasta („innsmouth"), elementach otoczenia („street", „marsh", „town", „road"), a także na słowach związanych z morzem i rybami („water", „fish", „island", „sea"), co idealnie oddaje morską i nadprzyrodzoną atmosferę tej książki. Ten sam temat 2 z „uncle", „manuscript" pojawia się ponownie.

Wspólne tematy i motywy

Interesujące jest to, że temat 2 (z wyrazami „uncle", „manuscript", „armitage" itd.) pojawia się we wszystkich trzech książkach. To sugeruje, że pewne motywy, postaci lub narracje są wspólne lub powtarzają się w tych tekstach, co jest typowe dla dzieł Lovecrafta, gdzie niektóre postacie i wątki się przeplatają.

Dokładność klasyfikacji

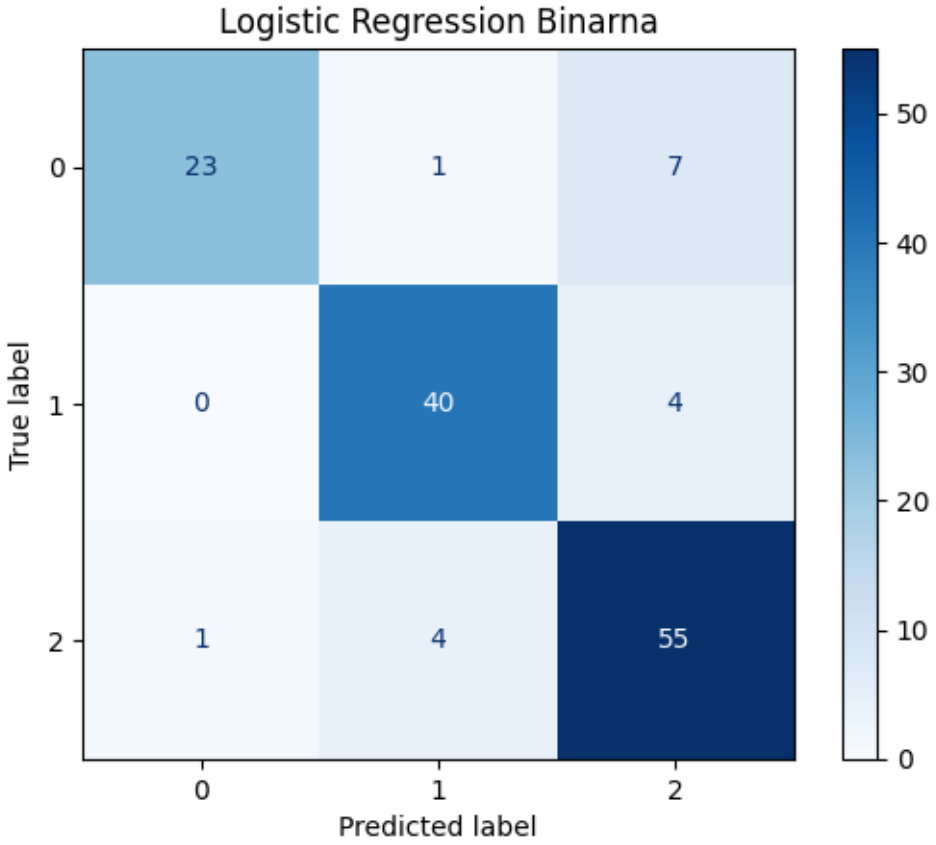
Klasyfikacja uzyskała 100% dokładności, co oznacza, że na podstawie tematycznej analizy LDA udało się idealnie przypisać fragmenty tekstów do ich właściwych książek.

Podsumowanie

- Model LDA skutecznie wyodrębnił tematy odpowiadające charakterystycznym motywom każdej z trzech książek.
- Pewne motywy są wspólne dla wszystkich trzech, co odpowiada fabularnym powiązaniom Lovecraftowskich opowiadań.
- Wysoka dokładność klasyfikacji potwierdza, że analiza tematyczna dobrze separuje teksty na podstawie ich charakterystycznych treści.
- Małe rozbieżności w przypisaniu niektórych tekstów są naturalne, biorąc pod uwagę, że opowiadania mogą mieć nakładające się wątki i podobne motywy.

6. Klasyfikacja

W kroku czwartym celem było zbudowanie i porównanie trzech klasyfikatorów, których zadaniem było automatyczne przypisywanie akapitów do jednej z 3 książek H.P. Lovecrafta. Do klasyfikacji wykorzystaliśmy reprezentacje tekstów oparte na czterech różnych sposobach ważenia częstości słów: binarnym, logarytmicznym, TfIdf oraz word embeddings. W każdym przypadku dokumenty zostały przekształcone do postaci wektorowej, a następnie zasilono nimi trzy niezależne klasyfikatory zbudowane na wybranym algorytmie (Random Forest, SVM i Logistic Regression).



-- Logistic Regression Binarna --

	precision	recall	f1-score	support
Cthulhu	0.96	0.74	0.84	31
Dunwich	0.89	0.91	0.90	44
Innsmouth	0.83	0.92	0.87	60
accuracy			0.87	135
macro avg	0.89	0.86	0.87	135
weighted avg	0.88	0.87	0.87	135

Podsumowanie wyników:

- **Najlepsze wyniki (najwyższa dokładność i f1-score) uzyskane zostały z Logistic Regression i MLP na wagach binarnych, logarytmicznych oraz TfIdf.** Dokładność oscyluje wokół 0.84–0.87, a makro f1-score w podobnym zakresie (ok. 0.85–0.87).
- **Najgorsze wyniki dają klasyfikatory SVM, szczególnie dla klasy Cthulhu, gdzie recall jest stosunkowo niski (~0.55 dla wag binarnych, logarytmicznych i TfIdf).** Co wpływa na niższą ogólną skuteczność (ok. 0.80 i niżej).
- **Word embeddings (Word2Vec) dają wyraźnie słabsze wyniki dla wszystkich trzech klasyfikatorów, z dokładnością w przedziale 0.56–0.76 i słabszym f1-score (około 0.54–0.77).**

Wybór najlepszego klasyfikatora

Spośród wszystkich wyników, najlepsze osiągi i najbardziej stabilne wyniki mają klasyfikatory oparte na **Logistic Regression z binarnymi, logarytmicznymi i TfIdf wagami**. Dokładność 0.87 i macro f1-score około 0.86–0.87 świadczą o dobrym dopasowaniu modelu.

Najlepszym okazuje się jednak Logistic Regression z binarnymi wagam (dokładność 87%, f1-score ~0.87), ze względu na lekko lepsze wyniki w porównaniu do wersji logarytmicznej i TfIdf (bardzo podobne, ale binarna wersja ma minimalnie wyższe precision dla klasy Dunwich i f1 dla Cthulhu).

Interpretacja klasyfikatora Logistic Regression, wagi binarne

- **Precision (precyzja)** mówi nam, jak dużo przewidzianych przykładów danej klasy faktycznie do niej należy. Najwyższa jest dla Dunwich (0.89) i Cthulhu (0.96), co oznacza, że model rzadko błędnie oznacza paragrafy jako z tych książek.
- **Recall (czułość)** mówi, jak dużo rzeczywistych przykładów danej klasy zostało poprawnie wykrytych. Najniższy recall jest dla Cthulhu (0.74), co wskazuje, że model nie wykrywa część klasyfikuje błędnie jako inne książki. Dla Dunwich i Innsmouth recall jest wyższy (odpowiednio 0.91 i 0.92), więc te klasy są lepiej wykrywane.
- **F1-score** to kompromis między precision i recall. Najniższe jest dla Cthulhu (0.84), najwyższe dla Dunwich (0.90), co oznacza, że Dunwich jest najlepiej rozpoznawaną klasą.
- **Skuteczność całkowita (~87%)** to bardzo dobry wynik w zadaniu klasyfikacji tekstów, zwłaszcza że mamy trzy klasy i teksty literackie, które są podobne stylistycznie.

Wnioski dodatkowe

- Wagi binarne, mimo że prostsze, dały wyniki nie gorsze, a często lepsze niż bardziej złożone wagi TfIdf czy logarytmiczne.
- Słabsze wyniki word embeddings mogą wynikać z faktu, że klasyfikator bazuje na uśrednionych reprezentacjach semantycznych, które mogły nie uchwycić specyficznych cech odróżniających poszczególne książki.
- SVM wypada gorzej, co może być spowodowane doborem parametrów lub naturą danych (np. liniowa separowalność może być słabsza w przestrzeni cech używanych przez SVM).

Najlepiej wybrać **Logistic Regression na binarnych wagach cech** jako klasyfikator, który:

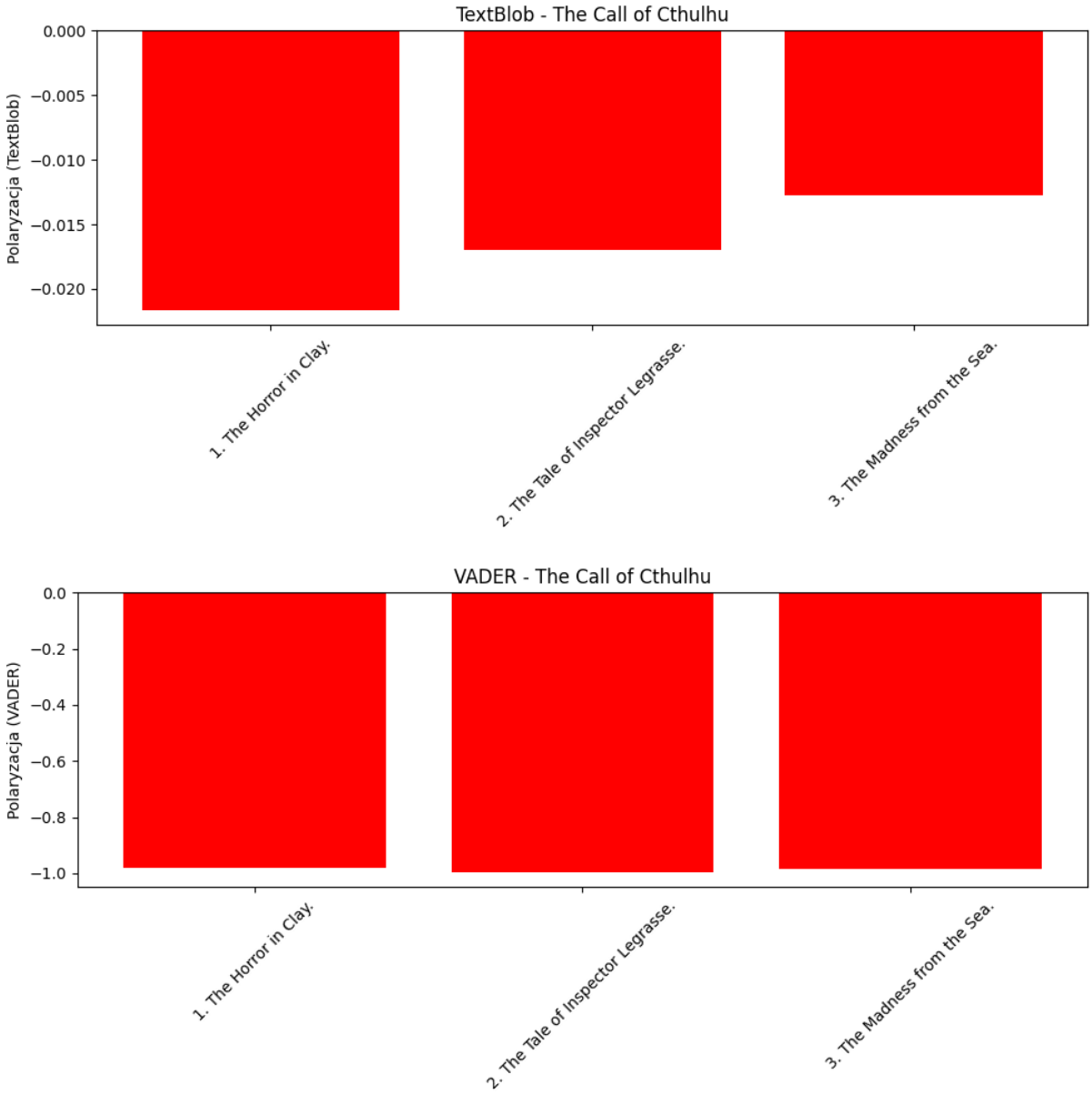
- osiąga najwyższą dokładność i dobry balans między precision a recall,
- dobrze radzi sobie z rozpoznawaniem każdej z trzech książek,
- jest prosty do interpretacji i efektywny obliczeniowo.

7. Analiza nastroju

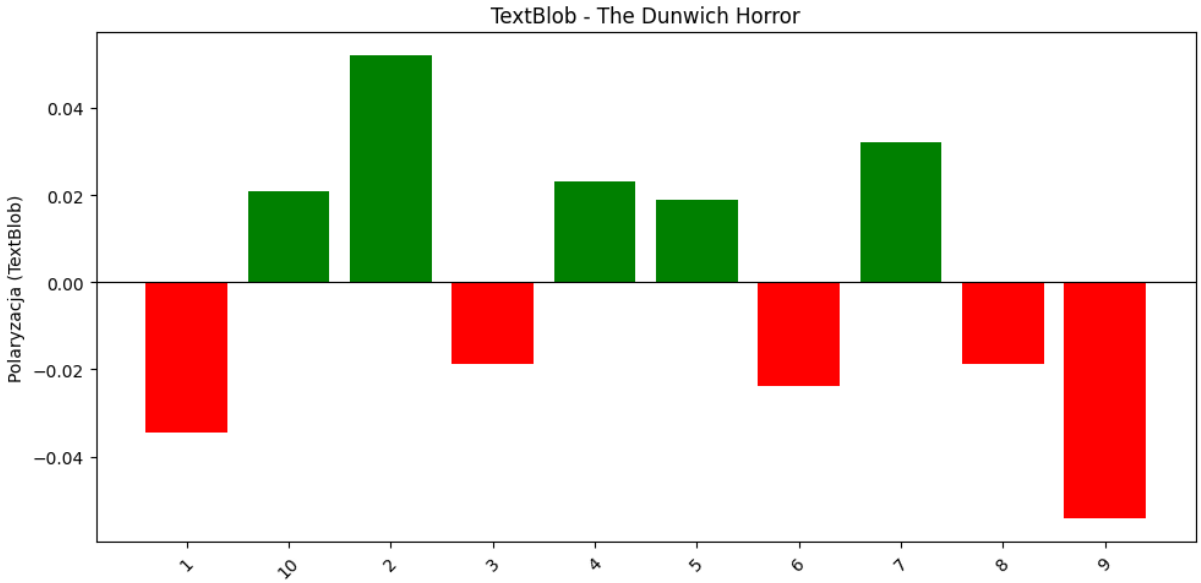
TextBlob i Vader

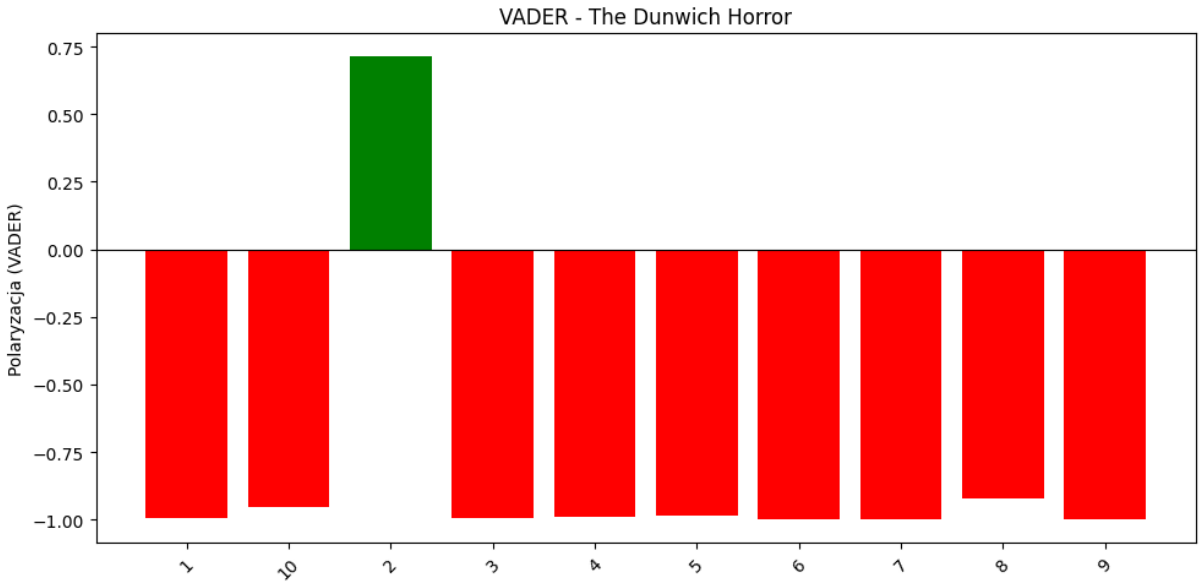
W analizie sentymentu trzech dzieł H.P. Lovecrafta: The Call of Cthulhu, The Dunwich Horror oraz The Shadow over Innsmouth wykorzystano dwie metody — TextBlob oraz VADER. Obie metody dają interesujące, choć częściowo sprzeczne wyniki, co wynika z różnic w ich podejściu do wykrywania emocji w tekście.

W przypadku The Call of Cthulhu, TextBlob wskazuje na bardzo neutralną polaryzację, bliską zeru, we wszystkich rozdziałach: pierwszy rozdział ma polaryzację -0,0217, drugi -0,0170, a trzeci -0,0127. Oznacza to, że tekst jest generalnie neutralny z lekkim odchyleniem ku negatywnemu, co pasuje do mrocznej tematyki horroru, choć TextBlob nie wykrywa silnej emocjonalności. Z kolei VADER pokazuje silnie negatywny charakter emocjonalny tekstu, z wynikami w zakresie od -0,9814 do -0,9982, co jest zgodne z fabułą opowieści o szaleństwie, kultach i kosmicznym horrorze.

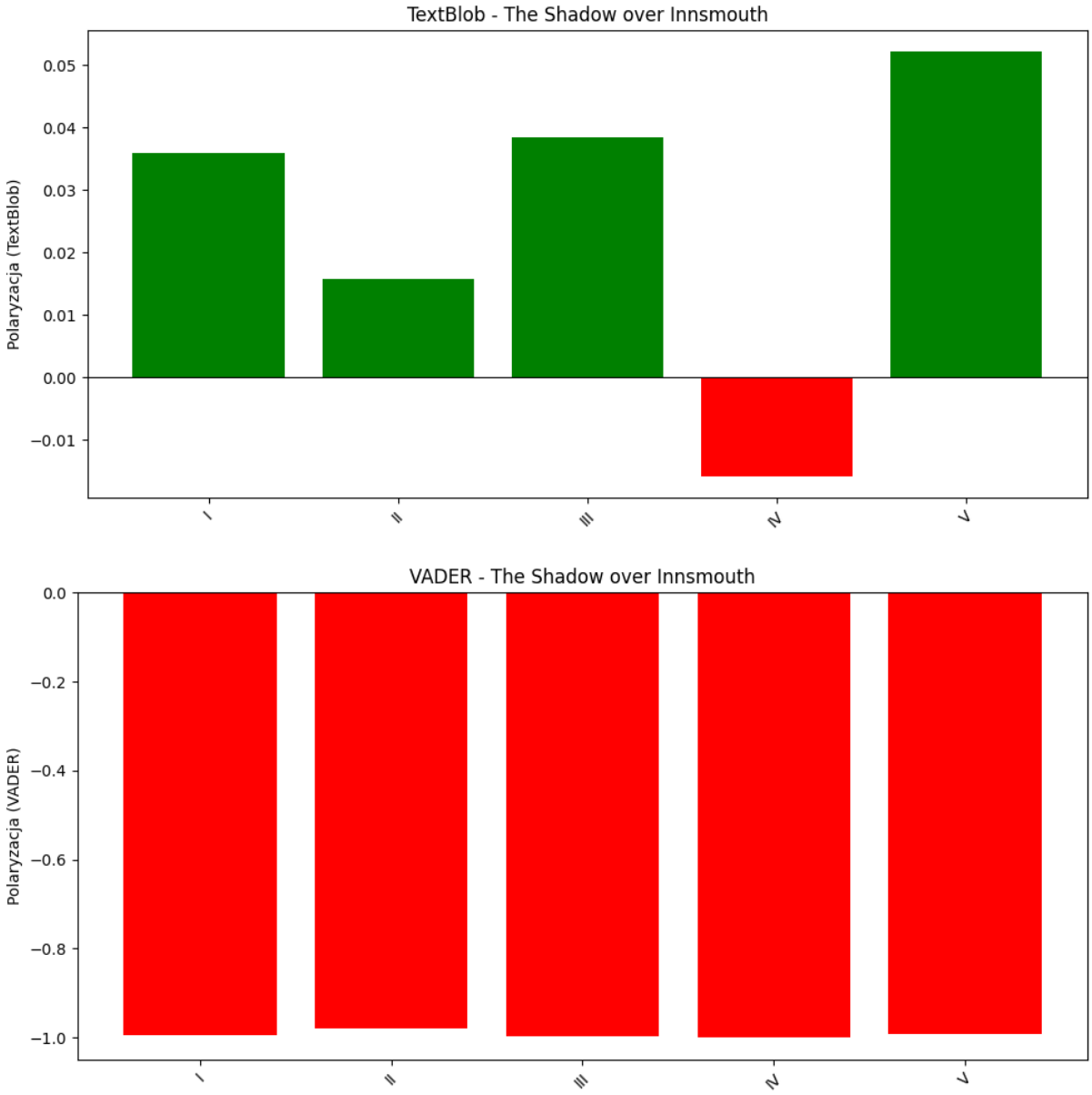


W przypadku The Dunwich Horror wyniki TextBlob są bardziej zróżnicowane, ale nadal oscylują wokół wartości zerowej, z przykładami rozdziałów: 2 (+0,0520) i 9 (-0,0541). Wahania te oznaczają, że TextBlob dostrzega lekki pozytyw lub negatyw, lecz nie wykrywa mocnych emocji, nawet gdy fabuła staje się mroczna. Natomiast VADER identyfikuje w większości rozdziałów silną negatywność, np. pierwszy rozdział osiąga wynik -0,9954, choć jest jeden wyjątek — rozdział drugi z wynikiem +0,7131, co może sugerować fragment mniej mroczny lub o bardziej neutralnym słownictwie. Ogólnie VADER wskazuje na obecność grozy i niepokoju.

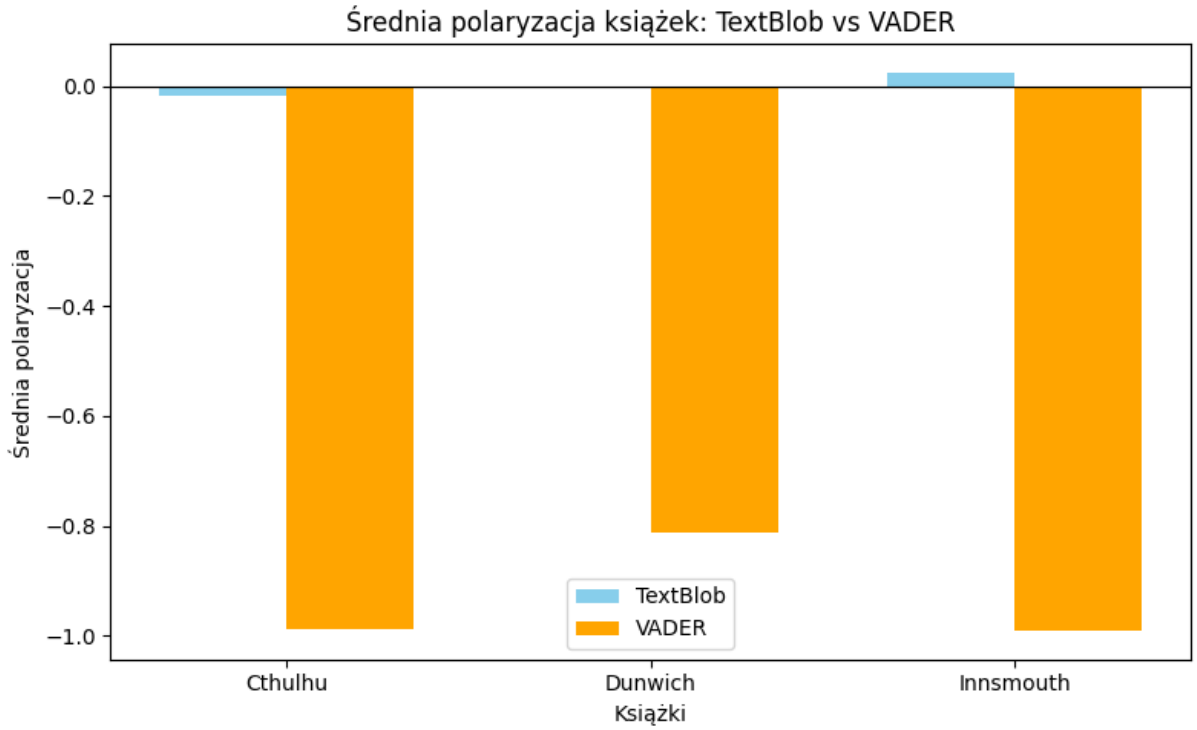




W przypadku The Shadow over Innsmouth TextBlob wykazuje lekkie pozytywne polaryzacje, np. rozdział pierwszy ma wynik +0,0360, a piąty +0,0521. W ten sposób metoda ta ocenia tekst jako lekko pozytywny, mimo że fabuła opowiada o przerażeniu i odkrywaniu koszmarnej prawdy. Z kolei VADER ponownie wskazuje na silnie negatywną emocjonalność w całym tekście, z wynikami bliskimi -1, np. rozdział czwarty (-0,9992) i piąty (-0,9919). Wyniki te dobrze odzwierciedlają tematykę degeneracji i fatalnego przeznaczenia bohaterów.



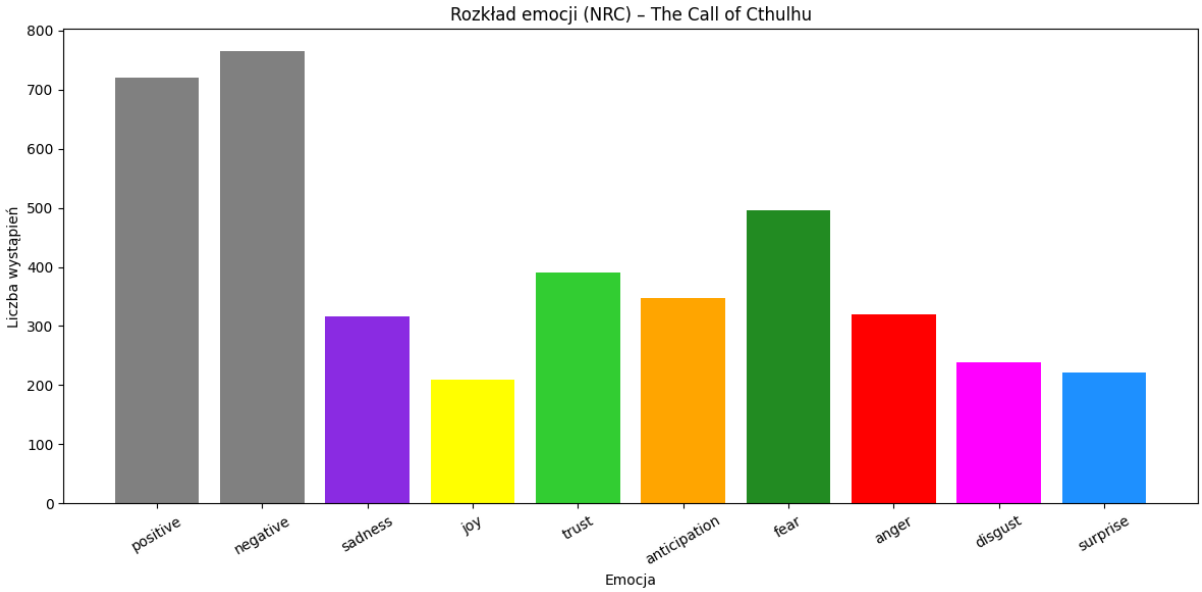
Podsumowując, TextBlob generalnie daje polaryzacje bliskie neutralności i nie wychwytuje głębokiej grozy czy niepokoju, co wskazuje na jego słabszą skuteczność w wykrywaniu subtelnej, literackiej grozy. Natomiast VADER konsekwentnie identyfikuje teksty Lovecrafta jako wybitnie negatywne emocjonalnie, co lepiej oddaje mroczny i niepokojący ton jego opowieści.



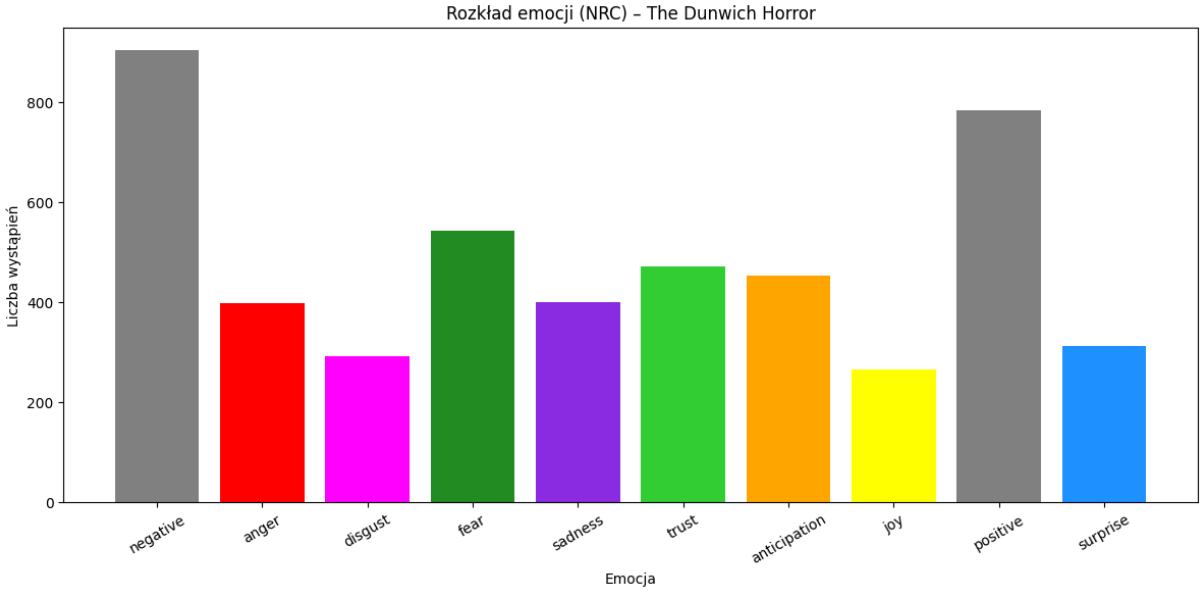
Dla kontekstu fabularnego, The Call of Cthulhu to opowieść o kultach, obłędzie i kosmicznym horrorze. The Dunwich Horror przedstawia historię narodzin potwora, śmierci i zniszczenia. Natomiast The Shadow over Innsmouth to podróż ku koszmarnej prawdzie i własnej degeneracji.

NRC-Emotion-Lexicon

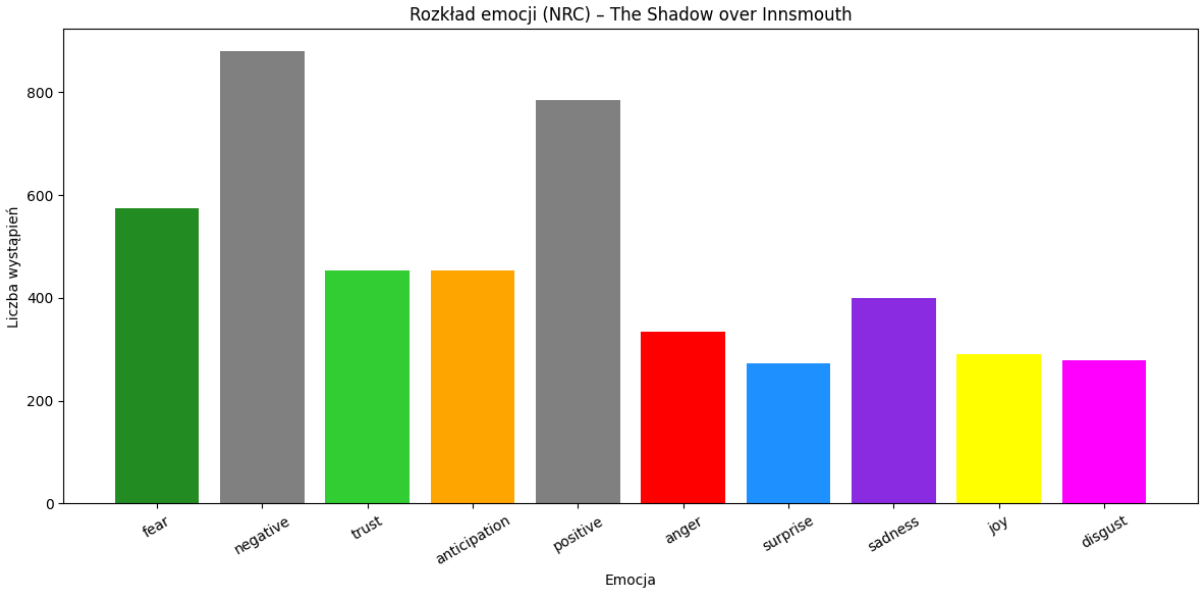
W **„The Call of Cthulhu”** dominują emocje negatywne, choć różnica między liczbą pozytywnych a negatywnych oznaczeń nie jest duża — 720 przypadków pozytywnych wobec 764 negatywnych. Strach (fear) okazuje się najczęściej występującą emocją, co świetnie wpisuje się w tematykę opowieści o kosmicznej grozie, niepojętych siłach i szaleństwie, które ogarnia bohaterów stykających się z tajemnicą Cthulhu. Co ciekawe, na drugim planie pojawia się zaufanie (trust), co może odzwierciedlać rolę wzajemnego wspierania się postaci w odkrywaniu tej przerażającej prawdy, a także wiarę w świadectwa i relacje świadków. Smutek, gniew i oczekiwanie (anticipation) pojawiają się również wyraźnie, co tworzy mieszankę emocji typową dla opowieści, gdzie groza przenika się z desperacją i niepewnością. Radość (joy) jest stosunkowo nieliczna, co podkreśla brak optymizmu w narracji.



W **„The Dunwich Horror”** proporcje są jeszcze bardziej przechylone w stronę negatywnego odbioru, z 903 przypadkami emocji negatywnych i 783 pozytywnych. Tutaj strach ponownie dominuje, co jest naturalne w historii o nadprzyrodzonym zagrożeniu, które budzi grozę wśród mieszkańców i zmusza ich do walki o przetrwanie. Bardzo silnie obecne są również gniew i obrzydzenie (anger i disgust), które odzwierciedlają niechęć, nienawiść i odrazę wobec Wilbura Whateleya i mrocznych sił zagrażających społeczności. Jednocześnie zaufanie oraz oczekiwanie są dość mocno obecne, co może wskazywać na współpracę ludzi przeciwko wspólnemu wrogowi i napięcie związane z próbą pokonania niewyobrażalnego zagrożenia. Radość, choć widoczna, pozostaje na marginesie emocjonalnego pejzażu tej opowieści.



W przypadku **„The Shadow over Innsmouth”** również obserwujemy przewagę emocji negatywnych (879) nad pozytywnymi (785). Najczęściej występującą emocją jest strach, co jest całkowicie spójne z fabułą, w której bohater stopniowo odkrywa przerażającą prawdę o mieście i o samym sobie. Smutek i gniew pojawiają się w związku z narastającym poczuciem osaczenia, odrazy i beznadziejności. Podobnie jak w dwóch pozostałych utworach Lovecrafta, zaufanie i oczekiwanie odgrywają istotną rolę – mogą odzwierciedlać nadzieje bohatera na ratunek lub wsparcie ze strony tych nielicznych, którym można zaufać. Mimo obecności pozytywnych emocji, całościowy obraz pozostaje przytłoczony grozą, fatalizmem i niepokojem.



Podsumowując, we wszystkich trzech utworach Lovecrafta mamy wyraźną dominację strachu jako emocji kluczowej dla budowania atmosfery kosmicznego horroru. Negatywna polaryzacja przeważa nad pozytywną, choć nie są to różnice przytłaczające – raczej subtelne, co pokazuje, że Lovecraft oprócz budowania grozy wplata w swoje historie elementy nadziei, lojalności czy napięcia związanego z walką o ocalenie. Rozkład emocji dobrze oddaje charakter tych opowieści: nieustające poczucie zagrożenia, bezradność wobec sił większych niż człowiek, ale też zmagania z losem i próbę stawienia czoła niewyobrażalnemu złu.

Ta analiza częściowo wyjaśnia wyniki TextBloba — mimo że zauważalna jest przewaga emocji negatywnych nad pozytywnymi, dla TextBloba różnica ta może być zbyt subtelna.