

Analiza Semantyczna i Ekstrakcja Kolokacji

Dependency Parsing w analizie tekstu literackiego

Przetwarzanie Języka Naturalnego – Zadanie 3

Robert Jacak, Maciej Jamroży

9 lutego 2026

Streszczenie

Niniejsze sprawozdanie przedstawia analizę semantyczną powieści *Moby Dick* Hermana Melville'a z wykorzystaniem dependency parsing (spaCy). Ekstrakcja kolokacji przymiotnik+rzeczownik i czasownik+rzeczownik pozwoliła na identyfikację dominujących wzorców językowych tekstu. Wyniki zwizualizowano jako grafy dwudzielne oraz udostępniono przez interaktywne CLI z kolorową skalą częstości.

Spis treści

| | |
|---|----------|
| 1 Wstęp | 3 |
| 2 Metodologia | 3 |
| 2.1 Narzędzia | 3 |
| 2.2 Przetwarzanie tekstu | 3 |
| 2.3 Dependency Parsing | 3 |
| 3 Analiza kolokacji przymiotnikowych | 4 |
| 3.1 Ranking Top 20 | 4 |
| 3.2 Interpretacja | 4 |
| 4 Analiza kolokacji czasownikowych | 5 |
| 4.1 Ranking Top 20 | 5 |
| 4.2 Interpretacja | 5 |
| 5 Wizualizacja – grafy dwudzielne | 6 |
| 5.1 Graf kolokacji przymiotnikowych | 6 |
| 5.2 Graf kolokacji czasownikowych | 7 |
| 6 Interaktywne CLI | 7 |
| 6.1 Skala kolorów | 7 |
| 6.2 Przykład sesji | 8 |
| 7 Wnioski | 8 |
| 7.1 Skuteczność metody | 8 |
| 7.2 Ograniczenia | 9 |
| 7.3 Możliwe rozszerzenia | 9 |

1 Wstęp

Celem projektu jest analiza semantyczna tekstu literackiego przy użyciu dependency parsing – metody pozwalającej na identyfikację zależności składniowych między słowami w zdaniu. W odróżnieniu od prostych metod n-gramowych, dependency parsing uwzględnia strukturę gramatyczną, co pozwala na ekstrakcję rzeczywistych kolokacji niezależnie od odległości między słowami.

Jako korpus wybrano powieść *Moby Dick* (1851) Hermana Melville'a, dostępną w domenie publicznej na Project Gutenberg. Tekst liczy 212 796 słów (2802 paragrafów) i reprezentuje bogaty, opisowy styl prozy XIX-wiecznej, obfitującej w rozbudowane konstrukcje przymiotnikowe i złożone relacje czasownik–dopełnienie. Tematyka morska, wielorybnicza i filozoficzna sprawia, że tekst jest szczególnie interesujący pod kątem analizy kolokacji.

Analiza obejmuje dwa typy relacji:

- **przymiotnik + rzeczownik** (relacja `amod`) – np. *white whale, old man*
- **czasownik + rzeczownik** (relacje `dobj` i `nsubj`) – np. *see whale, man say*

2 Metodologia

2.1 Narzędzia

- **spaCy** (model `en_core_web_sm`) – biblioteka NLP do tokenizacji, lematyzacji i dependency parsing
- **NetworkX** – biblioteka do tworzenia i analizy grafów
- **Matplotlib** – wizualizacja grafów dwudzielnich
- **colorama** – kolorowe wyjście CLI

2.2 Przetwarzanie tekstu

Tekst pobierany jest z Project Gutenberg i cachowany lokalnie. Następnie usuwane są nagłówki i stopki specyficzne dla Gutenberg (markery ***** START/END OF THE PROJECT GUTENBERG EBOOK**). Tekst jest normalizowany pod względem białych znaków, z zachowaniem interpunkcji potrzebnej do poprawnego parsowania.

2.3 Dependency Parsing

Oczyszczony tekst dzielony jest na paragrafy (separatorem `\n\n`) i przetwarzany wsadowo przez `nlp.pipe()` z rozmiarem batcha 50, co zapewnia wydajną obsługę dużych tekstów.

Dla każdego tokenu analizowana jest relacja syntaktyczna (`dep_`):

- `amod` (adjectival modifier) – token jest przymiotnikiem, jego `head` jest rzeczownikiem. Tworzy parę (przymiotnik, rzeczownik).
- `dobj` (direct object) – token jest rzeczownikiem-dopełnieniem, jego `head` jest czasownikiem. Tworzy parę (czasownik, rzeczownik).

- **nsubj** (nominal subject) – token jest rzeczownikiem-podmiotem (filtrowany dodatkowo przez `pos_ == NOUN`), jego `head` jest czasownikiem. Tworzy parę (czasownik, rzeczownik).

Wszystkie słowa są lematyzowane i konwertowane do małych liter. Filtr `.isalpha()` odrzuca tokeny zawierające znaki niealfabetyczne.

W wyniku analizy wyekstrahowano **12 033** unikalnych par przymiotnik+rzeczownik oraz **9 485** unikalnych par czasownik+rzeczownik.

3 Analiza kolokacji przymiotnikowych

3.1 Ranking Top 20

Poniżej przedstawiono 20 najczęściej występujących przymiotników wraz z ich najczęstszymi partnerami-rzeczownikami:

Tabela 1: Top 20 kolokacji przymiotnik + rzeczownik

| Ranga | Przymiotnik | Łączna częst. | Główni partnerzy |
|-------|-------------|---------------|------------------------|
| 1 | old | 323 | man, time, age |
| 2 | other | 283 | thing, whale, boat |
| 3 | great | 269 | whale, part, leviathan |
| 4 | same | 173 | time, way, instant |
| 5 | good | 165 | time, deal, luck |
| 6 | long | 164 | time, line, row |
| 7 | own | 163 | hand, person, heart |
| 8 | such | 163 | time, thing, man |
| 9 | white | 145 | whale, water, man |
| 10 | first | 124 | time, place, man |
| 11 | little | 123 | man, negro, craft |
| 12 | small | 117 | degree, fish, whale |
| 13 | many | 115 | thing, year, case |
| 14 | whole | 107 | world, life, length |
| 15 | last | 88 | night, year, man |
| 16 | large | 77 | whale, ship, number |
| 17 | strange | 76 | thing, sight, captain |
| 18 | wild | 73 | cry, wind, beast |
| 19 | full | 72 | ship, length, whale |
| 20 | poor | 70 | fellow, devil, lad |

3.2 Interpretacja

Wyniki doskonale odzwierciedlają charakter tekstu Melville'a:

- **Dominacja *old* (323), *other* (283), *great* (269)** – centralne motywów powieści: starość (Ahab, doświadczeni marynarze – *old man*[84]), wielkość (*great whale*[18]), różnorodność świata morskiego.

- **Kolokacja *white whale*[26]** – najbardziej specyficzna para dla tego korpusu, bezpośrednio identyfikująca tytułowego Moby Dicka. Przymiotnik *white* (145 wystąpień) łączy się też z *water*[8] i *man*[7].
- **Język opisowy** – przymiotniki takie jak *strange* (76), *wild* (73), *poor* (70) budują mroczną, dramatyczną atmosferę powieści (*wild cry*[5], *poor fellow*[11]).
- **Tematyka morska** – rzeczowniki *whale*, *ship*, *fish* pojawiają się jako partnerzy wielu przymiotników.

4 Analiza kolokacji czasownikowych

4.1 Ranking Top 20

Tabela 2: Top 20 kolokacji czasownik + rzeczownik

| Ranga | Czasownik | Łączna częst. | Główni partnerzy |
|-------|-----------|---------------|------------------------|
| 1 | be | 1270 | whale, thing, man |
| 2 | have | 489 | whale, man, look |
| 3 | take | 233 | turn, place, man |
| 4 | make | 206 | man, passage, whaleman |
| 5 | seem | 177 | man, thing, ship |
| 6 | see | 174 | whale, ye, world |
| 7 | give | 159 | chase, way, order |
| 8 | come | 139 | time, man, darkness |
| 9 | go | 133 | man, whale, ship |
| 10 | do | 132 | bidding, man, ye |
| 11 | say | 95 | ye, word, landlord |
| 12 | stand | 83 | man, harpooneer, tree |
| 13 | get | 81 | chance, peep, start |
| 14 | turn | 70 | head, thunder, face |
| 15 | know | 69 | one, ye, man |
| 16 | hold | 68 | peace, head, mate |
| 17 | hear | 67 | voice, ye, talk |
| 18 | keep | 64 | eye, ship, out |
| 19 | find | 62 | place, doubloon, house |
| 20 | show | 60 | visitor, face, tooth |

4.2 Interpretacja

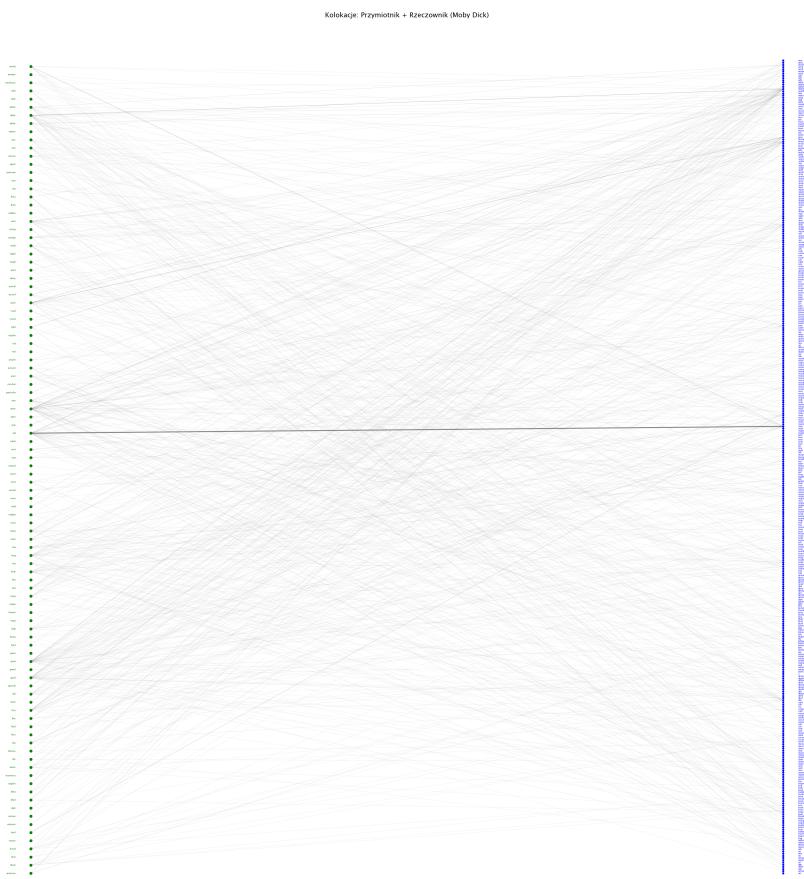
- **be dominuje (1270)** – jako kopula łączy się z podmiotami: *whale*[59], *thing*[31], *man*[26], *ship*[24] – odzwierciedla opisowy, encyklopedyczny styl Melville'a.
- **Czasowniki percepacji** (*see* 174, *seem* 177, *hear* 67) – odzwierciedlają obserwacyjny charakter narracji Ishmaela, który opisuje świat z pozycji świadka.
- **Czasowniki akcji** (*take* 233, *make* 206, *give* 159) – związane z pracą na statku: *give chase*[11], *take turn*[8], *make passage*[5].

- *whale* jako uniwersalny partner – pojawia się przy czasownikach *be*[59], *have*[13], *see*[12], *go*[6], co potwierdza centralną rolę wieloryba w fabule.

5 Wizualizacja – grafy dwudzielne

Wyniki zostały zwizualizowane jako grafy dwudzielne, gdzie lewa strona reprezentuje słowa kluczowe (przymiotniki lub czasowniki), a prawa strona – rzeczowniki. Grubość i przezroczystość krawędzi są proporcjonalne do częstości kolokacji.

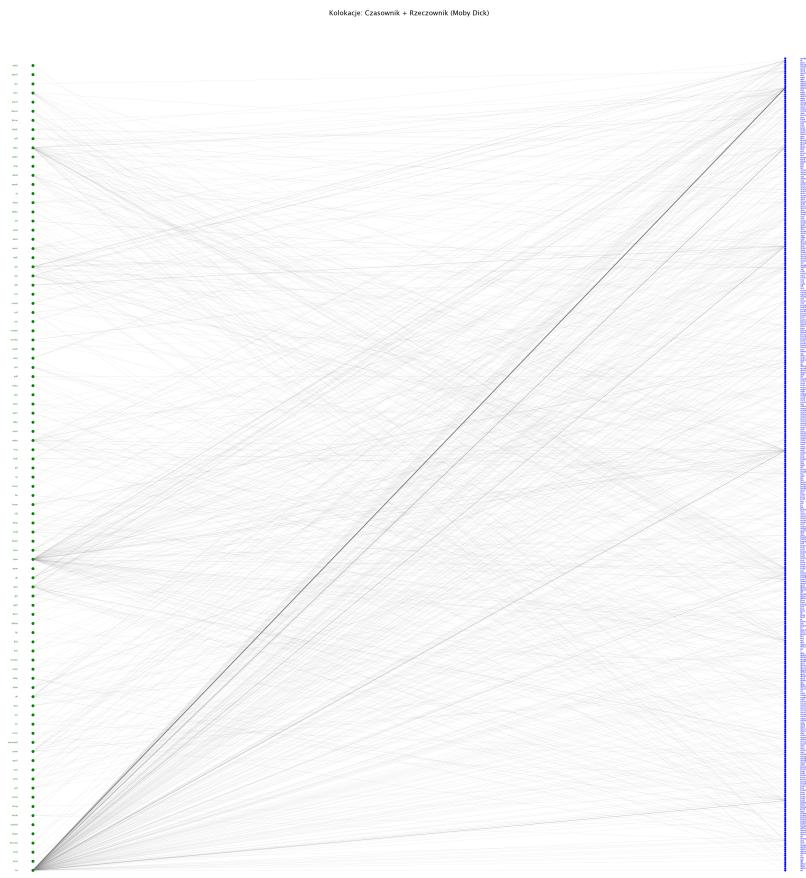
5.1 Graf kolokacji przymiotnikowych



Rysunek 1: Graf dwudzielny: przymiotnik + rzeczownik (top 100 przymiotników, waga ≥ 2). Zielone węzły – przymiotniki, niebieskie – rzeczowniki.

Na grafie widoczne są wyraźne klastry wokół centralnych rzeczowników: *whale*, *man*, *sea*. Przymiotnik *white* tworzy szczególnie silne połaczenie z *whale*, co odpowiada tytułowemu Moby Dickowi.

5.2 Graf kolokacji czasownikowych



Rysunek 2: Graf dwudzielny: czasownik + rzeczownik (top 100 czasowników, waga ≥ 2). Zielone węzły – czasowniki, niebieskie – rzeczowniki.

Rzeczowniki *whale* i *man* stanowią centralne węzły grafu, łączące się z wieloma czasownikami. Wskazuje to na ich dominującą rolę w strukturze narracyjnej tekstu.

6 Interaktywne CLI

Skrypt udostępnia interaktywne CLI do przeszukiwania wyekstrahowanych kolokacji. Użytkownik może:

- Wpisać **2 słowa** – sprawdzenie częstości konkretnej pary w obu typach relacji
- Wpisać **1 słowo** – wyszukanie wszystkich partnerów danego słowa (jako keyword i jako noun)

6.1 Skala kolorów

Wyniki są kolorowane według częstości wystąpień:

0 wystapien -- CZERWONY (nie występuje)
1 wystapienie -- ŻÓŁTY (rzadkie)

2-10 wystapien -- ZIELONY (umiarkowane)
>10 wystapien -- NIEBIESKI (częste)

6.2 Przykład sesji

```
> white whale
Para: 'white' + 'whale'
adj+noun: 26 częste
verb+noun: 0 nie występuje
Łącznie: 26 częste

> whale
Szukam kolokacji dla: 'whale'

Jako rzeczownik (adj+noun) - top 15 przymiotników:
white: 26
great: 18
other: 13
large: 7
small: 4

Jako rzeczownik (verb+noun) - top 15 czasowników:
be: 59
have: 13
see: 12

> old
Szukam kolokacji dla: 'old'

Jako przymiotnik (adj+noun) - top 15:
man: 84
time: 10
age: 8

> quit
Do widzenia!
```

7 Wnioski

7.1 Skuteczność metody

Dependency parsing okazał się znacznie skuteczniejszy niż proste metody n-gramowe w ekstrakcji kolokacji:

- **Niezależność od odległości** – relacja amod łączy przymiotnik z rzeczownikiem nawet gdy między nimi znajdują się inne słowa (np. “*the great white sperm whale*”).
- **Rozróżnianie ról syntaktycznych** – oddzielenie podmiotu (`nsubj`) od dopełnienia (`dobj`) daje bogatszy obraz relacji.

- **Lematyzacja** – sprowadzenie słów do form bazowych (np. “*whales*” → “*whale*”) konsoliduje statystyki.

7.2 Ograniczenia

- Model `en_core_web_sm` to mały model – większe modele (np. `en_core_web_lg`) mogą dawać dokładniejsze parsowanie.
- Analiza nie uwzględnia negacji ani modalności – “*no white whale*” jest liczone tak samo jak “*white whale*”.
- Archaiczny język XIX-wieczny może powodować błędy w parsowaniu współczesnym modelem NLP.
- Tekst w języku angielskim – analiza polskojęzycznego tekstu wymagałaby innego modelu.

7.3 Możliwe rozszerzenia

- Analiza kolokacji z uwzględnieniem rozdziałów – śledzenie zmian kolokacji w toku narracji.
- Porównanie z innymi powieściami tego samego okresu.
- Wykorzystanie większych modeli spaCy lub transformerów (np. spaCy z backendem transformer).
- Analiza sentymenu kolokacji – czy dane pary pojawiają się w kontekstach pozytywnych czy negatywnych.

8 Bibliografia

1. spaCy – Industrial-Strength Natural Language Processing. <https://spacy.io/>
2. Project Gutenberg – *Moby Dick* by Herman Melville. <https://www.gutenberg.org/ebooks/2701>
3. NetworkX – Network Analysis in Python. <https://networkx.org/>
4. Matplotlib – Visualization with Python. <https://matplotlib.org/>
5. Melville, H. (1851). *Moby-Dick; or, The Whale*. Harper & Brothers.