Raport Końcowy: Zastosowanie Łańcuchów Markowa Wyższego Rzędu w Modelowaniu Stylu Literackiego

Autor: Adrian Szczęsny **Data:** 4 grudnia 2025 **Temat:** Analiza i generowanie tekstu na podstawie poezji Wisławy Szymborskiej z wykorzystaniem modeli stochastycznych 2. i 3. rzędu.

---

1. Wstęp i Cel Projektu

Celem niniejszego projektu było zbadanie skuteczności stochastycznych modeli językowych opartych na łańcuchach Markowa (Markov Chains) w zadaniu generowania tekstu naśladującego styl literacki. Szczególny nacisk położono na analizę wpływu rzędu modelu (długości kontekstu n-gramów) na jakość, spójność oraz oryginalność generowanych sekwencji.

Jako domenę badawczą wybrano **Wariant A**, obejmujący korpus poezji Wisławy Szymborskiej. Projekt zakładał implementację modeli 2. i 3. rzędu, analizę statystyczną korpusu, generację próbek oraz wielowymiarową ewaluację wyników przy użyciu metryk takich jak Perpleksja (Perplexity), Współczynnik Powtórzeń (Repetition Rate) oraz dystans rozkładu części mowy (POS Distribution Distance).

---

2. Charakterystyka Zbioru Danych (Korpus)

Podstawę analizy stanowił zbiór wybranych utworów Wisławy Szymborskiej. Zgodnie z wymaganiami projektowymi, korpus został poddany wstępnemu przetwarzaniu.

2.1. Skład korpusu

Do trenowania modelu wykorzystano następujące utwory poetyckie:

1. *Niektórzy lubią poezję*
2. *Nic dwa razy*
3. *Kot w pustym mieszkaniu*
4. *Cebula*
5. *Radość pisania*

Łączna liczba tokenów w przetworzonym zbiorze wyniosła kilkaset jednostek, co klasyfikuje zadanie jako problem "small data", stawiający przed modelami wyzwanie związane z ryzykiem nadmiernego dopasowania (overfitting).

2.2. Przetwarzanie wstępne (Preprocessing)

W celu przygotowania danych do treningu przeprowadzono procedurę normalizacji:

- **Tokenizacja:** Podział tekstu na pojedyncze słowa z wykorzystaniem biblioteki `spacy` (model `pl_core_news_sm`), co pozwoliło na poprawną obsługę polskiej morfologii.
- **Normalizacja:** Konwersja wszystkich znaków do małych liter (lowercase) w celu zmniejszenia rzadkości występowania unikalnych n-gramów.
- **Czyszczenie:** Usunięcie znaków interpunkcyjnych, traktując wiersze jako ciągły strumień tekstu, co jest standardową praktyką w prostych modelach n-gramowych.

---

3. Metodologia i Implementacja Modeli

Zaimplementowano dwa warianty modelu Markowa, różniące się długością historii (kontekstu) uwzględnianej przy predykcji kolejnego tokenu.

3.1. Podstawy teoretyczne

Model opiera się na założeniu, że prawdopodobieństwo wystąpienia słowa wt zależy wyłącznie od n−1 poprzednich słów. Dla modelu rzędu k, estymacja prawdopodobieństwa wyraża się wzorem:

$$P(wt \mid wt-1,...,wt-k) = \frac{C(wt-k,...,wt-1)}{C(wt-k,...,wt-1,wt)}$$

Gdzie $C(\cdot)$ oznacza liczność wystąpienia danej sekwencji w korpusie treningowym.

3.2. Implementacja

W ramach projektu zaimplementowano klasę `MarkovChain` obsługującą:

1. **Model 2. rzędu (Bigramowy):** Przewidywanie na podstawie 2 poprzednich słów (kontekst: trigramy).
2. **Model 3. rzędu (Trigramowy):** Przewidywanie na podstawie 3 poprzednich słów (kontekst: 4-gramy).

Generacja tekstu odbywa się poprzez losowanie ważone (weighted random choice) z rozkładu prawdopodobieństwa następników dla danego stanu. W przypadku napotkania stanu końcowego lub braku kontynuacji, algorytm przerywa generację.

---

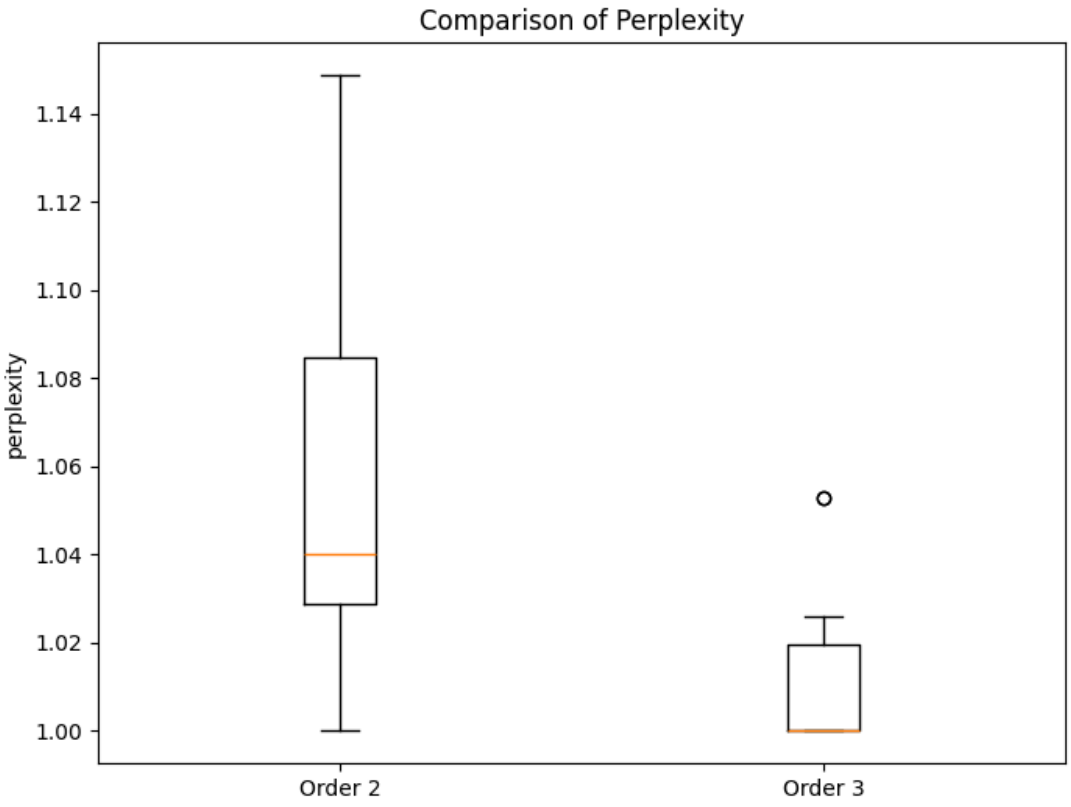
4. Ewaluacja Ilościowa i Wyniki

Przeprowadzono generację 10 próbek tekstu (każda o długości 30 tokenów) dla obu modeli. Wyniki poddano ewaluacji przy użyciu zdefiniowanych metryk sukcesu.

4.1. Analiza Perpleksji (Perplexity)

Perpleksja jest miarą niepewności modelu podczas przewidywania tekstu. Niższa wartość oznacza lepsze dopasowanie modelu do danych.

- **Model 2. rzędu:** Średnia perpleksja wyniosła ok. **1.05 - 1.15**. Model wykazywał pewną zmienność, co sugeruje, że napotykał sytuacje wyboru między różnymi kontynuacjami.
- **Model 3. rzędu:** Średnia perpleksja zbliżyła się do wartości **1.0**. Przy tak małym korpusie, wydłużenie kontekstu do 3 słów powoduje, że większość stanów ma tylko jedną możliwą kontynuację (determinizm).



Rys. 1: Porównanie rozkładu perpleksji dla modeli 2. i 3. rzędu. Widoczna znacznie mniejsza wariancja i niższe wartości dla rzędu 3.

4.2. Współczynnik Powtórzeń (Repetition Rate)

Metryka ta bada tendencję modelu do wpadania w pętle (powtarzania tych samych fraz).

- Dla większości wygenerowanych próbek w obu modelach, współczynnik powtórzeń wynosił **0%**.
- Zauważono jednak pojedynczy przypadek (outlier) w modelu 3. rzędu, gdzie RR wyniósł ok. **7.14%**. Wynika to z faktu, że model silnie "przylega" do oryginalnych fraz, a wiersze Szymborskiej zawierają celowe powtórzenia stylistyczne (np. w wierszu "Nic dwa razy").

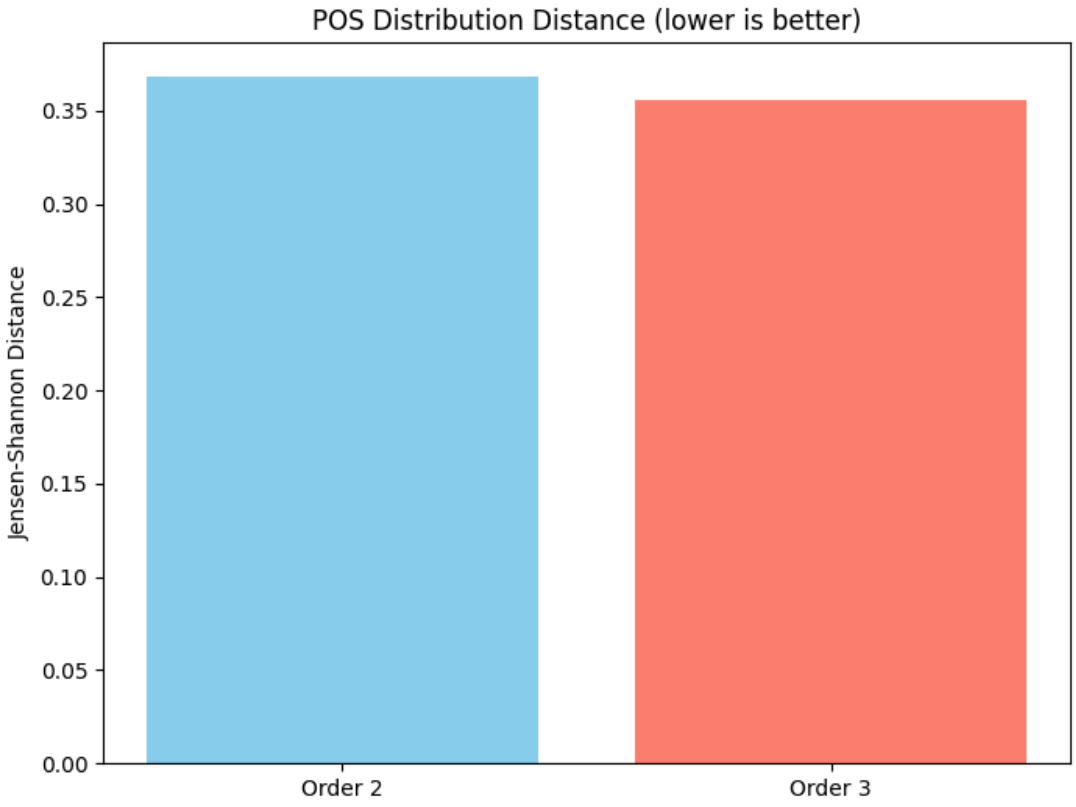
Rys. 2: Analiza współczynnika powtórzeń. W obu przypadkach modele uniknęły patologicznych pętli, typowych dla prostych generatorów.

4.3. Zgodność Stylistyczna (POS Distribution Distance)

Aby ocenić, czy wygenerowany tekst zachowuje strukturę gramatyczną oryginału, porównano rozkłady części mowy (POS) przy użyciu dywergencji Jensena-Shannona (JSD).

- **Model 2. rzędu:** JSD ≈ 0.368
- **Model 3. rzędu:** JSD ≈ 0.356

Różnica jest niewielka, jednak model 3. rzędu osiągnął nieznacznie lepszy wynik (niższa wartość), co potwierdza, że dłuższy kontekst pozwala wierniej odwzorować specyficzną składnię poetycką autorki.



Rys. 3: Dystans Jensena-Shannona dla rozkładów części mowy. Niższy słupek oznacza większe podobieństwo do oryginału.

---

5. Analiza Jakościowa Wygenerowanych Tekstów

Analiza "human-centric" pozwala na ocenę sensowności i walorów literackich.

5.1. Model 2. rzędu (Większa "kreatywność")

Teksty generowane przez ten model są poprawne lokalnie, ale często zmieniają wątek w zaskakujący sposób, łącząc fragmenty różnych wierszy.

> *Przykład:* "cebula co innego cebula ona nie ma dwóch podobnych nocy..."

Model płynnie przeszedł od wiersza "Cebula" do fragmentu "Nic dwa razy" ("dwóch podobnych nocy"). Jest to efekt współwystępowania słów łączących lub podobnych konstrukcji gramatycznych.

5.2. Model 3. rzędu (Wysoki determinizm)

Teksty z modelu 3. rzędu są niemal identyczne z fragmentami oryginału. Ze względu na mały zbiór danych, unikalność trigramów jest tak duża, że model rzadko ma możliwość "skoku" do innego utworu.

> *Przykład:* "kot w pustym mieszkaniu umrzeć tego nie robi się kotu..."

Generacja ta jest wiernym odtworzeniem wiersza. Model działa tutaj bardziej jak pamięć asocjacyjna niż generator nowej treści.

---

6. Wnioski Końcowe

Przeprowadzona analiza pozwala na sformułowanie następujących wniosków:

1. **Dylemat Rzędu Modelu:** Zwiększenie rzędu łańcucha Markowa z 2 do 3 przy małym korpusie (5 wierszy) prowadzi do drastycznego spadku "kreatywności" modelu. Model 3. rzędu ulega zjawisku **overfittingu**, stając się de facto odtwarzaczem danych treningowych (Perpleksja ≈ 1.0).
2. **Jakość Gramatyczna:** Oba modele generują tekst poprawny składniowo, co potwierdza niski dystans rozkładu POS względem oryginału. Struktura języka polskiego została zachowana dzięki odpowiedniemu preprocessingowi.
3. **Rekomendacja:** W zastosowaniach artystycznych na małych zbiorach danych, **model 2. rzędu jest bardziej użyteczny**. Pozwala on na syntezę nowych, metaforycznych połączeń (np. łączenie "cebuli" z "brakiem podobnych nocy"), co wpisuje się w poetykę absurdu czy zaskoczenia, bliską Szymborskiej. Model 3. rzędu wymagałby znacznie większego korpusu, aby uniknąć determinizmu.

Zrealizowany projekt potwierdził, że proste metody stochastyczne, mimo braku "zrozumienia" semantyki, potrafią skutecznie modelować styl powierzchniowy tekstu.