Generacja tekstów piosenek

Maciej Krzyżanowski, Sebastian Kutny, Tomasz Lewandowski

Kwiecień 2023

Spis treści

1 Wstęp

Celem projektu było stworzenie modelu generującego tekst piosenki na podstawie wybranych danych jako tekstów innych utworów. Wykorzystaliśmy 2 metody: łańcuchy Markova oraz rekurencyjne sieci neuronowe.

Projekt zawiera narzędzie "scraper"do pozyskiwania danych ze stron:

- `https://www.tekstowo.pl`
- `https://www.azlyrics.com`

Implementacja została wykonana w języku Python oraz wykorzystuje biblioteki:

- pandas
- BeautifulSoup
- nltk
- request
- queue
- re

Dostępna jest opcja łączenia zbiorów danych do jednego pliku w celu wykorzystania ich jednocześnie.

Przed rozpoczęciem przetwarzania danych są one oczyszczane poprzez ujednolcenie wielkości liter, usunięcie niepotrzebnych znaków interpunkcyjnych, słów zakazanych (np:. określających składowe tekstu utworu) oraz wyrażeń ze szczególnymi znakami interpunkcyjnymi jako określających zawartość tekstu.

Tekst jest generowany jako dowolna liczba wersów o dowolnej ilości słow.

2 Łańcuchy Markova

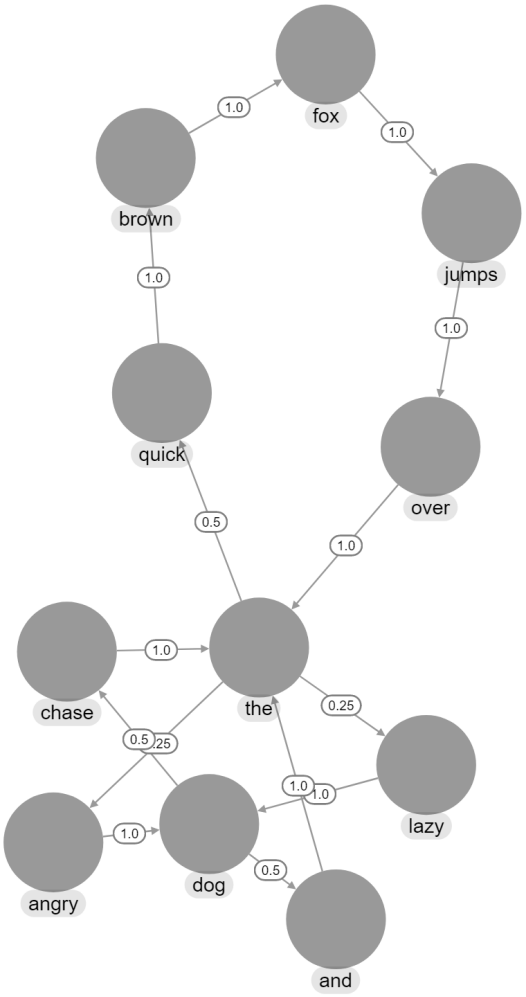
Łańcuchy Markova to matematyczny model służący do generowania tekstu lub sekwencji innych elementów, takich jak dźwięki lub obrazy. Ideą modelu jest analiza sekwencji istniejących elementów i wykorzystanie tych informacji do przewidywania, jakie elementy powinny pojawić się następnie.

W przypadku generowania tekstu, łańcuchy Markova są zwykle stosowane do analizy sekwencji słów w tekście źródłowym i generowania nowych sekwencji słów na podstawie tych informacji. Proces zaczyna się od wyboru deterministycznego lub losowego początkowego stanu łańcucha, a następnie generowania kolejnych stanów na podstawie informacji o pradwopodobieństwie wystąpienia po sobie stanów w analizowanym tekście w kontekście poprzednio wygenerowanych stanów. Przykładowo, jeśli w tekście źródłowym po słowie "generator"często pojawia się słowo "piosenek", to model łańcucha Markova przypisze wysokie prawdopodobieństwo wystąpienia słowa "piosenek"po słowie "generator".

Istnieją różne sposoby implementacji modelu łańcuchów Markova, ale zwykle opierają się one na analizie pewnej liczby poprzednich elementów, zwanej "stopniem"modelu. Na przykład, w przypadku modelu pierwszego stopnia, prawdopodobieństwo wystąpienia danego elementu zależy tylko od poprzedniego elementu, w modelu drugiego stopnia, prawdopodobieństwo zależy od dwóch poprzednich elementów, a w modelu trzeciego stopnia, prawdopodobieństwo zależy od trzech poprzednich elementów itd. Stopień łańcucha nazywamy N-gramem.

W naszym projektcie N-gram jest parametryzowany i bazuje na N uprzednio wygenerowanych słowach w wersie, losując następne słowo na podstawie prawdopodobieństwa jego wystąpienia po N sekwencji słów uprzednio wygenerowanych. Dodatkowo przy każdym wersie o nieparzystym numerze podejmowana jest próba stworzenia rymującego się wersu na podstawie ostatniej sylaby poprzedniego. Najpierw znajdowane są wszystkie rymujące się zakończenia wersu niebędące ostatnim słowem poprzedniego, a następnie - jeśli takie istnieją - losujemy jedno z nich zamiast z wszyskich pozycji. Przy nieznalezieniu rymujących się słów generacja odbywa się tak jak w zwykłym wypadku. W praktyce szansa na stworzenie rymu jest mała i powinna rosnąć z ilością danych przetwarzanych przez model.

Model łańcuchów Markova nie jest idealny i może generować sekwencje, które nie są sensowne lub poprawne gramatycznie. Dopiero przy wglądzie modeli w setki stanów wstecz oraz przy bardzo dużej ilości danych można wygenerować tekst podobny do pisanego przez człowieka.



Rysunek 1: Przykład łańcucha Markova, dla zdania "The quick brown fox jumps over the lazy dog and the angry dog chase the quick brown fox.", dla wartości $ngram = 1$, oznaczającej stany jako pojedyncze słowama oraz wartościami prawdopodobieństw przejść pomiędzy stanami obliczonych na podstawie zdania wejściowego.

```
Able one picks his broken down devotion i see pretty
Youd been and im a man oh oh like a
Cole and dick van pattern new york queen a bird
Fists of fury are cowardly now running through a self
Man meet me in deep space ask your mates but
Rough when things are looking good on the fame and
Young kentucky girl in a passing den no bloodiness no
But to say owt shes in the morning light the
Difference of right and i got to have some mail
Songs rap aint nothin but a sweet flower blossoms in
```

```
Ing yeah bird yeah someone took a gamble and risk
House cat got your facts all that ive been wakin
Southside muthafuckas get smoked i for the losers bless them
Dam wont be anybody after you i get lonely people
Kill registered at the ratio is the only thing i
Coupon my backwoods packs i aint trippin but i dont
Ghetto malukus in tha hoppin this ones a fuckin clique
School was alright i wan na talk to him when
Tag itwit more loot that jimmy got paid and repeat
Pretty sue i cant sing but ask for much too
```

Rysunek 2: Przykładowe wyniki generacji 10 wersów po 10 słów, kolejno dla zbiorów danych: *english_mixtape.csv* oraz *somemix.csv*