**Przetwarzanie Języka Naturalnego**

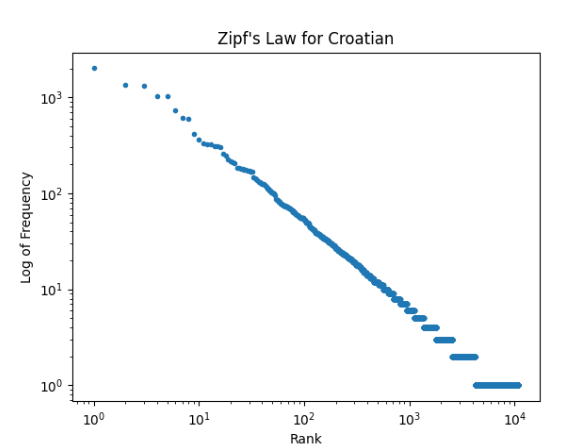
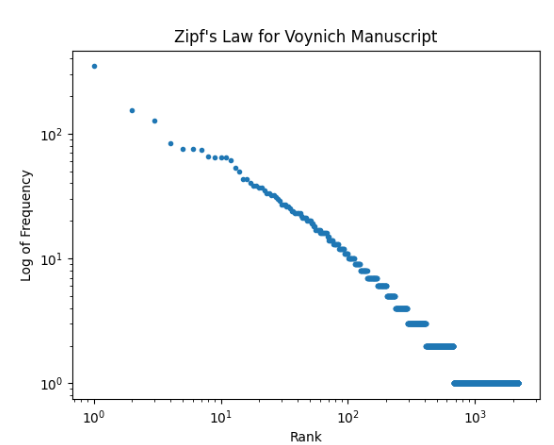
Natalia Niedziałek, Filip Gieracki

Projekt Zaliczeniowy

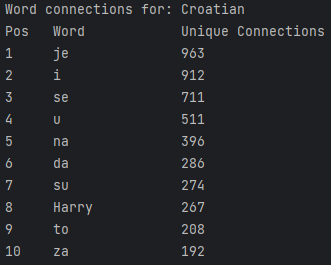
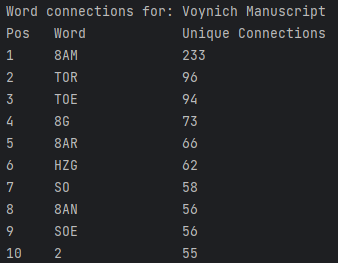
Celem projektu było badanie Manuskryptu Voynicha, jak również zapoznanie się z zależnościami językowymi. Do realizacji wspomnianych zadań został wykorzystany drugi język, przydzielony przez prowadzącego - chorwacki. W celu przeprowadzenia badań, wybrany został tekst “Harry Potter” w języku chorwackim, pierwsze 80 stron.

Projekt został podzielony na kilka etapów. Pierwszy obejmował pozyskanie oraz przygotowanie danych do analiz; w ramach tego kroku zostały usunięte wszelkie białe znaki oraz znaki interpunkcyjne, aby finalnie uzyskać listę czystych wyrazów, które stanowiły argument funkcji statystycznych.

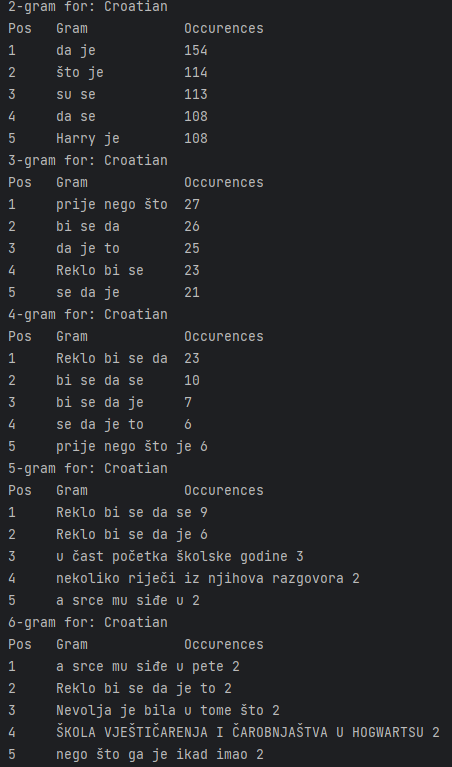
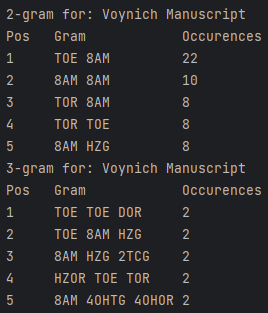
W kolejnym kroku przeprowadzona została analiza kolejno manuskryptu Voynicha jak również pozyskanego tekstu w języku chorwackim. Obejmowała ona zliczanie słów, wyliczanie ich pozycji jak również kalkulacje związane z wykresem prawa Zipfa, który prezentuje się następująco:



Kolejną częścią etapu analizy języków było utworzenie grafów dwudzielnych, w celu odnalezienia słów, które najczęściej powtarzają się w języku w sąsiedztwie innych, unikalnych słów. Ze względu na liczbę słów, walory prezentacyjne oraz czytelność, graf został przygotowany w postaci liczbowej; stworzona została lista 10 najczęstszych słów, wraz z liczbą “unikalnych połączeń” z innymi słowami w podanych językach.



Kolejny krok obejmował przygotowanie n-gramów, oraz zliczenie ich oraz wybranie takich, które pojawiły się w badanym tekście co najmniej 2 razy oraz przesortowanie ich i wyświetlenie 5. najbardziej popularnych.



dla N większych niż 3, dla Manuskryptu Voynicha - brak wyników

W przypadku tekstu w języku chorwackim - n-gramów jest znacznie więcej, dla maksymalnego badanego N - 10, nadal uzyskano wyniki.

Wnioski

Mając na uwadze niektóre przesłanki oraz częściowe wyniki analiz można dojść do wniosku, iż Manuskrypt Voynicha mógł zostać napisany rzeczywistym, sensownym językiem. Analiza wykresu prawa Zipfa częściowo potwierdza postawioną tezę, ze względu na bardzo duże podobieństwo wykresu do wykresu uzyskanego dla języka chorwackiego.

Analizując n-gramy można dojść do wniosku, że rozbieżność pomiędzy badanymi tekstami znacznie zwiększa się, lecz może to sugerować złożoność języka, w jakim został napisany Manuskrypt Voynicha.

W przypadku analizy diagramów dwudzielnych - tutaj zróżnicowanie jest znaczne, co może sugerować, iż Manuskrypt Voynicha nie jest tekstem posiadającym sens, znaczenie. Mając na uwadze poniższe obserwacje i badania, skłaniamy się ku tezie, iż manuskrypt Voynicha nie został napisany rzeczywistym, mającym sens językiem, jednakże nie wykluczamy jego ewentualnej sensowności.