Piotr Węsiora 06.01.2021

Komputerowe przetwarzanie tekstów - projekt

Badanie ewolucji języka na podstawie przemówień inauguracyjnych prezydentów Stanów Zjednoczonych Ameryki

Celem projektu było przeprowadzenie serii eksperymentów na tekstach przemów inauguracyjnych prezydentów USA w celu zaobserwowania zmian w języku angielskim. Wyniki badań mogą służyć jako element badań lingwistycznych bądź politologicznych. Wnioski wyciągnięte z testów mogą stanowić część dowodu na stawiane hipotezy z wyżej wymienionych nauk bądź jako podstawę do dalszych rozważań i badań.

W celu wykonania testów skorzystano z następujących narzędzi:

- Język programowania Python w wersji 3.8
- Zintegrowane środowisko programistyczne PyCharm Community w wersji 11.0.8
- Pakiety:
    - NLTK 3.5 w celu wykonania szeregu badań na tekście i wykorzystania korpusu *inaugural* zawierającego przemówienia inauguracyjne prezydentów USA
    - MatPlotLib 3.3.2 w celu stworzenia wykresów dla otrzymanych wyników
    - PyPhen 0.10.0 by obliczyć ilość sylab w przemowach (https://pyphen.org/).

Wymienione pakiety są dostępne w repozytorium pakietów PyPi.

Przeprowadzenie testów

```
import nltk
import matplotlib.pyplot as plt
from nltk.corpus import inaugural
import pyphen

names, num_words, num_sents, num_vocab, word1, word2, lexical_diversity,
num_avg_words_per_sentence, num_avg_syl, \
    word_list = ([] for i in range(10))

symbols = [',', '.', ':', ';', "'", '(', ')', '"', '!']
stop_words = nltk.corpus.stopwords.words('english')
stop_words.extend(symbols)
stop_words = set(stop_words)
sylab_dict = pyphen.Pyphen(lang='en')
```

Wstępnie zostają zaimportowane wymagane biblioteki, stworzone zmienne, słownik PyPhen oraz stop lista z pakietu NLTK uzupełniona o wybrane symbole interpunkcyjne.

```
def get_syllables_count(text):
    n = 0
    for i in text:
        s = sylab_dict.inserted(i)
        n = 1 + n + s.count('-')
    return n
```

Stworzono metodę get_syllables_count podająca ilość sylab znajdującą się w przekazanym tekście. Tekst powinien być listą słów. Każde słowo z listy ma obliczaną ilość sylab za pomocą słownika PyPhen. Zwracana jest ilość sylab dla tekstu.

```
cfd = nltk.ConditionalFreqDist((target, fileid[:4])
                               for fileid in inaugural.fileids()
                               for w in inaugural.words(fileid)
                               for target in ['america', 'law', 'world', 'USA', 'US']
                               if w.lower().startswith(target))
plt.gcf().canvas.set_window_title("Word count 1")
cfd.plot()

cfd2 = nltk.ConditionalFreqDist((target, fileid[:4])
                                for fileid in inaugural.fileids()
                                for w in inaugural.words(fileid)
                                for target in ['unity', 'heal', 'together']
                                if w.lower().startswith(target))
plt.gcf().canvas.set_window_title("Word count 2")
cfd2.plot()
```

Pierwsze wykonane testy mają na celu określenie liczby wybranych słów dla każdej przemowy inauguracyjnej. Dla każdego tekstu, każde słowo po zrównaniu do małych liter jest sprawdzane, czy zaczyna się z wybranymi słowami, a następnie stworzono wykres za pomocą metody *plot()*.

```
for fileid in inaugural.fileids():
    word_list = list(inaugural.words(fileid))
    for i in stop_words:
        if i in word_list:
            while i in word_list:
                word_list.remove(i)

    names.append(fileid[:4])
    # words = len(inaugural.words(fileid)) #wraz z słowami z stop listy
    words = len(word_list)  # bez słów z stop listy
    num_words.append(words)
    sents = len(inaugural.sents(fileid))
    num_sents.append(sents)
    vocab = len(set(w.lower() for w in word_list))
    num_vocab.append(vocab)
    lexical_diversity.append((vocab / words))
    num_avg_words_per_sentence.append(round(words / sents))
    avg_syl = get_syllables_count(word_list) / words
    num_avg_syl.append(avg_syl)
    word1.append(len(nltk.re.findall(r'(united states of america)',
str.lower(inaugural.raw(fileid)))))
    word2.append(len(nltk.re.findall(r'(united states)(?!.*(of america))',
str.lower(inaugural.raw(fileid)))))
    # print(len(nltk.re.findall(r'\b(U\.S\.A)\b', str(inaugural.raw(fileid)))))
```

Dla każdego tekstu znajdującego się w nltk.inaugural zebrano dane. Każde przemówienie zostaje zapisane jako lista słów z której następnie usunięto wszystkie elementy znajdujące się w wcześniej zrobionej stop liście. Oczyszczenie tekstów z stop słów pozwala nam na pobraniu bardziej interesujących danych z których można wyciągnąć celniejsze wnioski. Następnie pobrano i zapisano w listach następujące dane (nie biorąc pod uwagę interpunkcji i stop słów dzięki przygotowanej liście słów bez elementów z stop listy) dla każdego przemówienia:

- Rok przemówienia
- Liczba słów
- Liczba zdań
- Liczba unikatowych słów (ilość słów występujących w tekście bez duplikatów)
- Liczba hapaks legomon w kontekście przemówienia (unikatowe słowa występujące tylko raz w badanym przemówieniu)
- Różnorodność leksykalna (wyrażona w zakresie od 0.00 do 1.00)
- Średnia ilość słów w zdaniu

- Średnia ilość sylab w przemówieniu

Na końcu wykonano podliczenie ilości wystąpień ciągów znaków „united states of america” oraz „united states” (nie licząc wystąpień z końcówką of america.

```
def create_plot(x, xlabel, ylabel, title, y, l, y2='', l2='', y3='', l3=''):
    plt.figure(figsize=(9, 3))
    plt.rcParams['axes.facecolor'] = 'white'
    plt.rcParams['axes.edgecolor'] = 'white'
    plt.rcParams['axes.grid'] = True
    plt.rcParams['grid.alpha'] = 1
    plt.rcParams['grid.color'] = "#cccccc"
    plt.grid(True)
    plt.xticks(rotation='vertical')
    plt.plot(x, y, label=l, linestyle='-', marker='o')
    if y2 != '':
        plt.plot(x, y2, label=l2,  linestyle='-', marker='o')
        if y3 != '':
            plt.plot(x, y3, label=l3,  linestyle='-', marker='o')
    plt.legend(loc='upper right')
    plt.gcf().canvas.set_window_title(title)
    plt.suptitle(title)
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    return plt
```

Celem stworzenia wykresów stworzono metodę create_plot pobierającą dane do wypełnienia wykresów oraz tworzącą wykres na podstawie podanych parametrów. Zwracana jest figura PyPlot.

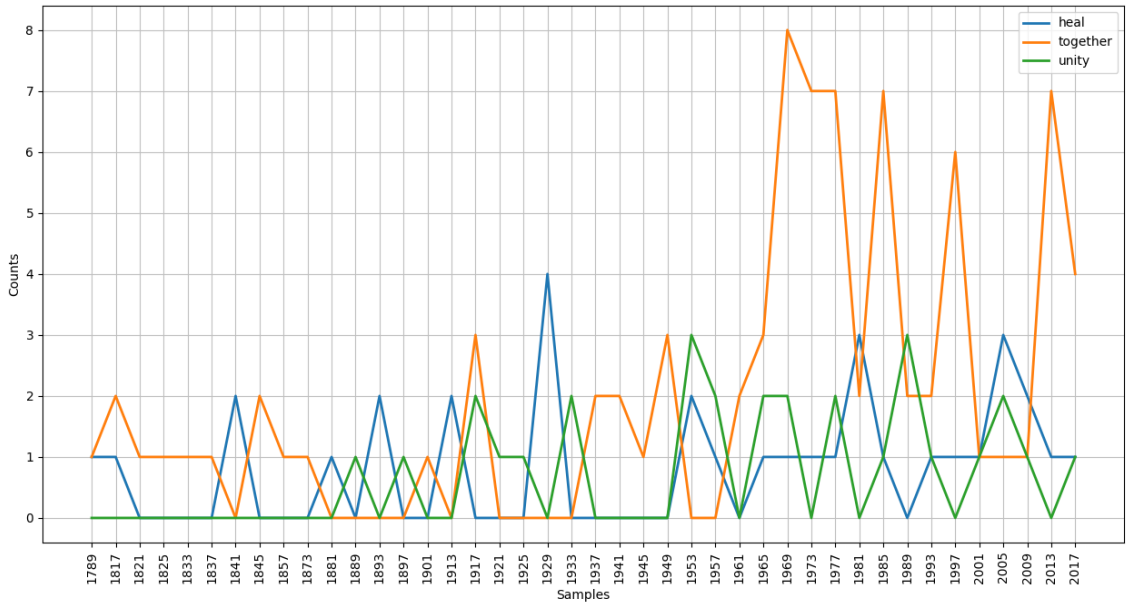
```
create_plot(names, 'Rok', 'Liczba słów', 'Liczba słów w przemowie', num_words, 'Słowa',
num_vocab, 'Unikatowe słowa',
            num_hapaxes, "Hapaksy").show()
create_plot(names, 'Rok', 'Liczba słów', 'Średnia liczba słów w zdaniu',
num_avg_words_per_sentence, 'Słowa',
            num_avg_vocab_per_sentence, 'Unikatowe słowa', avg_hapaxes,
"Hapaksy").show()
create_plot(names, 'Rok', 'Różnorodność', 'Różnorodność leksykalna', lexical_diversity,
            'Różnorodność leksykalna').show()
create_plot(names, 'Rok', 'Liczba zdań', 'Liczba zdań w przemowie', num_sents,
'Zdania').show()
create_plot(names, 'Rok', 'Liczba sylab', 'Średnia liczba sylab na słowo w przemowie',
num_avg_syl,
            'Średnie sylaby').show()
create_plot(names, 'Rok', 'Ilość wystąpień', 'US vs USA', word1, 'United States of
America', word2,
            'United States').show()
```

Wykorzystując stworzoną metodę zostają stworzone i wyświetlone wykresy.

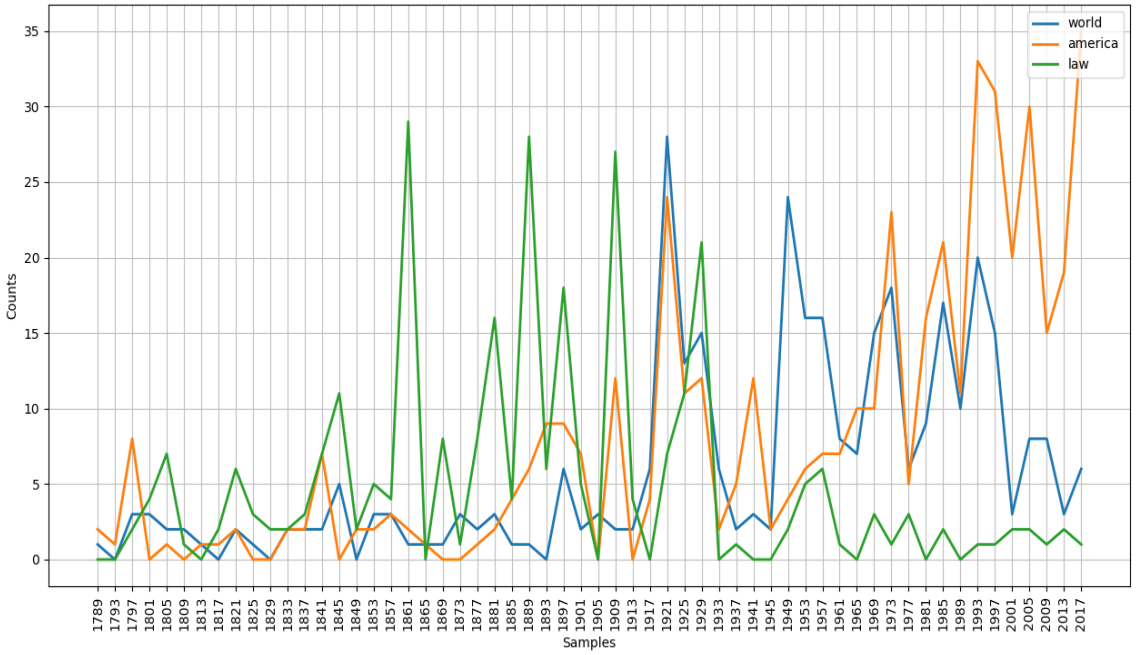
Wyniki badań

Otrzymane rezultaty obliczeń zostały przedstawione w formie wykresów.

BADANIE 1 – SPRAWDZENIE LICZBY WYSTĄPIEŃ OKREŚLONYCH SŁÓW DLA KOLEJNYCH TEKSTÓW



Rys.1. Wykres przedstawiający ilość wystąpień słów ‘heal’, ‘together’, ‘unity’

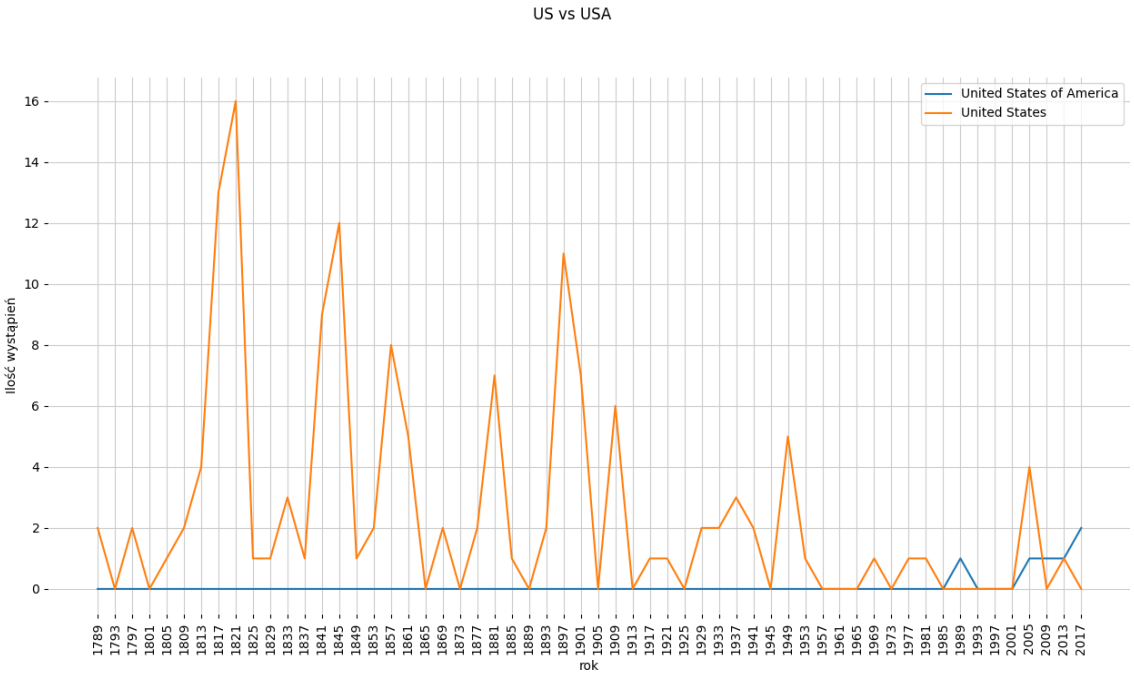


Rys.2. Wykres przedstawiający ilość wystąpień słów ‘world’, ‘america’, ‘law’

Na podstawie informacji zawartych na rysunku pierwszym można zauważyć stałą ilość słów „heal”, „together” oraz „unity” wśród wszystkich przemówień do lat sześćdziesiątych. Wyjątkiem jest słowo „together”, które stało się znacznie częstsze w latach sześćdziesiątych i występuje w każdej przemowie od tamtej pory.

Drugi wykres przedstawia liczbę wystąpień słów „world”, „america” i „law” w przemówieniach inauguracyjnych. Widać znaczącą ilość wystąpień słowa „law” w latach 1857-1929, oraz znaczny wzrost popularności słów „world” i „america” po roku 1917 z wyjątkiem lat 1933-1945. Można też zobaczyć, że od roku 1981 słowo „america” pojawia się co najmniej dziesięć razy w każdym przemówieniu.

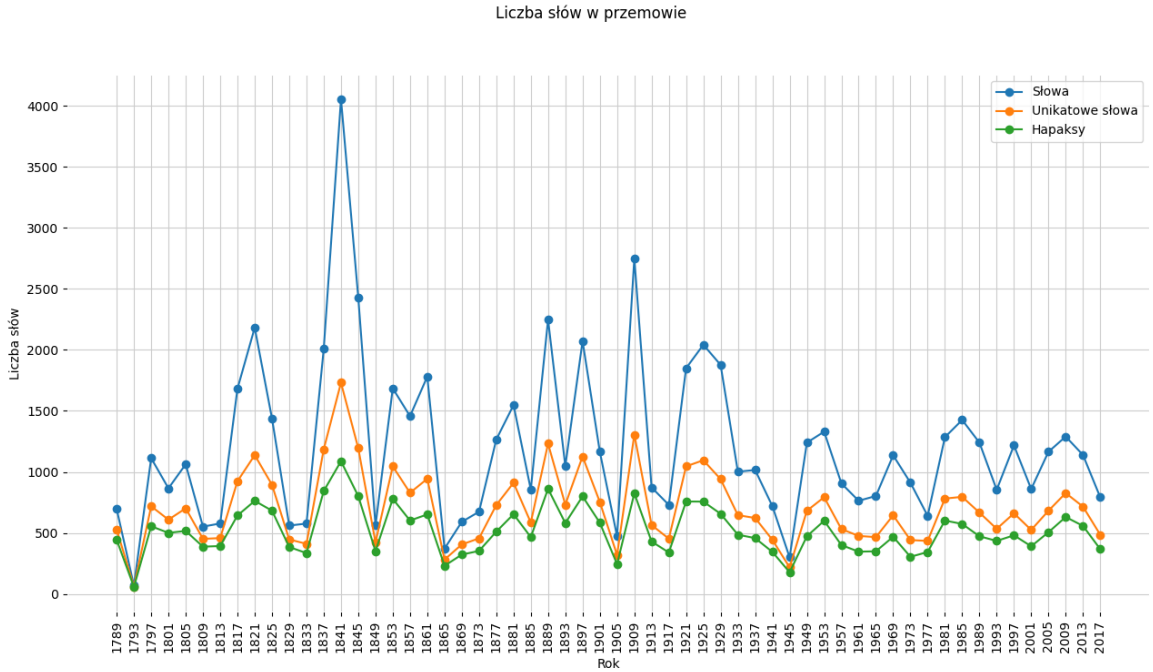
Wyniki powyższych obliczeń mogą służyć do obserwacji zmian języka polityki i tematów podkreślanych przez inaugurowanych prezydentów. Wykresy przedstawiają próby zaobserwowania zmian w charakterze wypowiedzi prezydentów i zbadanie czy przejawia się w nich pojednanie ponad podziałami politycznymi, uszanowania prawa bądź sprawy zagraniczne. Z wykresu drugiego można zauważyć okres znaczącej popularności słowa „law" i pojawienie się niewielkich ilości słowa „together" w niedawnych przemowach. W celu wyciągnięcia zdecydowanych wniosków należy poszerzyć bazę badanych słów o synonimy i wyrazy pokrewne.



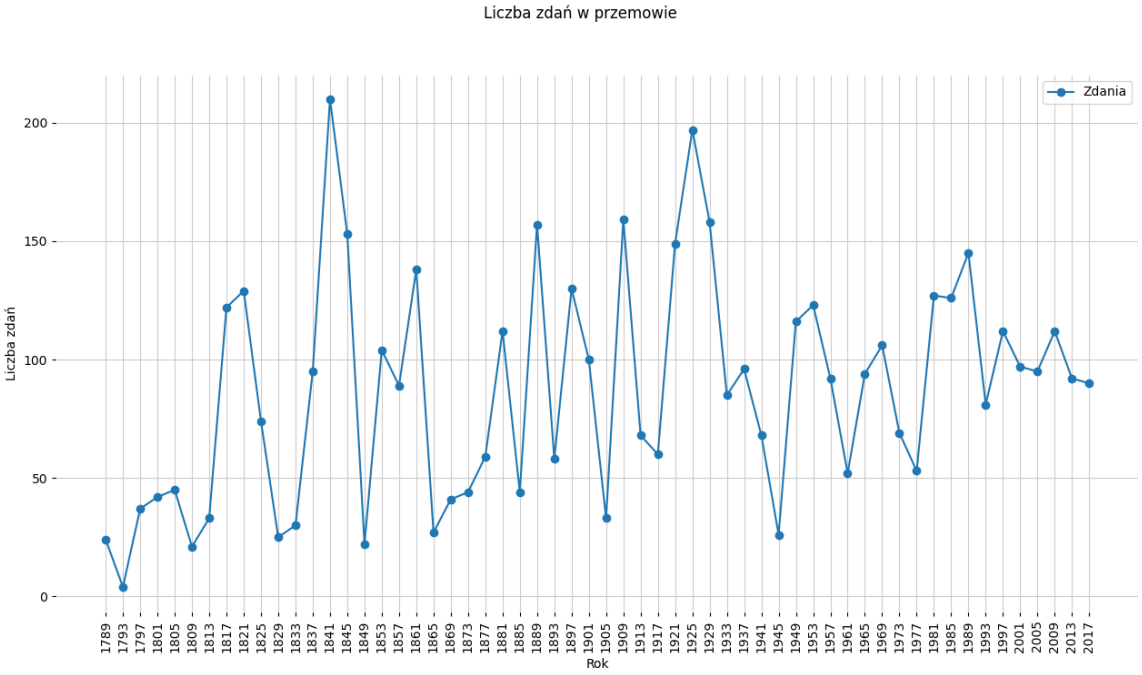
Rys.3. Wykres przedstawiający ilość wystąpień ciągów 'United States of America' oraz 'United States'

Dane zamieszczone na wykresie wskazują na wyraźny spadek popularności frazy „United States" po 1909 roku oraz głębszy spadek po roku 1949. Sformułowanie „United States of America" nie pojawiało się w przemówieniach do roku 1989, gdy zostało po raz pierwszy użyte. Na podstawie tego wykresu i poprzedniego rysunku widać, że nieformalne słowo „america" wyparło dłuższą formę „United States". Można przypuszczać, że jest to przykład uproszczenia języka politycznego. Pełna nazwa Stanów Zjednoczonych Ameryki jest wykorzystywana niezwykle rzadko. Wyniki badań świadczą również o braku akronimów USA, bądź US. Wynika to z uroczystego charakteru przemów inauguracyjnych.

BADANIE 2 – SPRAWDZENIE ILOŚCI WSZYSTKICH SŁÓW, ZDAŃ I ZŁOŻONOŚCI LEKSYKALNEJ



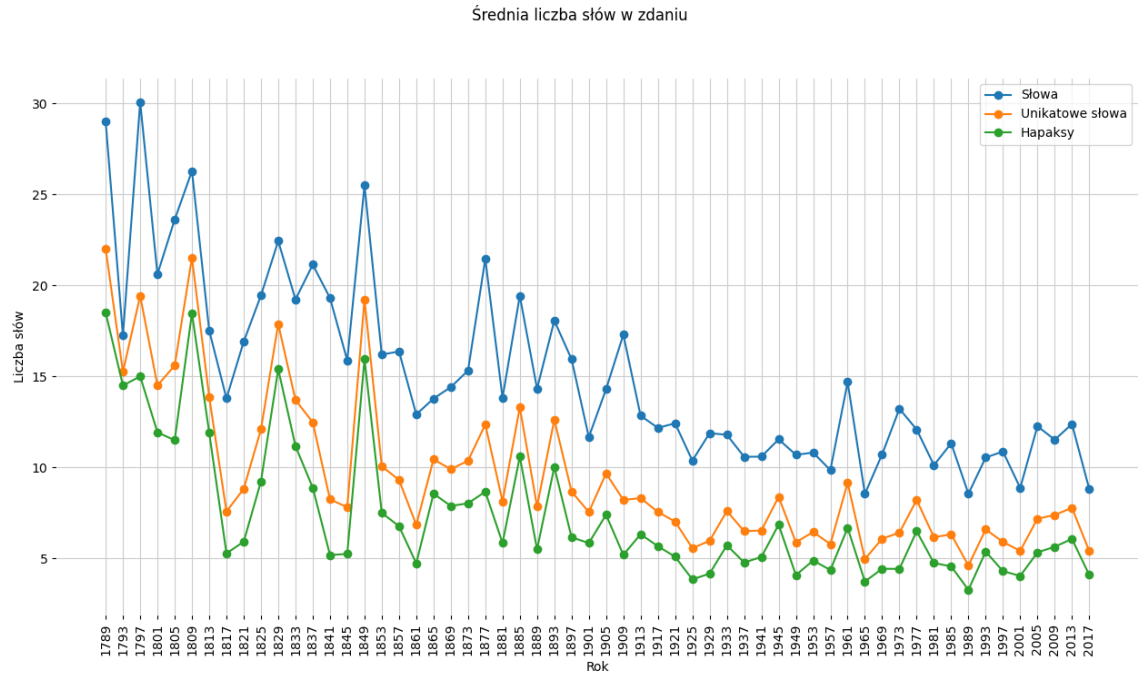
Rys.4. Wykres przedstawiający ilość słów, unikatowych słów oraz hapaksów w przemówieniach



Rys.5. Wykres przedstawiający ilość zdań w przemowach

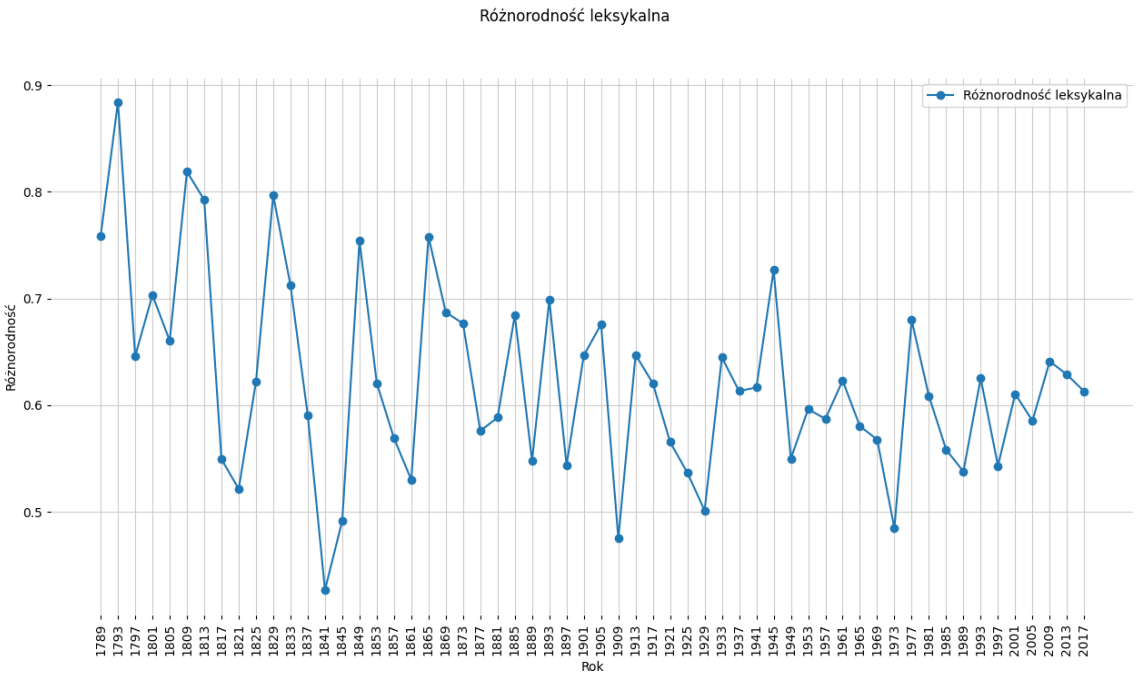
Dane zamieszczone na wykresie czwartym i piątym nie wskazują na istnienie istotnej tendencji w długości przemów, bądź ilości zdań. Należy jedynie zauważyć mniejsze ekstrema dla podanych wartości po roku 1953. Dłuższe przemówienia zawierają znacznie większą ilość słów powtarzających się i mniej hapaksów. Przemówienia od 1953 mają więcej duplikatów.

Najkrótsze przemówienie pochodzi z reelekcji George'a Washingtona w 1793 roku, którego większość miało charakter objaśniający i przypominający o obowiązkach prezydenta, natomiast najdłuższą mowę wygłosił William Henry Harrison w 1841 roku i liczyło ono 8445 słów (wartości na wykresie nie uwzględniają elementów znajdujących się w stop liście).



Rys.6. Wykres przedstawiający przeciętną liczbę słów przypadających na zdanie w tekście

Z wykresu można odczytać wyraźną tendencję spadkową w średniej ilości słów w zdaniu. Świadczy to o większej dynamice przemów.

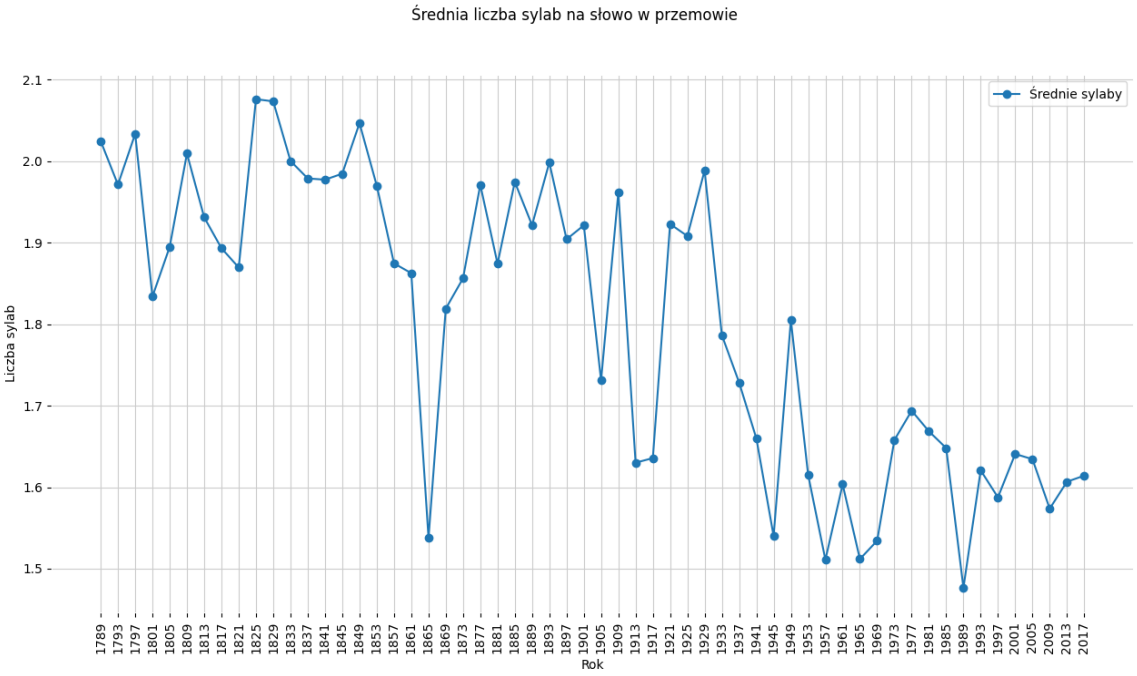


Rys.7. Wykres przedstawiający różnorodność leksykalną dla przemówieniach

Na podstawie otrzymanych wyników można zauważyć niewielki spadek w różnorodności słów w przemowach.

BADANIE 3 – SPRAWDZENIE ŚREDNIEJ LICZBY SYLAB W SŁOWIE W TEKŚCIE.

Celem badania było zbadanie jak zmienia się dynamika i skomplikowanie przemów inauguracyjnych. Niestety NLTK nie zawiera ujednoliconego narzędzie pozwalającego na zbadanie ilości sylab w dowolnym słowie. Celem pozyskania danych zastosowano paczkę PyPhen bazującego na słownikach Hunspell dla dzielenia wyrazów na części. Słowniki Hunspell zawierają niestety błędy i braki w wyniku czego wiarygodność otrzymanych danych jest niepewna.



Rys.8. Wykres przedstawiający średnią ilość sylab w tekstach przemówień

Otrzymane wyniki świadczą o tendencji spadkowej ilości sylab w słowie co oznacza, że wzrasta ilość słów krótkich w języku politycznym i świadczy o jego uproszczeniu. Otrzymane wyniki z wszystkich badań pozwalają na zaobserwowanie uproszczenia i ustandaryzowania przemów inauguracyjnych prezydentów Stanów Zjednoczonych w przeciągu ostatnich dekad.