# Przygotowanie danych

# Usunięcie wierszy z pustym speech\_raw

* 1. Pozostało w bazie 272 221

# Algorytm przypisania autora tekstu (SQL).

Podczas przeglądania danych zidentyfikowano błędy w przypisaniu id autorów przemówień. Konsekwencją tego jest brak możliwości łatwego przypisania autora danego wystąpienia do jego tekstu. Błędy występowały dla większości wpisów w bazie i mimo tego, że sprawiały wrażenie systematycznej zmiany (poprawna wartość wydawała się bliska z kontekście porządku leksykograficznego), nie udało się ustalić algorytmu, który mógłby przypisać poprawne wartości w sposób automatyczny.

W związku z powyższym przypisanie autora (*author\_final*) wykonano w następujących krokach (10\_ author\_final.sql):

1. Wiele wystąpień zaczyna się od schematu „Poseł Imię Nazwisko:”. Pierwsze występujące w takim kontekście imię i nazwisko przypisano jako dane autora z tekstu (kolumna *author\_by\_text*).
2. Przypisano imię i nazwisko autora w oparciu o *id\_* z bazy (kolumna *author\_by\_id*).
3. Dla przypadków gdzie dla danego id, chociaż raz *author\_by\_text* jest równe *author\_by\_id*, przypisano te wartości jako *author\_final*.
4. Dla pozostałych przypadków sprawdzono jaka wartość *author\_by\_text* pojawia się najczęściej w obrębie danego id. Jeśli wartość była niepusta, została przypisana jako *author\_final*.
5. Pozostałe przypadki to takie, gdzie najczęściej pole *author\_by\_text* było puste, tj. w tekście wystąpienia nie występowało zestawienie „Poseł Imię Nazwisko:”. Wynika to z różnej postaci plików transkrypcją – dla części z nich fragmenty pogrubione na poniższym obrazku nie znalazły się w tekście dostępnym w bazie.



1. W trakcie analizy danych niektóre id były weryfikowanych ręcznie w oryginalnych plikach pdf. Wartości dla nich wprowadzono ręcznie w kodzie.

Po zastosowaniu powyższych kroków w bazie brakowało przypisania *author\_final* dla 795 id (na 3828). Dla pojedynczych fragmentów wystąpień uzupełnienie wyniosło ok 83%.

# Czyszczenie fragmentów przemówień (Python)

Wykonano następujące etapy czyszczenia danych:

* usunięto znaki specjalne \n \r \t
* wykasowano tytuły przemówień pojawiające się na początku każdego fragmentu
* usunięty fragmenty tekstu w nawiasach (np. (Oklaski.), (Dzwonek.))
* Usunięto fragmenty identyfikujące mówcę, tj. fragment „Poseł Imię Nazwisko:”

Następnie przygotowano drugi zestaw tekstów, które oczyszczono jeszcze bardziej z fragmentów mało informacyjnych.

* Wypowiedzi innych posłów i komentarze typu "Oklaski" pojawiające się w nawiasach
* "Poseł Imię Nazwisko:"
* Panie Marszałku!
* Pani Marszałek!
* Wysoka Izbo!
* Panie Ministrze!
* Dziękuję bardzo.

Po tym etapie czyszczenia w bazie pozostało 272 217 fragmentów przemówień, z których 225 385 ma przypisanego autora. Oba zestawy tesktów z różnym poziomem oczyszczenia będą stosowane w różnych analizach.

# Łączenie fragmentów przemówień (python)

Fragmenty tych samych wypowiedzi, zgrupowane po dacie, autorze i tytule został y połączone.

Po usunięciu tekstów, które po wszystkich modyfikacjach stały się puste, w bazie zostało 158 885 tekstów z przypisanym autorem.

# Wstępna analiza danych

Dla tekstów z przypisanym autorem przeprowadzono wstępną analizę danych. Po usunięciu cyfr, znaków interpunkcyjnych oraz słów z listy polskich *‘stop-words’* analizowano częstości w celu weryfikacji i rozszerzenia listy *‘stop-*words’. W wyniku analizy dodanie do niej następujące elementy, ze względu na niską wartość informacyjną i ryzyko wprowadzenia zaburzeń do dalszych analiz:

zł

pkt

art

ustawy

r

Po wykonaniu tego czyszczenia, w korpusie znajduje się nieco ponad 42 mln słów. Przed lematyzacją w korpusie występuje 367 716 różnych słów.

Słów, które można określić jako „rzadkie”, tj. występujące w całym korpusie nie więcej niż 5 razy jest 226 599 (w tym 126382 występuje tylko jeden raz), czyli stanowią one istotną większość. Przykładowe słowa pojawiające się dokładnie jeden raz w całym korpusie:

zaanektować

babette

opowiadaną

odejmowana

trawiona

kolektywistyczna

zagonów

troić

maćkowy

centralizowany

Słowa pojawiające się najczęściej, to:

[('rząd', 95603),

('projektu', 95657),

('prawa', 96092),

('chodzi', 99480),

('panie', 108932),

('projekt', 113847),

('pytanie', 116217),

('pracy', 118301),

('komisji', 146153),

('państwa', 148339)]

# Generatory przemówień bazujące na modelach n-gramowych

Na tym etapie podjęto decyzję o budowie prostego generatora przemówień opartego o modele n-gramowe. Takie modele bazują na statystykach występowania n-gramów w analizowanym korpusie i mogą służyć do przewidywania kolejnego elementu sekwencji jak również do generowania nowych sekwencji w oparciu o zaobserwowane zależności.

Przykładowo prawdopodobieństwo, że w naszej sekwencji, po słowie „praca” wystąpi słowo „zaliczeniowa” wynosi:

P(zaliczeniowa|praca) = cnt(„praca zaliczeniowa”)/cnt(„praca \_\_\_”)

Gdzie \_\_\_ oznaca dowolne słowo.

# Ogólny generator

Model powstał w dwóch wersjach dla 1-gramów oraz 2-gramów. Dla obu opcji w pierwszym kroku przygotowano słownik zawierający wszystkie występujące kombinacje odpowiednich n-gramów ze słowami po nich następującymi.

Generowanie przemówień polega na przypisaniu słowa początkowego (lub wybraniu losowego) a następnie w sposób losowy wybieraniu słów kolejnych na bazie słownika zbudowanego w poprzednim punkcie.

Wraz ze wzrostem n, rośnie zapotrzebowanie na dane do modelu. Jednak takie modele potrafią generować sekwencje lepszej jakości, ponieważ zachowują więcej zależności ze zbioru uczącego.

Kod przygotowujący te najprostsze modele znajduje się w pliku „02\_generator\_1.0.ipynb”. Do ich wykonania nie były potrzebne żadne z pakietów związanych z NLP. Niemniej nie udało się w ten sposób zbudować modelu 3-gramowego. W kolejnych etapach ta możliwość będzie ponownie analizowana. Poniżej przykładowe wyniki.

Przykładowe przemówienie o długości 60 słów dla modelu 1-gramowego, rozpoczynające się od słowa „My”:  
„My wychodzimy z rocznicami sienkiewiczowskimi i doświadczenie i 7,5%. Tuba propagandowa kampania antynikotynowa ma współdziałanie pracodawców i rozważa, rząd raczy projekt. W związku z senackich do wyborów bezpośrednich uprawnień. Na dzisiaj otrzymaliśmy, zawiera następujące zmiany: Art. 33 tys. altan ogrodowych pod kątem legalności źródeł finansowania. Pani Marszałek! Wysoki Sejmie! Dziesięć państw członkowskich Rady Radiofonii i stanowi poważne przeobrażenia, jakie substancje związków.”

Przykładowe przemówienie o długości 60 słów dla modelu 1-gramowego, rozpoczynające się od losowego słowa:

„Urobek podkomisji, mają stosować się zadania i tylko chęci ani upraw na czym operatorzy i przytułków dla polskiej żeglugi śródlądowej do Prawa Podatkowego, która zwalnia urzędników. Muszę zresztą widać w którym funkcjonuje już wyżej wymienionym projektem ustawy o których niesławnymi bohaterami narodowymi funduszami spójności terminologicznej. Wydaje się, że nasilają się zgodnie z mechanizmów, które toczą się teraz tego demo potrafi przeprowadzić.”

Przykładowe przemówienie o długości 100 słów dla modelu 2-gramowego rozpoczynające się od wyrażenia „Dziś jest”:

„Dziś jest już ustalony system finansów publicznych? Czy może mi pan minister mówi, że ta krytyka zasługuje na aprobatę argument zakładający, iż prezes Izby Cywilnej oraz Izby Kontroli w dniu 2 marca 2006 r. o świadczeniach przedemerytalnych też jest możliwe. Zdaniem prezydenta narusza to uprawnień przedsiębiorcy, naprawienie szkody wyrządzonej czynem niedozwolonym oraz pracowników cywilnych. Dla obydwu grup wyważyć. Pamiętajmy także o zwykłą większość głosów. Co to oznacza? Jest to bardzo ważna kwestia, to mianowicie, że w tej sytuacji pogłębia fakt, że pomoc państwa dla firm oraz poparcie dla tego rodzaju działalności tak specyficznych instytucji, jakimi są Policja, ośrodki pomocy społecznej regulują oczywiście.”

Przykładowe przemówienie o długości 60 słów dla modelu 2-gramowego rozpoczynające się od losowego wyrażenia:

„Ilość poprawek, które rzeczywiście będą oni niezależni, niezawiśli, nieusuwalni, odpowiadać będą również wiodącymi, jeśli chodzi o cywilizowanie prawa pracy, proszę państwa, czy też w krajach Unii Europejskiej. Liga Polskich Rodzin ocenia pozytywnie działania Ministerstwa Łączności do poselskiego projektu ustawy. Dziękuję bardzo. Pani Marszałek! W imieniu Klubu Parlamentarnego Prawo i Sprawiedliwość nagle zmieniło zdanie, nagle postanowiło pokazać, że jest to naprawdę bardzo wiele polskich firm na rynkach światowych powoduje, że dla wprowadzenia takiego rozwiązania, które służą wszystkim obywatelom. Omawiana dziś regulacja jest szczególnie wrażliwa na punkcie tej opinii BAS-u są kuriozalne albo w spółkach górniczych, trzeba przyznać lojalnie - dwie doprecyzowujące, dwie troszeczkę.”

Widać, że w drugim modelu tekst bardziej przypomina realnie wystąpienia i prawdziwe zdania.

# Badanie prawdopodobieństwa danego zdania

Mamy również możliwość przeanalizowania spójności danej sekwencji z modelem. Tj. jak prawdopodobne jest, że dane zdanie pochodzi z danego modelu. Można to szacować za pomocą wartości perplexity, która pozwala ocenić jakość modelu n-gramowego na danych testowych.

P=*p*(*wtest*)−1*N*, where *p*(*wtest*)=∏*i*=1*N*+1*p*(*wi*|*wi*−1*i*−*n*+1)

Aby zaadresować kwestię zerowania się iloczynu w przypadku pojawienia się n-gramów spoza korpusu, na którym trenowano model, wprowadza się różne metody wygładzania tej miary. Tutaj zastosujemy „add-1 smoothing”, która modyfikuje kalkulację pojedynczych prawdopodobieństw w następujący sposób:

*p*(*wi*​∣*wi*−*n*+1*i*−1​)= *c*(*wi*−*n*+1*i*​)+1​ / *c*(*wi*−*n*+1*i*−1​)+*V*   
Kalkulacja perplexity może uwzględniać również pozycję początkową i końcową wyrazu w zdaniu. Należy wówczas uwzględnić w konstrukcji modelu dodatkowe tokeny oznaczające początek i koniec zdania.

# Badanie prawdopodobieństwa danego zdania w zależności od autora wystąpienia