# 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)]

Najczęściej wypowiadający się posłowie to:

| **author\_final** | **Liczba slów** | **Liczba wypowiedzi** |
| --- | --- | --- |
| **Stanisław Stec** | 477862 | 1368 |
| **Mirosław Pawlak** | 203379 | 1134 |
| **Romuald Ajchler** | 401879 | 1120 |
| **Jan Kulas** | 547001 | 1101 |
| **Andrzej Szlachta** | 242306 | 982 |

Posłowie z największą liczbą słów w wypowiedziach to:

| **author\_final** | **Liczba slów** | **Liczba wypowiedzi** |
| --- | --- | --- |
| **Jan Kulas** | 547001 | 1101 |
| **Józef Zych** | 514652 | 878 |
| **Jerzy Jaskiernia** | 481512 | 658 |
| **Stanisław Stec** | 477862 | 1368 |
| **Tadeusz Tomaszewski** | 412781 | 917 |

# 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 a liczniki „cnt” zostały wygenerowane na korpusie, na którym trenowano model.

# Konstrukcja modeli

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 wybieraniu słów kolejnych na bazie słownika prawdopodobieństw zbudowanego w poprzednim punkcie.

Wraz ze wzrostem n, rośnie zapotrzebowanie na dane do modelu. W przypadku gdy danych do uczenia jest zbyt mało, model będzie po prostu odtwarzał zdania, które pojawiały się w danych uczących. Z drugiej strony takie modele potrafią generować sekwencje lepszej jakości, bardziej przypominające prawdziwe zdania, ponieważ zachowują więcej logicznych powiązań pomiędzy kolejnymi słowami.

Do zbudowania najprostszych modeli nie były wykorzystane żadne z pakietów związanych z NLP.

Aby uprościć i zautomatyzować konstrukcję modeli, przygotowano klasę ng\_models (ng\_models.py). Obiekt tej klasy jest modelem n-gramowym utworzonym zgodnie z przekazanymi parametrami *n* oraz *corpus*. Tworzenie modelu odbywa się z uwzględnieniem znaków specjalnych oznaczających początek i koniec zdania. Podczas jego konstrukcji następuje przetworzenie korpusu metodą *process\_text*, która m.in. usuwa znaki specjalne, duże litery, wielokrotne spacje.

Za generowanie wypowiedzi odpowiada metoda **generate** pozwalająca generować sekwencje zdań zadanej długości. Metoda zaczyna od losowego słowa z listy słów pojawiających się na początku zdań. Każde zdanie generowane jest oddzielnie do momentu aż zostanie wylosowany znacznik końca zdania, lub zostanie przekroczony parametr ograniczający długość zdania.

Konstrukcja modelu 1 gramowego na lokalnym komputerze trwa ok. 4 minuty a 2 gramowego 6 minut .

Poniżej przykładowe wyniki.

Przykładowe przemówienie dla modelu 1-gramowego wygenerowane z parametrami (steps=10, max\_sent=10):

'Panie marszałku. Poza tym pracownikom projekt ordynacji podatkowej dla polski 1. Nie przekreślamy sprawę jasno sformułowane są niezgodne z dwiema. Jak zostanie zwrócony jako szef zespołu na wykreśleniu pkt. Boję się o swoich list to zasłużone dla strażaków. Chciałbym skupić wyłącznie do spożycia alkoholu oraz refundacji. Wysoka izbo. Jakie gwarancje zastawnika niż pobierających świadczenia z góry rozdzielanie. Swoich dzieci. W praktyce zostało wyartykułowane.'

Przykładowe przemówienie dla modelu 1-gramowego wygenerowane z parametrami (steps=10, max\_sent=20):

'Natomiast takich uchwał dotyczących kwestii warunków prawnych w budowie. Szanowny panie marszałku. Pierwsza. W rozporządzeniu rady sądownictwa dopiero na swoich kompetencji i. To kolejny projekt i szczególnej państwowej służby ochrony państwa. W takiej potrzeby zwiększania środków europejskich demokratów formułuje wnioski. Szanowni państwo dzisiaj o szczególne znaczenie specjalnych jest wierną. Powiedzieliśmy do spółki na stworzenie nowej rezerwy zostanie skazany. Panie marszałku. Manipulowanie tą wersją projektu skłania do kilkuletnich zaniedbań i. Pierwszy pracę. Dopiero w związku z przeznaczeniem a przede wszystkim informacyjne. To proszę państwa dla poszczególnych powiatów oraz zniesienie ograniczeń. Przypuszczam że każda ze strony młodzieży oraz upoważnienie na. Później próbuje ogarnąć tej sprawie wyjścia naprzeciw wspomnę np. Można było. Potrzeba uchwalenia zaproponowanego przez radę ministrów. Jej do pana ministra tchórzewskiego byłego likwidatora majątku zagrożonego. Wprowadza tę izbę nie udało mi wiadomo że albo. Wysoki sejmie.'

Łatwo jest ocenić, że powyższe wypowiedzi nie pochodzą z prawdziwych wystąpień sejmowych. Możliwych kombinacji słów następujących po sobie jest zbyt dużo aby taki model mógł zachować logikę pomiędzy kolejnymi elementami zdania.

Przykładowe przemówienie dla modelu 2-gramowego wygenerowane z parametrami (steps=10, max\_sent=10):

'Mam kłopot z budowaniem nowych mieszkań w złym świetle respektowanie. To podchodzi pod administrację. Chodzi zapewne o wiele trudniejsza jest sytuacja w państwie członkowskim. Jeżeli obie strony uznały za obowiązujące zajmowanie przez 10 lat. Panowie osiągnęli a my słyszymy sto takich zgromadzeń. Jeżeli uzyskaliśmy technologie na warunkach partnerskich; wtedy jest ona nam. Do orzekania... Tyle jesteś samorządny o tyle obecnie średnio 133 osoby na. Przyjęcie kompromisowych rozwiązań łączących projekt rządowy nad którym dyskutujemy to. Jak kupuję jakiś towar.'

Przykładowe przemówienie dla modelu 2-gramowego wygenerowane z parametrami (steps=10, max\_sent=20):

'Chciałbym wykreślić ze sprawozdania... Powoli zanika a mogła to być może wykorzystawszy także niezły poziom techniczno-technologiczny jest znacznie lepszy niż ten który złożył przysięgę. Jednak prenumerować przez pocztę przez swoich i dlatego nie ma powodu do jakichkolwiek sytuacji które mogą naruszać wolności wykonywania zawodu. Czy zrewaloryzowana ale też o rozruchu. Nie ustosunkował.właśnie do tego że komitet badań naukowych lub organizacji społecznej o czym mówił pan prezes socha był łaskaw pan. To przekształcane w pracowniczy program emerytalny jest instytucją samofinansującą się która jest w pełni popiera i w imieniu posłów z. Będziemy kontrolować waszą działalność trybunał stanu który decyzją sejmu termin ten wydaje się tu toczy będzie miała zastosowanie w postępowaniu. Nie ochronę uprawnień najemców i dzierżawców które już dzisiaj plagi społecznej. Mówię wewnętrzna organizacja tego ratownictwa także. Gdyby obecne napięcia między przedstawicielami ministerstwa skarbu czy po przyjęciu będzie miała jedno pytanie: co pani poruszyła są prawnie chronione.'

Widać, że w drugim modelu tekst bardziej przypomina realnie wystąpienia i prawdziwe zdania. Oczywiście w dalszym ciągu nie ma żadnych wątpliwości, że nie są to realne wypowiedzi, ale ich czytanie jest nieco przyjemniejsze niż dla poprzedniego modelu.

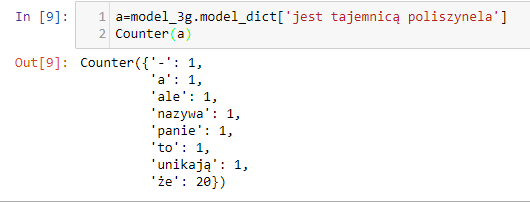
Podjęto również próbę zbudowania modelu 3-gramowego. Czas kalkulacji wydłużył się istotnie. Budowa słownika do modelu trwała prawie 30 minut.

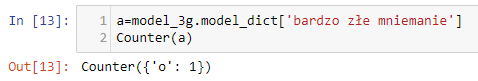
Przykładowy tekst wygenerowany z parametrami (steps=10, max\_sent=10):

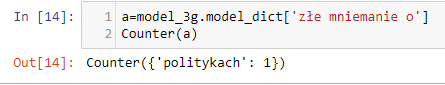
'Czy elementem realizacji polityki gospodarczej. Już zadała to pytanie bo jest tajemnicą poliszynela unikają one płacenia miliardowych podatków i nadmiernie wykorzystują pracowników. I marek belka nie budzi naszego zaufania. Dziękuję unii wolności za to że ma wygrać najlepszy. Chciałabym przytaczać w tej chwili problem bezrobocia mimo że ten dokument został odtajniony decyzją wiceministra spraw wewnętrznych i administracji i innym. Wiadomo są jedne z głównych czynników utrudniających udany proces integracji repatriantów z polskim społeczeństwem z obywatelami żeby przez zaniechania rządu nie. Przecież trudno się spodziewać tego że przy wyważeniu argumentów ograniczenie czasu zgłaszania weta też ma pewne koncepcje które należałoby wysokiej izbie. Połowa to mogą być grupy członkowskie. Wtedy ta nowelizacja rzeczywiście mogła wejść w życie na drugim etapie jego nowelizacji. Panie prokuratorze zwrócić uwagę że istnieje bardzo złe mniemanie o politykach i o polityce wobec cudzoziemców polityce imigracyjnej jest takim dzwonem.'

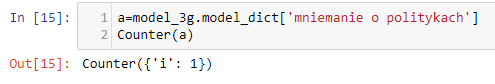
Zdania sprawiają wrażenie lepiej skonstruowanych niż dla poprzednich modeli. Jednak stosunkowo mały zbiór danych uczących sprawia, że model w dużym stopniu powtarza całe zdania ze zbioru uczącego.

Przyjrzyjmy się jak wygląda słownik dla niektórych 3-gramów występujący w powyższym tekście.









Z powyższego wynika, że do budowy dobrego modelu dla większych wartości n dla języka polskiego potrzebne są dużo większe zestawy danych. Dla języków z prostszymi gramatykami, gdzie nie występuje tak wiele odmian słów, wymagania te mogą być niższe.

# Badanie prawdopodobieństwa wystąpienia 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 generalnie służy do oceny jakości modelu n-gramowego na danych testowych. Dla pojedynczego zdania możemy wyliczać:

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

Gdzie N jest długością zdania testowanego. Liczniki i indeksowanie są tak dopasowane aby uwzględniać pozycję początkową i końcową wyrazu w zdaniu poprzez uwzględnianiu w analizach dodatkowych tokenów oznaczających początek i koniec zdania.

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*

Za kalkulację tej wielkości odpowiada metoda **perplexity** zaimplementowana dla klasy *ng\_models*.

Przykładowo dla modelu 2-gramowego z poprzedniego punktu otrzymujemy dla zdania ‘'Dzisiejsze posiedzenie ma na celu omówienie projektu ustawy'’ wartość 109921 a dla zdania 'Lwy to zwierzęta kotowate żyjące na sawannie' 4240107.

# Badanie prawdopodobieństwa danego zdania w zależności od autora wystąpienia

Metodę opisaną w powyższym punkcie wykorzystano do weryfikacji jakości modeli jakie można zbudować dla konkretnego autora. Ponieważ dzieląc zbiór uczący na poszczególnych autorów, zasób danych staje się mocno ograniczony, analizy te wykonywano dla modeli 1-gramowych.

Zbudowano oddzielne modele dla 5 polityków z największą liczbą słów w danych (400 – 500 tys.) oraz dla trzech z istotnie mniejszą ilością słów (<100 tys.). Porównano wyniki dla kilku mniej lub bardziej prawdopodobnych zdań.



Na szaro zaznaczone są osoby, ze stosunkowo dużym zbiorem danych uczących. Widać, że ma to istotny wpływ w sytuacjach gdy oceniane są zdania „mało prawdopodobne”. Wówczas w kalkulacji prawdopodobieństwa ciągu wyrazów pojawiają się bardzo małe wartości, co z kolei przekłada się na wysokie perplexity. Ogółem prawdopodobieństwa mają podobne trendy dla wszystkich autorów, ale warto zauważyć zaburzenie kolejności pomiędzy zdaniami „Demokracja jest najlepszym systemem politycznym” a „Zupełnie przypadkowy układ pięciu słów”. Niemniej cieszy, że sekwencja „lew czarownica i stara szafa” w każdym przypadku są opcją najmniej prawdopodobną.

Niemniej przeglądając słowniki dla poszczególnych modeli budowanych w ten sposób, widać, że cechuje je duża przypadkowość. Zbyt wiele n-gramów nie będzie się w nich odnajdywać i będą kontrybuowały do miary z tą samą, najniższą wartością. Zdecydowanie też miara ta nie nadaje się do porównywania modeli budowanych na korpusach o istotnie różnej liczności.

Do konsultacji – czy ma sens modyfikacja miary aby wyniki można było porównywać? Nie.

# Budowa wypowiedzi w oparciu o maksymalizację prawdopodobieństwa

Poniższe analizy wykonano na tekście oczyszczonym z najczęściej pojawiających się, mało informacyjnych fraz typu „Panie Marszałku!”.

Znając prawdopodobieństwa następowania kolejnych słów po sobie, można spróbować zbudować zdanie minimalizujące wartość perplexity, tj. maksymalizujące prawdopodobieństwo „zobaczenia go w modelu”. W tym celu rozszerzono klasę *ng\_models* o metodę *mprob\_sent*. Celem tego ćwiczenia była możliwość otrzymania zdania, które byłoby najbardziej charakterystyczne dla danego korpusu, czyli na przykład można by je określić jako dewizę związaną z danym tematem, osobą czy też grupą.

Metoda zakłada pobranie pierwszego słowa zdania od użytkownika a następnie wybieraniu kolejnych słów wg prawdopodobieństw w modelu. Czyli jeśli słowem początkowym jest słowo „poseł”, metoda sprawdzi jakie n-gramy pojawiały się w modelu zaczynające się od słowa poseł i wybierze najczęstszy. Przykładowe najczęstsze następniki słowa „poseł” w modelu 1-gramowym dla wybranego mówcy.

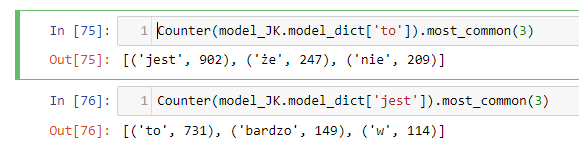


Niestety okazuje się, że taki algorytm bardzo łatwo wpada w pętle. Przykładowe zdania otrzymane w ten sposób:

'Poseł sprawozdawca mówił o tym że w tym że w tym że w tym że w tym że w tym że w tym że w tym że w tym że w tym.'

'Ja to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to jest to.’

Powód jest zobrazowany na poniższym obrazku.



W takim przypadku, jeśli model trafi w pewnym momencie na słowo „to”, zapętli się i będzie powtarzał 2-gram „to jest”.

Jednym z możliwych rozwiązań jest zakaz powtarzania słów w obrębie jednego zdania. W przypadku trafienia na słowo, które już się pojawiło w trakcie generowania, wybieramy kolejne wg częstości pojawiania się. Kończymy generowanie, gdy sytuacja powtarza się zbyt często. Wówczas otrzymujemy np.:

'Poseł sprawozdawca mówił o tym że w tej ustawy.'

'Ja to jest bardzo ważne.'

W przypadku, gdy do metody nie zostanie przekazane pierwsze słowo, generowanie rozpocznie się od słowa, które najczęściej pojawia się jako pierwsze w zdaniach w danym korpusie. Dla modeli budowanych dla pojedynczych mówców są ta zdania m.in.:



Długość otrzymanych zdań jest bezpośrednią pochodną skłonności do zapętlania się słów w danym korpusie.

Jakoś usunięcie stop wordów?