1. Manual.py

Ten kod wyodrębnia tekst z pliku PDF, a następnie przetwarza go za pomocą biblioteki spaCy w celu tokenizacji oraz usunięcia stop-słów i znaków interpunkcyjnych. Następnie za pomocą TF-IDF (Term Frequency-Inverse Document Frequency) analizuje częstość występowania słów w przetworzonym tekście, aby znaleźć te, które są najbardziej reprezentatywne dla dokumentu

Słowa kluczowe (TF-IDF): ['ust', 'numer', 'dnia', 'mowa', 'celnych', 'poz', 'towaru', 'wartość', 'zgłoszenie', 'art', 'celne', 'ewidencję', 'finansów', 'podatku', 'podstawie', '2019', 'dz', 'przesyłek', 'przesyłki', 'składane', 'sprawie', 'towarów', 'walucie', '14', '18', '19', '23', 'adres', 'dziennik', 'identyfikacyjny', 'kod', 'minister', 'ministra', 'należności', 'osoba', 'osoby', 'pkt', 'polskiej', 'rozporządzenia', 'rozporządzenie', 'samej', 'składająca', 'towary', 'ustaw', 'zgłoszeniem', 'zgłoszeń', 'znany', '10', '11', '1169', '150', '2004', '2016', '2018', '2020', '2070', '2262', '2265', '24', 'administracji', 'brzmieniu', 'celnego', 'celnemu', 'celny', 'celnym', 'cn', 'części', 'dane', 'danych', 'dni', 'dodaje', 'dokonane', 'dopuszczenia', 'dotychczasową', 'dotyczącej', 'dza', 'działa', 'działalności', 'działania', 'działem', 'dzonej', 'elektronicznej', 'euro', 'ewidencja', 'ewidencji', 'finanse', 'fizyczne', 'gospodarczej', 'grudnia', 'ilość', 'imię', 'inansó', 'inistra', 'jednej', 'kiego', 'kieruje', 'kodeksu', 'kwoty', 'listopada', 'mające', 'marca', 'ministrów', 'nadawcy', 'nastę', 'następuje', 'nazwisko', 'nazwę', 'nieobjęte', 'nieprowadzące', 'obejmujących', 'objęcie', 'objętych', 'obniżające', 'obrotu', 'odbiorcami', 'odbiorcy', 'odrębnych', 'ogłoszenia', 'określenie', 'operator', 'operatorów', 'opis', 'organ', 'organowi', 'oznacza', 'pierwszy', 'pocztowych', 'podatkowych', 'podlegają', 'podwyższające', 'postaci', 'pośredni', 'prawidłowe', 'prawo', 'prezesa', 'pro', 'procedurą', 'prowa', 'prowadzi', 'prowadzącemu', 'prowadzący', 'przedstawiciel', 'przekracza', 'przepisów', 'przesyłce', 'przewożonych', 'przywozowych', 'publiczne', 'pujące', 'rady', 'ro', 'rozporządzeniu', 'rzeczypospolitej', 'rządowej', 'rządzenie', 'równowartości', 'skiba', 'spełnione', 'sprzedanych', 'sprzedaży', 'stawce', 'stawkę', 'stycznia', 'system', 'szczegółowego', 'transakcji', 'treść', 'trybie', 'udostępnia', 'udostępniana', 'unijnego', 'upływie', 'ustanawiającego', 'ustawy', 'usług', 'uzna', 'wadzącemu', 'waluty', 'warszawa', 'wchodzi', 'wniosek', 'wpływ', 'wraz', 'września', 'wspólnotowy', 'wymogi', 'wyrażona', 'wysokości', 'wysokość', 'wyznaczony', 'wz', 'zakresu', 'zamówienia', 'zarzą', 'zatwierdzana', 'zawartych', 'zawierających', 'zbiorczo', 'zgodnie', 'zgłaszanych', 'zgłoszenia', 'zgłoszeniu', 'zmieniające', 'zp', 'zwolnieniem', 'zwolnień', 'zwolnione', 'złożeniem', 'łączna', 'łącznie', 'życie']

1. lda.py

Ten kod wyodrębnia tekst z pliku PDF i przetwarza go przy użyciu biblioteki spaCy, wykonując tokenizację, usuwanie stop-słów, lematyzację oraz usuwanie cyfr i znaków specjalnych. Następnie tworzy słownik i korpus dla modelu LDA (Latent Dirichlet Allocation) przy użyciu biblioteki Gensim, który jest trenowany na jednym temacie (num\_topics=1) przez 15 iteracji. Na końcu kod wyświetla 10 najważniejszych słów dla wyodrębnionego tematu, które reprezentują główne idee zawarte w dokumencie.

(0, '0.031\*"celny" + 0.024\*"towar" + 0.024\*"zgłoszenie" + 0.022\*"ustęp" + 0.018\*"numer" + 0.018\*"dzień" + 0.016\*"mowa" + 0.016\*"ewidencja" + 0.016\*"przesyłka" + 0.013\*"finanse" + 0.013\*"poz" + 0.013\*"rozporządzenie" + 0.013\*"wartość" + 0.011\*"podatek" + 0.011\*"artykuł" + 0.011\*"składać" + 0.011\*"podstawa" + 0.011\*"waluta" + 0.011\*"osoba" + 0.009\*"minister" + 0.009\*"sprawa" + 0.007\*"objąć" + 0.007\*"działać" + 0.007\*"adres" + 0.007\*"wysokość" + 0.007\*"kod" + 0.007\*"organ" + 0.007\*"należność" + 0.007\*"identyfikacyjny" + 0.007\*"operator" + 0.007\*"Minister" + 0.007\*"sam" + 0.007\*"punkt" + 0.007\*"zwolnienie" + 0.007\*"odbiorca" + 0.007\*"udostępniać" + 0.004\*"stawka" + 0.004\*"system" + 0.004\*"styczeń" + 0.004\*"szczegółowy" + 0.004\*"stawiec" + 0.004\*"DZIENNIK" + 0.004\*"sprzedać" + 0.004\*"spełnić" + 0.004\*"składane" + 0.004\*"równowartość" + 0.004\*"rządowy" + 0.004\*"rada" + 0.004\*"publiczny" + 0.004\*"przywozowy" + 0.004\*"przewozić" + 0.004\*"przesyłeka" + 0.004\*"przepis" + 0.004\*"przekraczać" + 0.004\*"sprzedaż" + 0.004\*"transakcja" + 0.004\*"ustanawiać" + 0.004\*"tryb" + 0.004\*"zarzą" + 0.004\*"zatwierdzać" + 0.004\*"zawierać" + 0.004\*"zawrzeć" + 0.004\*"zbiorczo" + 0.004\*"zgodnie" + 0.004\*"zamówienie" + 0.004\*"zgłaszany" + 0.004\*"zmieniać" + 0.004\*"znany" + 0.004\*"znać" + 0.004\*"zwolnić" + 0.004\*"złożeć" + 0.004\*"łącznie" + 0.004\*"zgłosić" + 0.004\*"zakres" + 0.004\*"wyznaczyć" + 0.004\*"wyrażić" + 0.004\*"unijnego" + 0.004\*"upływ" + 0.004\*"ust" + 0.004\*"przedstawiciel" + 0.004\*"ustawa" + 0.004\*"usługa" + 0.004\*"uzny" + 0.004\*"wadzący" + 0.004\*"wchodzić" + 0.004\*"wniosek" + 0.004\*"wpływ" + 0.004\*"wraz" + 0.004\*"wrzesień" + 0.004\*"wspólnotowy" + 0.004\*"wymóg" + 0.004\*"treść" + 0.004\*"prowadzącemu" + 0.004\*"polski" + 0.004\*"prowadić" + 0.004\*"dotychczasowy" + 0.004\*"dotyczyć" + 0.004\*"dzać" + 0.004\*"działalność" + 0.004\*"działo" + 0.004\*"dziennik" + 0.004\*"dopuszczeć" + 0.004\*"dzoć" + 0.004\*"euro" + 0.004\*"fizyczny" + 0.004\*"gospodarczy" + 0.004\*"grudzień" + 0.004\*"ilość" + 0.004\*"imię" + 0.004\*"elektroniczny" + 0.004\*"jeden" + 0.004\*"dokonać" + 0.004\*"dany" + 0.004\*"INANSÓ" + 0.004\*"INISTRA" + 0.004\*"POLSKIEJ" + 0.004\*"Prawo" + 0.004\*"RZECZYPOSPOLITEJ" + 0.004\*"RZĄDZENIE" + 0.004\*"dodawać" + 0.004\*"Skiba" + 0.004\*"Ustaw" + 0.004\*"Warszawa" + 0.004\*"administracja" + 0.004\*"brzmienie" + 0.004\*"część" + 0.004\*"dane" + 0.004\*"USTAW" + 0.004\*"ki" + 0.004\*"kierować" + 0.004\*"kodeks" + 0.004\*"pierwszy" + 0.004\*"pocztowy" + 0.004\*"podatkowy" + 0.004\*"podlegać" + 0.004\*"podwyższające" + 0.004\*"łączny" + 0.004\*"oznaczać" + 0.004\*"postać" + 0.004\*"pośredni" + 0.004\*"prawidłowy" + 0.004\*"prezes" + 0.004\*"pro" + 0.004\*"procedura" + 0.004\*"prowa" + 0.004\*"pować" + 0.004\*"opis" + 0.004\*"określić" + 0.004\*"ogłosić" + 0.004\*"kwota" + 0.004\*"listopad" + 0.004\*"marzec" + 0.004\*"mieć" + 0.004\*"nadawca" + 0.004\*"nastę" + 0.004\*"następować" + 0.004\*"nazwa" + 0.004\*"nazwisko" + 0.004\*"nieobjęte" + 0.004\*"nieprowadzyć" + 0.004\*"obejmować" + 0.004\*"obniżać" + 0.004\*"obrót" + 0.004\*"odrębny" + 0.004\*"prowadzić" + 0.004\*"życie"')

1. extract.py

Ten kod używa biblioteki KeyBERT do wyodrębniania słów kluczowych z tekstu wyciągniętego z pliku PDF. Po załadowaniu tekstu za pomocą pdfminer, model KeyBERT jest inicjalizowany, a następnie stosowane są dodatkowe parametry, takie jak analiza zarówno pojedynczych słów, jak i dwuwyrazowych fraz (ngramów), oraz zwiększenie liczby wyników do 40 słów kluczowych.

40

Słowa kluczowe (KeyBERT): ['sprawie zgłoszeń', 'ust zgłoszenie', 'dnia ogłoszenia', 'zgłoszeniem rozporządzenie', 'zgłoszenie', 'zgłoszeń', 'zgłoszenia', 'zgłoszenie celne', 'zgłoszenie działa', 'zgłoszeniu', 'przed złożeniem', 'zgłoszeń celnych', 'rozporządzenie sprawie', 'brzmieniu zgłoszenie', 'składane zgłoszenie', 'złożeniem', 'ogłoszenia', 'zgłoszeniem', 'wpływ na', 'kiego zgłoszenia', 'złożeniem po', 'działem', 'przesyłce', 'zmieniające rozporządzenie', 'tym zgłoszeniem', 'rozporządzenie wchodzi', 'zakresu działania', 'ustaw rzeczypospolitej', 'pkt rozporządzenia', 'składająca zgłoszenie', 'zgłoszenia oraz', 'przesyłki oraz', 'przesyłce nie', 'zatwierdzana przez', 'od zgłaszanych', 'zgłoszenie prowadzi', 'celnych dz', 'dotyczącej przesyłek', 'przesyłki', 'przesyłek przewożonych']

1. maska.py

Ten kod używa modelu językowego polish-roberta-large-v2 w ramach pipeline'u z biblioteki transformers do zadania typu "fill-mask" (wypełnianie maski). Pipeline jest używany do przewidywania słowa, które pasuje w miejsce <mask> w podanym tekście.

Tekst: Ustawa o <mask> publicznych.

Wyniki pipeline (wypełnianie maski): [{'score': 0.6053970456123352, 'token': 8062, 'token\_str': 'finansach', 'sequence': 'Ustawa o finansach publicznych.'}, {'score': 0.3608054220676422, 'token': 18615, 'token\_str': 'zamówieniach', 'sequence': 'Ustawa o zamówieniach publicznych.'}, {'score': 0.017560111358761787, 'token': 5368, 'token\_str': 'drogach', 'sequence': 'Ustawa o drogach publicznych.'}, {'score': 0.008454409427940845, 'token': 6827, 'token\_str': 'zamówień', 'sequence': 'Ustawa o zamówień publicznych.'}, {'score': 0.001021631876938045, 'token': 23112, 'token\_str': 'usługach', 'sequence': 'Ustawa o usługach publicznych.'}]

1. rake.py

Ten kod używa biblioteki RAKE (Rapid Automatic Keyword Extraction) do wyodrębnienia słów kluczowych z tekstu pobranego z pliku PDF. Po załadowaniu tekstu za pomocą pdfminer, model RAKE analizuje tekst, identyfikuje frazy kluczowe i sortuje je według istotności. Na końcu kod wypisuje 10 najważniejszych fraz

Frazy kluczowe (RAKE): ['które prowadzący ewidencję uzna za mające wpływ na prawidłowe określenie należności celnych', '2 pkt 2 rozporządzenia prezesa rady ministrów z dnia 18 listopada 2019 r', 'których łączna wartość w jednej przesyłce nie przekracza równowartości 150 euro', 'rozporządzenie wchodzi w życie po upływie 14 dni od dnia ogłoszenia', 'obrotu przesyłek przewożonych przez operatorów pocztowych innych niż operator wyznaczony', 'kiego zgłoszenia oraz udostępniana wraz z tym zgłoszeniem organowi celnemu', 'zawierających towary zwolnione z należności celnych przywozowych na podstawie art', 'w rozporządzeniu ministra finansów z dnia 8 września 2016 r', 'minister finansów kieruje działem administracji rządowej ‒ finanse publiczne', 'jest składane na podstawie danych zawartych w ewidencji prowa']

1. test.py

Ten kod wykonuje kilka kroków, aby wyodrębnić i przetworzyć słowa kluczowe z pliku PDF przy użyciu dwóch narzędzi: RAKE do ekstrakcji fraz kluczowych i spaCy do dalszej analizy tych fraz.

Słowa kluczowe (RAKE - pojedyncze słowa): {'kwoty', 'raz', '‒', 'inansó', 'operator', 'zwolnieniem', 'przesyłce', 'zgłoszenie', 'walucie', 'dza', 'zarzą', 'ustawy', 'następuje', 'ministrów', 'dnia', 'rządzenie', 'należności', 'ewidencja', 'określenie', 'udostępnia', 'stawce', 'działa', 'rzeczypospolitej', 'mowa', 'wysokości', 'sprzedaży', 'rozporządzenia', 'rady', 'towary', 'imię', 'wchodzi', 'przedstawiciel', 'działalności', 'osoba', 'życie', 'wartość', 'złożeniem', 'usług', 'zakresu', 'nazwę', 'numer', 'marca', 'zgłoszeniem', 'listopada', 'procedurą', 'prawo', 'objęcie', 'transakcji', 'ogłoszenia', 'osoby', 'rozporządzenie', 'operatorów', '§', 'nazwisko', 'podstawie', 'działem', 'obrotu', 'minister', 'wymogi', 'dane', 'waluty', 'ust', 'równowartości', 'sprawie', 'organ', 'podatku', 'dziennik', 'podlegają', 'działania', 'finanse', 'stawkę', 'postaci', 'kodeksu', 'organowi', 'oznacza', 'może', 'nadawcy', 'warszawa', 'brzmieniu', 'przepisów', 'ewidencję', 'administracji', 'części', 'odbiorcami', 'września', 'ewidencji', 'dodaje', 'treść', 'przesyłki', 'rozporządzeniu', 'danych', 'prowadzi', 'trybie', 'odbiorcy', 'zamówienia', 'dopuszczenia', 'towarów', 'zwolnień', 'ilość', 'wysokość', 'kod', 'inistra', 'prezesa', 'finansów', 'opis', 'kieruje', 'grudnia', 'wniosek', 'zgłoszeniu', 'zgłoszeń', 'poz', 'wpływ', 'przesyłek', 'adres', 'towaru', 'dni', 'ustaw', 'system', 'pro', 'euro', 'stycznia', 'ministra', 'zgłoszenia', 'przekracza', 'zp', 'upływie'}