Projekt 2 - raport

Sebastian Deręgowski, Bartosz Jamroży, Dawid Janus

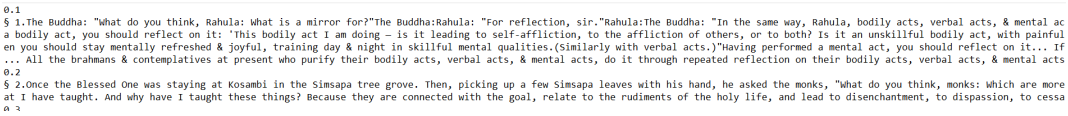
28 maja 2021

1 Opis problemu

Podczas prac nad tym projektem nasza grupa zajmowała się problemem przyporządkowania rozdziałów do ksiąg relijnych, dokładniej pogrupowania 590 rozdziałów w 8 ksiąg. Wykorzystaliśmy w tym celu modele uczenia maszynowego korzystających z metod klastrowania danych.

2 Opis zbioru danych

Zbiór danych dotyczy azjatyckich ksiąg religijnych. Dane pobraliśmy ze strony:
https://archive.ics.uci.edu/ml/datasets/A+study+of++Asian+Religious+and+Biblical+Texts
Większość świętych tekstów w tym zbiorze danych została pobrana z projektu „Gutenberg".



Rysunek 1: Fragment oryginalnych tekstów

Surowe dane, czyli rozdziały ksiąg religijnych zostały przez autorów przetransformowane do ramki danych. Utworzono kolumnę dla każdego słowa ze zbioru. Jeżeli dane słowo występuje w rozdziale, w odpowiadającą mu kolumnę wpisano liczbę wystąpień. Jeżeli nie to 0.

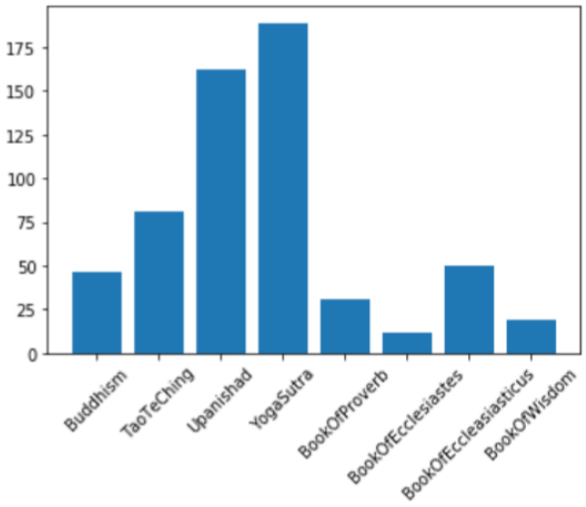
	# foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	convict	...	erred	thinkest	modern	reigned	sparingly	visual	tho
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
585	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
586	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
587	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	
588	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
589	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	

590 rows × 8266 columns

Rysunek 2: Ramka danych przygotowana przez autorów

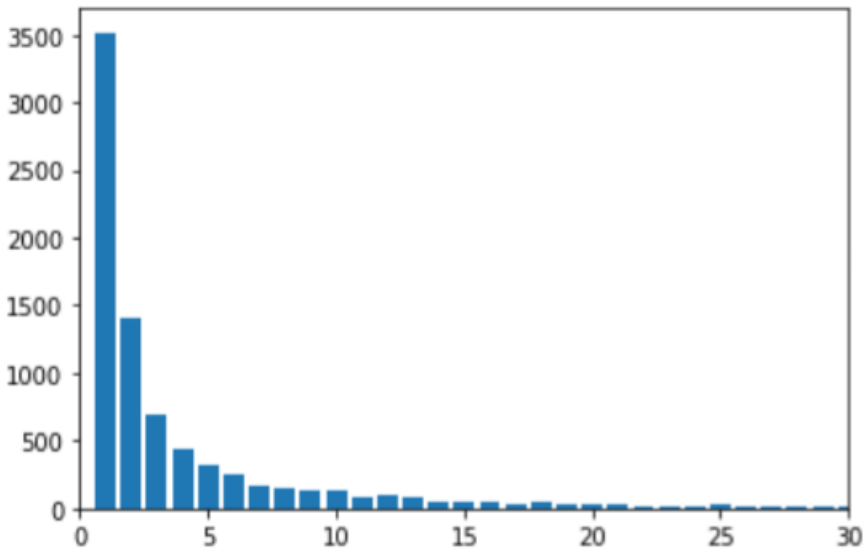
3 EDA

Na początku pracy popełniliśmy pewien błąd. Początkowo nie zdawaliśmy sobie sprawy z tego ze nie powinniśmy korzyść z informacji o tym, do jakiej religii przynależy dany rozdział. Korzystając z „niedozwolonych" informacji sprawdziliśmy liczność poszczególnych kategorii. Najliczniejsze okazały się pierwsze cztery księgi.



Rysunek 3: Ilość rozdziałów dla poszczególnych ksiąg

Doszliśmy także do wniosku, że słowa które pojawiają się tylko raz, nie będą pomocne przy podejmowaniu decyzji przez model. Zliczyliśmy słowa występujące daną ilość razy. Słów, które pojawiły się tylko raz, mamy około 3500. Jest to prawie połowa liczby wszystkich kolumn. Są to nieprzydatne cechy, nadające się do usunięcia. Jednak tego dokonaliśmy dopiero po stemmingu i lematyzacji opisanych w inżynierii cech.



Rysunek 4: Zliczenie liczby słów, które wystąpiły daną ilość razy

4 Inżynieria cech

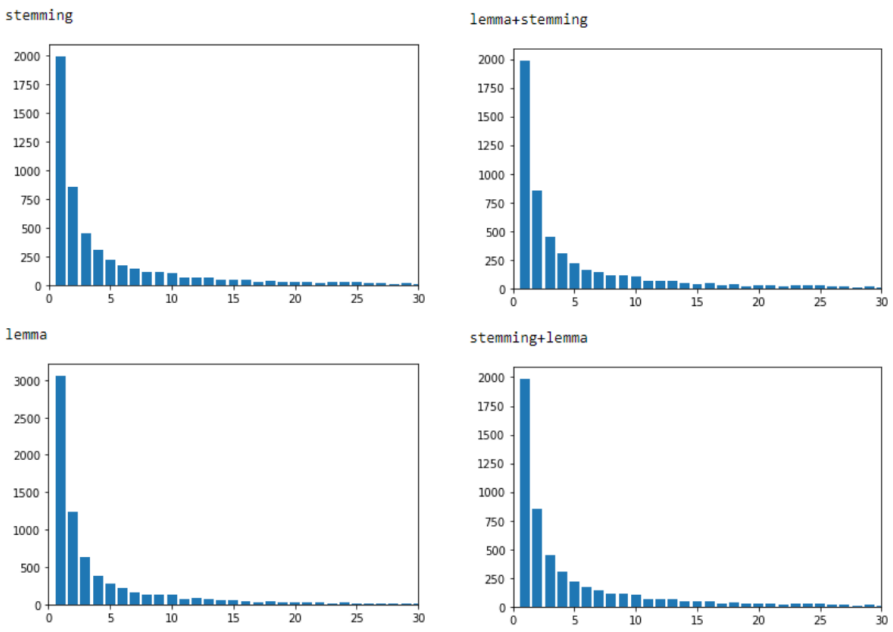
Podstawowym problemem, który musieliśmy rozwiązać, były słowa o tym samym bądź podobnym znaczeniu, ale zapisanych w różnych formach. Język angielski na szczęście nie rozróżnia przypadków rzeczownika, ale wciąż problematyczne były np. różne końcówki czasownika (*-ing,ed,s*) bądź rozróżnienie liczby pojedynczej od mnogiej.

W tym celu musieliśmy dokonać procesów lematyzacji i stemmingu słów z naszego zbioru. Procesy te są do siebie zbliżone, ale jednak nie takie same. Stemming ucina prefiksy i sufiksy, sprowadzając słowo do jego rdzenia. Lematyzacja z kolei sprowadza słowa do ich podstawowej formy, np. czasowniki do bezokolicznika.

Dokonaliśmy różnych transformacji naszej listy słów, zmieniając procesy i ich kolejność. Okazało się, że każda z transformacji kończy się inną liczbą słów:

1. Wyjściowa lista: 8266 słów
2. Lista po stemmingu: 5512 słów
3. Lista po lematyzacji: 7403 słów
4. Lista po stemmingu i lematyzacji: 5500 słów
5. Lista po lematyzacji i stemmingu: 5506 słów

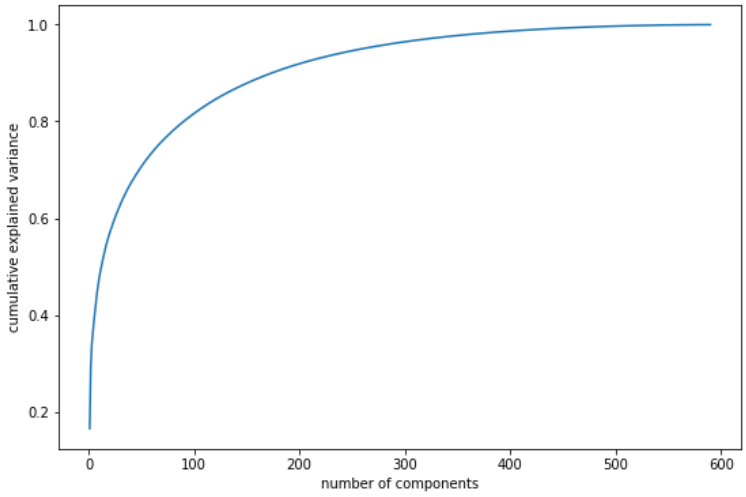
Przed rozwiązaniem tego problemu, zajęliśmy się kolejnym, nakreślonym już w poprzednim rozdziale, czyli słowami, które pojawiają się w tylko jednej obserwacji (rozdziale). Oczywistym jest, że słowa te nie pomogą nam w klasteryzacji, a z racji na to, że takich słów jest najwięcej (pokazuje to rys. 3) pozbycie ich jest dość pożądane. Sprawdziliśmy zatem jak zmieni się wykres z rys. 4 po zastosowaniu procesów transformacji słów wspomnianych wyżej:



Rysunek 5: Liczba słów powtarzających się daną liczbę razy po zastosowaniu odpowiedniej transformacji

Jak widzimy, sama lematyzacja usuwa dość mało pojedynczych wyrazów. Z tego powodu zdecydowaliśmy się na zastosowanie najpierw stemmingu, a potem lematyzacji. Następnym zaś krokiem było usunięcie słów, które pojawiły się w tylko jednym rozdziale.

Następnie próbowaliśmy wykorzystać algorytm PCA dla naszego zbioru. Na podstawie wykresu z rysunku 6. oszacowaliśmy liczbę komponentów na 500, co daje nam stopień kompresji równy około 0.14. PCA nie dało nam jednak żadnej pomocy - podczas modelowania szybko zauważyliśmy, że ramka danych po PCA jest tak samo klastrowana jak ta bez tego algorytmu.

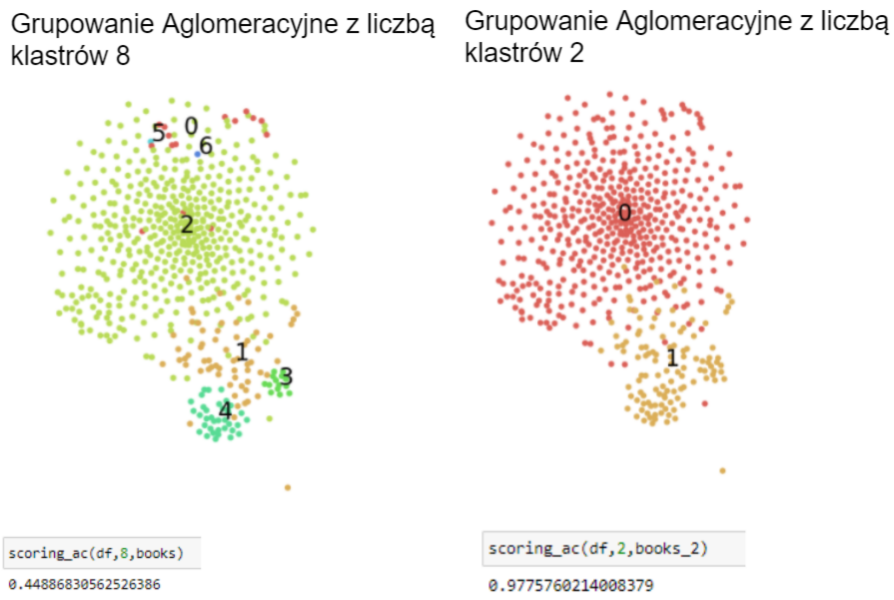


Rysunek 6: Algorytm PCA

5 Modelowanie

Pierwszym wyborem był algorytm DBSCAN, tak by pozwolić modelowi samemu zdecydować o liczbie klastrów. Jednak nie udało nam się tak dobrać parametrów promienia sąsiedztwa oraz liczby sąsiadów, by uzyskać więcej niż jeden klaster. Zaczęliśmy więc eksperymentować z innymi metodami. Podczas prac z różnymi metodami klasteryzacji z różnymi parametrami, zauważyliśmy powtarzającą się zależność - najbardziej naturalną liczbą klastrów jest 2. Stąd nasunęło nam się przypuszczenie, że 8 ksiąg możemy podzielić na dwie grupy, w obrębie których poszczególne księgi są niemal nierozróżnialne, natomiast obie grupy są ze sobą dość mało "spokrewnione".

Finalnie użyliśmy klastrowania aglometacyjnego, z parametrami `(linkage='ward',n_clusters=2)`. uznając, księgi *Buddhism*, *TaoTeChing*, *Upanishad* i *YogaSutra* za jedną księgę, a pozostałe cztery za drugą. W celu zbadania poprawności klasteringu, użyliśmy `v_measure_score` porównując wyniki modelu do oryginalnych etykiet. Otrzymaliśmy skuteczność blisko 98%. Jest to wynik o wiele lepszy niż ten sam algorytm dla ośmiu klastrów (czyli docelowej liczby ksiąg), gdzie wynik był bliski 45% (porównanie na rys. 7):



Rysunek 7: Porównanie klasteryzacji dla 2 i 8 klastrów

6 Podsumowanie

Podsumowując okazuje się, że wykonanie tego zadania poprawnie jest niesamowicie trudnym zadaniem. Być może zaawansowane formy przetwarzania języka naturalnego byłyby w stanie doprowadzić nas do lepszych rezultatów, ale przy użyciu znanych nam technik musimy zadowolić się tym, co uzyskaliśmy. Pocieszający jest natomiast fakt, że przy uproszczonym modelu dwóch ksiąg udało nam się osiągnąć blisko 100% skuteczność.