Books Clustering

Drazkowski Hubert, Wilk Marcin

Czerwiec 2021

1 Wstep

Pierwszym celem projektu jest zastosowanie metod analizy grupowania danych (ang. clustering).
Drugim celem projektu jest odpowiedź na pytanie jak podobne sa do siebie starożytne teksty religijne.
Analiza statystyczna została wykonana w programie Python. Dla czytelności i zwiezłości raportu niektóre opisy metod lub rysunki wykorzystywane podczas analizy zostały pominiete. Można je znaleźć w noatatniku jupyter. Wykresy do modelowania także w prezentacji.

Inspiracja do stworzenia danych wyjściowych był "Project Gutenberg".
Czyszczenie danych przeprowadzili PREETI SAH i ERNEST FOKOUE w (2019) "WHAT DO ASIAN RELIGIONS HAVE IN COMMON? AN UNSUPERVISED TEXT ANALYTICS EXPLORATION".

Dostepne były dwa źródła danych:

- Macierz rzadka gdzie zmiennymi byla liczba wystapien danego slowa w danym fragmencie
- Orginalny tekst ze wszystkimi fragemntami skonkatenowanymi

Fragmenty pochodza z 8 tekstów z czterech religii. Stad potencjalnie rodza sie dwie warstwy etykietek

- Hinduism : Yogasutras, Upanishads
- Buddishm : Four Noble Truth of Buddhism
- Taoism : Tao Te Ching
- Christianity : Book of Proverb, Book of Ecclesiastes, Book of Ecclesiasticus, Book of Wisdom

2 Eksploracyjna analiza danych i inżynieria cech

Wnioski po eksploracyjnej analzie danych macierzy zliczeń słów:

1. Jest bardzo dużo słów, które rzadko wystepuja, ponad 50% z nich wystepuje 2 razy lub rzadziej.
2. Mamy do czynienia z problemem dużej ilosci wymiarów. Obserwacji jest 589, a zmiennych 8266
3. Niektóre słowa pojawiaja sie czesto, rzadko zdażaja sie słowa, które wystepuja wiele ilości razy we fragmencie
4. Fragmenty różnia sie iloscia zawieranych słów. Zdecydowana wiekszość jest w przedziale 0-200. Potem jest kilkadziesiat do 400, powyżej 400 jest niewiele obserwacji.

Można zauważyć, że mamy tutaj problem przekleństwa wymiarowości. Mamy do czynienia z macierza o wymiarach 590 na 8266, na dodatek jest to macierz rzadka, to jest pełna zer. Kolumny stanowia indykatory zliczajace wystapienie danego słowa w korpusie. Każdy jeden fragment korpusu jest obserwacja. Zobaczmy czy jakieś charakterystyczne słowa wystepuja najcześciej w tych tekstach.
Niektóre pojawiaja sie naprawde czesto, nie sa one jednak na pierwszy rzut oka charakterystyczne dla jakichkolwiek tekstów. Może być jednak tak, że okaże sie zupełnie co innego, bo tylko cześć z ksiag bedzie używała składni do której beda wymagane niektóre z tych słów. Przyjrzyjmy sie z kolei tym rzadziej wystepujacym słowom, bedziemy tutaj szukać czy da sie tutaj jakoś zmniejszyć wymiar naszych danych przez usuniecie słów, które wystepuja bardzo rzadko albo w ogóle.

Po przeczyszczeniu orginalnego tekstu tak aby pozbyć sie dziwnych znaków i uzyskać mniej wiecj orginalny podział (z dokładnościa do pewnych słów nieinformatywnych usunietych przez autorów artykułu) tworzymy nowa ramke danych. Ilość korpusów zgadza sie z wielkościa macierzy. Analiza sentymentu zawiera miedzy innymi dwa wskaźniki polarity i subjectivity. Polarity wpada w przedział [1,1] , a im niższa wartość tym tekst oceniany jest za bardziej nacechowany negatywnie zaś subjectivity w [0,1] , a interpretować należy te skale w taki sposób, że wartości bliższe zera powinny wskazywać na mniej nacechowane subiektywnie treści.

Im wieksze ARI tym bardziej skomplikowany tekst Im wieksze FRI tym tekst łatwiejszy Powinny być ze soba ujemnie skorelowane ze wzledu na wspólny czynnik we wzorze

$$ARI = 4.71(\frac{characters}{words}) + 0.5(\frac{words}{sentences}) - 21.43,$$

$$FRE = 206.835 - 1.015(\frac{words}{sentances}) - 84.6(\frac{syllabes}{words}).$$

Sentiment : https://github.com/sloria/TextBlob, leksykon zdefiniowany subiektywnie przez współczesnych ekspertów.

Nie ma sensu zliczać ilości znaków lub słów. Bardziej zależy nam na cechach jezyka i używanego słownictwa, przekazywanych treściach. Te sama myśl przekazuja w swojej konstrukcji współczynniki ARI i FRE. Dodamy jedynie średnia długość użytego słowa w zdaniu.

Pozwoliliśmy sobie na rzucenie okiem także na orginalny plik tekstowy, w nastepnych krokach rysujemy ładnie wygladajaca mape wyrazów najcześciej sie pojawiajacych i przykładamy do owego tekstu dwa narzedzia rodem z NLP to jest wskaźnik subjectivity i polarity

- Polarity (użycie wiekszości słów nacechowanych negatywnie albo pozytywnie)
- Subjcetivity (użycie słów uważanych za opiniotwórcze)
- FRE (indeks mierzacy skomplikowanie tekstu)
- ARI (indeks mierzacy skomplikowanie tekstu)
- avg_wordlength (długość średnia słowa we fragmencie)

Naszym głównym celem w tej cześci była redukcja wymiarów oraz dodanie kilku cech przy pomocy NLP na podstawie oryginalnych tekstów. Pierwszym podejściem było znalezienie słów, które wystepuja tylko w jednej ksiedze. Dla każdej ksiegi policzylismymy ile takich różnych słów zawiera oraz ile w sumie. Nastepnie usuneliśmy te słowa z naszej ramki danych i przeprowadzimy PCA.

Na orginalnej macierzy dokonaliśmy kilku przekształceń:

- Aby zmniejszyć liczbe wymiarów usuwamy słowa ktre wystapiły co najywżej dwa razy w całym korpusie. (zostało 4394)
- Aby jakoś zachować informacje mierzymy ile razy wystapiło jakiekolwiek unikatowe (¡2) słowo w danym fragemncie. Do tego dodajemy ile takich unikatowych słów miał fragment.
- Na tym robimy PCA, uzyskujac 91% tłumaczonej wariancji zostajemy przy 200 składowych głónych

Ostatecznie przed modelowaniem

- Łaczymy ze soba ramke danych po PCA i z orginalnej analizy tekstu za pomoca narzedzi NLP.
- Wystandaryzowujemy niektóre zmienne, bo nowo dołaczone cechy miały znacznie wieksze wartości od tych zwróconych przez PCA, tak aby wszystkie kolumny były mierzone w podobnej skali.

3 Modelowanie

Patrzyliśmy na beterie metod :

1. Kmedioids
2. Kmeans
3. Agglomerative z połaczneiami
   - Single
   - Average
   - Complete
   - Ward
4. DBSCAN
5. GMM

Nastepnie dodaliśmy zdefiniowana prze nas metryke $sum(|x-y|/|x+y|)$. Znowu patrzyliśmy na wyżej wymieione metody. Inspiracja do jej stworzenia była wieksza cheć zaakcentowania różnic na małych liczbach słów. W metryce euklidesowej jeśli w jednej ksiedze słowo x jest 0 razy a w drugiej 4, natomiast inne słowo w pierwszej ksiedze jest 48 razy a w drugiej 52, to wpływaja one tak samo na odległość miedzy tymi ksiegami. W nowej metryce pierwsze słowo dodawałoby 1 do metryki natomiast drugie słowo tylko 0.04.
Porównywaliśmy metody za pomoca wykresu łokciowego i wykresu miary Silhouette. Silhouette score by dla własnej metryki dużo wyższy niż w przypadku metryki euklidesowej, jednak nie powinniśmy jednoznacznie stwierdzać wyższości tej metody, gdyż silhouette jest wrażliwy na metryke użyta do jego obliczania. Na sam koniec pojedliśmy sie próby wizualizacji wyników za pomoca t-SNE i policzenia homogeniczności używajac orginalnych klas.

4 Wnioski

Porównanie wykonywaliśmy na dwóch etykietach. Etykietach przynależnosci do religii i etykietach przynależnosci do konkretnej ksiegi. DBSCAN wyłapał tutaj tylko jeden duży klaster i ewentualne pojedyncze klasy odstajace. Nie ma sensu tutaj nawet rozważać homogenity. Nasza własna metryka pomimo dobrego silhouette daje bardzo słabe wyniki dla naszego zbioru danych. Nie potrzebujemy liczyć homogenity, żeby to zobaczyć, wystarczy spojrzeć na t-SNE. Wybraliśmy do ostatecznego porównania modele :

- AgglomerativeClustering(n_clusters=6,linkage="average")
  - Homogenity religion : 0.38
  - Homogenity book : 0.29

- GaussianMixture(n_components=8, covariance_type="full")
  - Homogenity religion :0.40
  - Homogenity book : 0.32
- KMeans(n_clusters=6,random_state=0)
  - Homogenity religion : 0.39
  - Homogenity book : 0.30