WikiDocker - sprawozdanie

Autorzy: Jan Kornacki, Hubert Lewandowski, Grzegorz Pozorski

Wstęp

Wikidocker, czyli szybki i nie tak naiwny klasyfikator Bayesa, jakby mogło się wydawać. Program ma kwalifikuje artykuł anglojęzyczny do najbardziej prawdopodobnej kategorii z 6 głównych, dostępnych na Simple Wikipedia.

Kategoria wyjściowa byłaby rekomendacją, która mogłaby znaleźć zastosowanie w kategoryzowaniu choćby nowych artykułów. Przykładowo taki program mógłby mieć zastosowanie do automatyzacji tworzenia nawigacji na stronach internetowych z olbrzymimi bazami danych.

Przygotowanie

Wykorzystana technologia

W projekcie użyty został naiwny klasyfikator Bayesowski wsparty techniką *stemmingu* (usuwanie końcówki wyrazu zostawiając tylko jego temat) do kwalifikacji wiadomości.

Wzór, którego użyliśmy, wygląda następująco:

$$P(X_i|y) = \frac{N_{X_i|y} + \alpha}{N_y + \alpha \cdot |V|}$$

Opis symboli:

Symbol	Opis
$P(X_i\|y)$	Prawdopodobieństwo warunkowe słowa X_i należącego do klasy y
$N_{X_i\|y}$	Liczba wystąpień słowa X_i we wszystkich fragmentach należących do klasy y
N_y	Liczba słów we wszystkich fragmentach należących do klasy y
$\|V\|$	Całkowita liczba słów w słowniku
α	Parametr służący uniknięciu prawdopodobieństw zerowych

Biblioteki

Do wyciągnięcia artykułów dla zbioru uczącego posłużyliśmy się bibliotekami:

- requests - obsługuje żądania, które były kierowane do strony Wikipedii
- Beautiful Soup - w celu wyciągnięcia konkretnych danych
- sklearn - do wygenerowania zbioru testowego

> W celu uniknięcia każdorazowego pobierania danych i uczenia kwalifikatora użyliśmy serializacji słowników do pliku.

Zbiór uczący

Zbiorem uczącym słowniki są **wszystkie** artykuły z 6 głównych kategorii na stronie Simple Wikipedia.

Zbiór testowy

Dane o poprawności niniejszego kwalifikatora opierały się na zbiorze testowym wygenerowanym ze zbioru uczącego.

Proces obliczeniowy kwalifikatora

Pobranie artykułów

Jest to **najdłuższy** proces ze wszystkich. W obecnej wersji ten proces może zająć nawet kilkadziesiąt minut!

Nie jest on jednak brany pod uwagę, ponieważ można go zrobić tylko raz (lub też w sytuacji gdy chcemy zaktualizować bazę artykułów) i bazować na zserializowanym "dumpie".

Przy pobieraniu, każdy artykuł przechodzi przez proces usuwania półsłówek (np przysłówków, przymiotników, określniki, itp), które źle wpływają na rezultaty. Dzięki temu, słowa kluczowe mają większą szansę w słownikach.

Dane o wyrazach są wyciągnięte z *wybranych* artykułów dotyczących kategorii "Angielskie Lematy".

Generowanie zbioru uczącego i testowego

Drugi proces pod względem prędkości. Jest również pomijalny w użytku "codziennym" ponieważ po wygenerowaniu zbiorów i nauczeniu słowników, proces jest wymagany tylko przy zaktualizowanym dumpie.

Uczenie słowników

Najszybszy proces inicjalizacji całego kwalifikatora. Mimo, że najszybszy to również pomijalny dzięki serializacji słowników do plików.

Podsumowując

Program kolejno:

1. Generuje słowniki z plików
2. Pobiera paragrafy z artykułu podanego na wejściu
3. Przetwarza dane, w celu uzyskania formy akceptowalnej przez algorytm
4. Kwalifikuje artykuł do kategorii, dla której określono największe prawdopodobieństwo

> Mając gotowe słowniki klas proces kwalifikacji będzie bardzo szybki, co jest kluczowe dla tego projektu.

Wyniki

Wyniki dla zbioru testowego

Wynik poprawności nauczonego kwalifikatora odnosi się do proporcji 0.8 w generowaniu zbiorów testowych. Przy kwalifikowaniu artykułów użyta jest proporcja 0.999 tak aby zbiór treningowy był jak największy.

Poniższe dane określa najlepszą poprawność decyzji, podjętej przez kwalifikator przy proporcji zbiorów 0.8.

```
### CORRECTNESS ###
50,373% of articles was qualified correctly.
```

Przykłady artykułów testowych

Applied sciences:

- Alaska: The Final (Architectural) Frontier
- Modern Architecture

People and social studies:

- Scientists Say: Placebo
- Common Mental Health Problems Students May Face This Year

Government and law :

- Poland's judges forced into retirement purgatory – another blow to justice
- The Effectiveness of the International Criminal Court: Challenges and Pathways for Prosecuting Human Rights Violations

Natural sciences and maths:

- The pebbled path to planets
- Scientists Say: Pollen

Daily life, art and culture:

- Pop art
- In the Kitchen: Artist Jean Shin Shares the Korean Dumpling Recipe That Embodies the Same No-Waste Philosophy as Her Art

Religions and beliefs:

- Buddhism: Basic Beliefs
- Islam

Czy rezultat jest zadowalający?

Biorąc pod uwagę, że szansa wylosowania prawidłowej kategorii bez żadnej wiedzy, jest równa ~17%, wyniki sięgające nawet 50% (średnio 47%) są ogromną wartością dodaną, przy tak niskim czasie samej decyzji.

Wynik jest na poziomie losowego prawdopodobieństwa wybrania jednej z dwóch klas, mimo że mamy 6!