Politechnika Krakowska

Wydział Informatyki i Telekomunikacji

Kierunek: Informatyka

Kraków, 15.05.2023

Semestr: VIII

Przetwarzanie Języka Naturalnego

Wyszukiwanie duplikatów

Autor: Aleksander Kuzmich

Abstrakt

W ramach realizowanego projektu zostanie podjęta próba zrealizowania nieskomplikowanego obliczeniowo sposobu na sprawdzenie czy podane dwa zdania są duplikatami, czyli niosą taką samą informację.

Wykorzystane dane

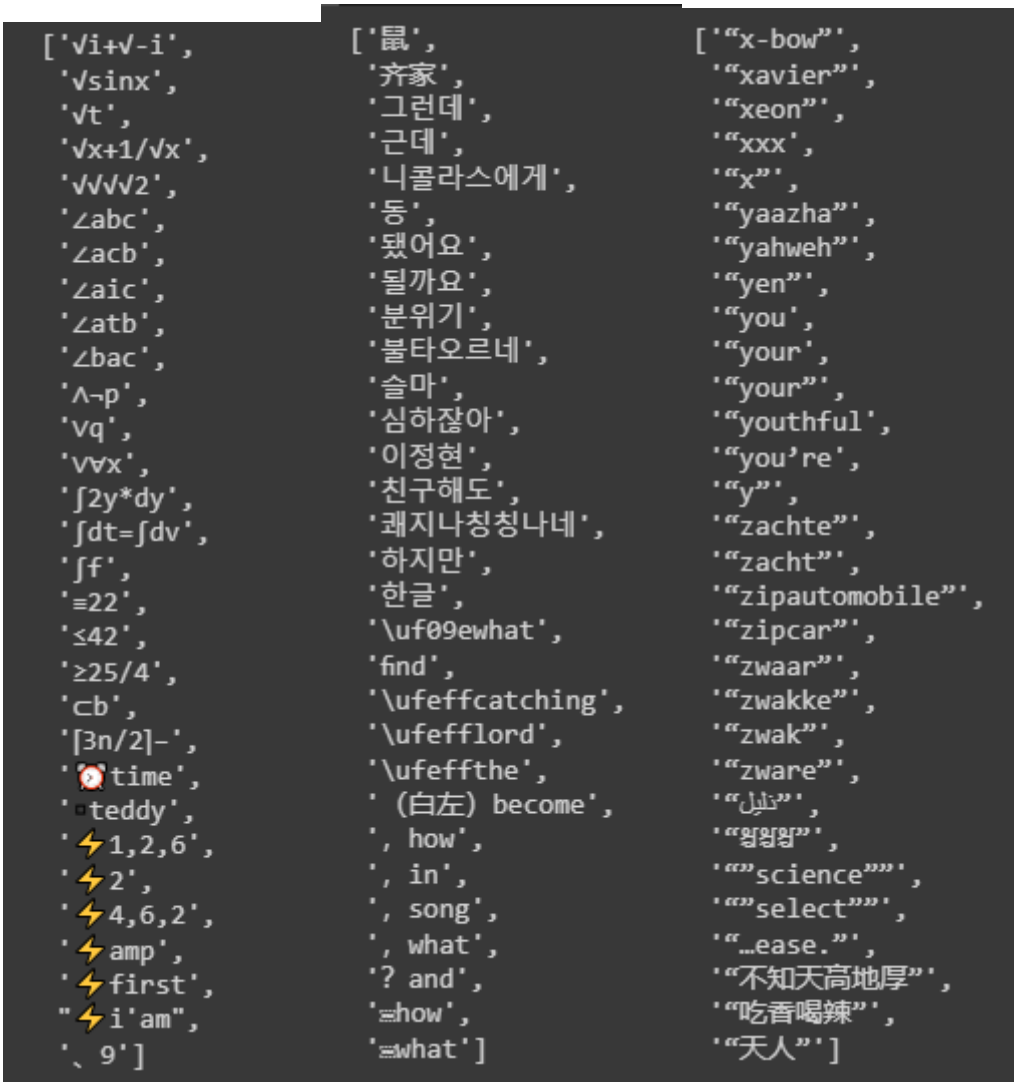
Dla realizacji projektu jest potrzebny zbiór danych (dalej nazywany datasetem), który zawierałby pary zdań oraz informację o tym, czy są one duplikatami. Tym wymaganiom odpowiada zbiór pytań „Quora Question Pairs” (link). Pary podanych pytań zostały olabelowane przez wybrane osoby, co powoduje, że jest to ocena subiektywna konkretnej osoby [1].

Zbiór danych m. in. posiada kolumny „question1”, „question2” – tekstówe dane oraz kolumnę „is_duplikate”, która zawiera liczbę 1 w przypdaku, gdy dwa pytania są duplikatami i 0, gdy zostało stwierdzone, że nie są duplikatami. Wczytany do ramki DataFrame z biblioteki pandas plik z danymi przezentuje się następująco:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and c...	I'm a triple Capricorn (Sun, Moon and ascendan...	1
6	6	13	14	Should I buy tiago?	What keeps childern active and far from phone ...	0
7	7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	8	17	18	When do you use シ instead of し?	When do you use "&" instead of "and"?	0
9	9	19	20	Motorola (company): Can I hack my Charter Moto...	How do I hack Motorola DCX3400 for free internet?	0

Przygotowanie zbioru danych

Dla oceny, czy podane dwa zdania są duplikatami, one powinny zostać przetworzone na wektory liczbowe. Przede wszystkim, one muszą być rozdzielone na osobne słowa, które dalej będą nazywane „tokenami”. Istotne jest, że dane z wybranego zbioru są zadanymi przez użytkowników strony Quora pytaniami, co oznacza, że mogą zawierać literówki, znaki punktuacji, różne odmiany tego samego słowa. Te zdania mogą zawierać też różnego rodzaju znaki, które mogą zakłócać dalszą pracę. Na poniższych zrzutach ekranu są przedstawione niektóre słowa, które znalazły się w zbiorze i mogą stanowić problem przy przetwarzaniu tekstu:



Na potrzeby przetworzenia tak zróżnicowanych tekstów, w ramach projektu została stworzona klasa „Tokenizer”, która odpowiada za podział pytań na tokeny, wykorzystując TreebankWordTokenizer z biblioteki scikit-learn, zamianę skrótów takich jak „i’m” na „i am”, usunięcię często powtarzających się słów, które mają mały wpływ na sens zdania. Ważnym etapem jest zamiana dużych literw słowach na małe w taki sposób, żeby zminimalizować utratę ich znaczenia – zaimplementowana została zamiana dużych liter w pierwszym zdaniu słowa. Alternatywnie jest dostępna opcja zamiany wszystkich dużych liter. W celu sprowadzenia różnych odmian słowa do „standardowej” postaci został użyty WordNetLemmatizer z biblioteki scikit-learn.

Po zastosowaniu wszystkich etapów czyszczenia tekstu, ważne jest sprawdzenie, czy ilość pozostałych tokenów jest wystarczająca dla ich dalszego użycia. W przypadku, gdy ilość tokenów zdania jest mniejsze określonej ilości, która w przypdaku opisywanego projektu została ustalona na 4, zdanie wraz z jego parą zostaje usunięte z datasetu. Opisane metody czyszczenia zostały wybrane i zaimplementowane na podstawie [2].

Po zastosowaniu wyżej opisanych etapów czyszczenia tekstu, ilość unikatowych słów wyniosła 91260.

Wektoryzacja

Wyczyszczone i przetworzone na listy tokenów zdania, mogą być skonwertowane na liczbowe wektory, z wykorzystaniem metody TFIDF, która oblicza wartośc dla każdego tokena na podstawie jego częstotliwości w zdaniu oraz w całym zbiorze zdań [2]. Tak przedstawione zdania są wektorami w wielowymiarowej przestrzeni, co umożliwia zastosowanie różnych metod matematycznych, jedną z których jest obliczenie wartości kosinusa kąta między dwoma wektorami. W przypdaku, gdy ta wartość jest równa 1, wektory są skierowane w jednym kierunku i można uznać, że mają takie same znaczenie. W praktyce należy przyjąć jakąś granicę [-1;1], wartości powyżej której oznaczają, że zdania mają ten sam sens i są duplikatami [2].

Metryki

Użyty dataset cechu się tym, że ma tylko dwie klasy: 0 – zdania nie są duplikatami, 1 – zdania są duplikatami. Dla wygodnego oszacowania skuteczności wybranych i zaimplementowanych metod, została wybrana metryka „confussion matrix" z biblioteki scikit-learn. W przypadku klasyfikacji dwóch klas, ta metryka zwraca cztery wartości: ile par nie duplikatów zostało oszacowanych jako nie duplikaty (prawidłowo oszacowane nie duplikaty), ile nie duplikatów zostało oszacowanych jako duplikaty (błędne oszacowania), ile duplikatów zostało oszacowane jako nie duplikaty (błędne oszacowania) oraz ile duplikatów zostało oszacowane jako duplikaty[3].

Analiza wyników

Została wybrana próbka par o ilości 10000 i dla takiego zbioru zaimplementowana metoda jest w stanie prawidłowo oszacować około 65% elementów. Biorąc pod uwagę to, że przypdakowe przydzielanie klas 0 i 1 daje 50% prawidłowych wyników, prosta wektoryzacja list tokenów za pomocą metody TFIDF pokazuje lepsze wyniki, które można spróbować podnieść ulepszeniem klasy Tokenizer oraz doborem dokładnej granicy, powyżej której para zdań jest uznawana za duplikat

Bibliografia

[1] https://www.kaggle.com/competitions/quora-question-pairs/overview dostęp [15.05.2023]

[2] Lane Hobson, Cole Howard, Hannes Hapke, 'Przetwarzanie języka naturalnego w akcji', PWN 2021

[3] https://en.wikipedia.org/wiki/Confusion_matrix dostęp [15.05.2023]