Komputerowe systemy rozpoznawania

Mateusz Grotek 186816 Paweł Tarasiuk 186875

1 Wersja wstępna algorytmu

Główną ideą na której chcemy oprzeć naszą klasyfikację są słowa kluczowe. Zbiór słów kluczowych dzielimy na następujące podzbiory:

- nazwy geograficzne
- imiona i nazwiska
- nazwy walut
- nazwy instytucji państwowych
- nazwy/skróty nazw firm wraz z typem spółki (jeśli występuje)
- produkty powiązane z konkretnymi krajami

Klasyfkacja ma za zadanie przyporządkować każdy tekst do jednej z poniższych klas:

- west-germany
- usa
- france
- uk
- canada
- japan

Tworzymy 36 zbiorów rozmytych, po jednym dla każdej kombinacji podzbioru słów kluczowych i klasy. Do każdego zbioru rozmytego dla określonej kategorii i określonej klasy wrzucamy wszystkie słowa kluczowe z tej kategorii i początkowo przypisujemy im 0,0 jako wartośc funkcji przynależności do zbioru. Następnie przeglądamy zbiór uczący i do każdego elementu elementu przypisujemy stosunek pojawiania się danego słowa kluczowego w danej klasie względem jego pojawiania się we wszystkich klasach. Na przykład, jeżeli słowo pojawiło się w tekstach dotyczących kanady x razy, a we wszystkich tekstach pojawiło się y razy, to stosunek wynosi x/y. Stosunek ten mówi jak dobrze słowo pasuje do podanej klasy. Na wejście kNN podajemy wektory uczące zawierające 36 wartości. Wektory te obliczamy dla tekstów uczących jako sumę wartości funkcji przynależności wystąpień każdego słowa podzieloną przez długość danego wektora.

Algorytm ten jest rozszerzoną wersją prostszego algorytmu, który po prostu zlicza wystąpienia słów kluczowych w tekście. Dzięki uwzględnieniu rozmytości nie ma problemu ze słowami, które dobrze opisują więcej niż jedną klasę. Na przykład angielskie nazwiska pasują równie dobrze do wielkiej brytanii, jak i usa, więc będą miały mniejszą wagę, niż nazwiska japońskie przy rozpoznawaniu kraju.

Dodatkową regułą, którą chcielibyśmy zastosować jest domyślne przypisanie do klasy usa w wypadku, gdy wartości dla innych klas nie przekraczają pewnej wartości. Reguła ta będzie szczególnie użyteczna w wypadku rozróżnienia między usa i uk.