Case study 1.11 (ZED – TPD 2009)

Klasyfikacja wiadomości tekstowych

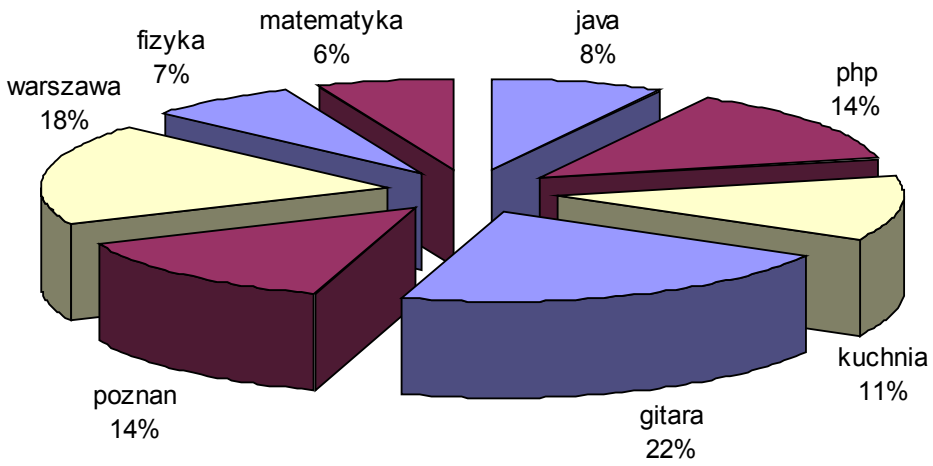
Ogólnym celem jest stworzenie systemu automatycznie klasyfikującego wiadomości tekstowe (typu emajle) pochodzące z grup dyskusyjnych.

Opis danych:

Wiadomości tekstowe pochodzą z archiwum polskojęzycznych Newgroups. Wiadomości obejmują okres czasu od 1 marca 2005r. do 20 kwietnia 2005r i zostały pobrane z archiwów poniższych grup dyskusyjnych:

- pl.comp.lang.java,
- pl.comp.lang.php,
- pl.rec.kuchnia,
- pl.rec.muzyka.gitara,
- pl.regionalne.poznan,
- pl.regionalne.warszawa,
- pl.sci.fizyka,
- pl.sci.matematyka.

Nazwy tych grup tworzą etykiety klas decyzyjnych. Łącznie zbiór danych NewsGroups zawiera 36260 wiadomości lecz rozkład wiadomości w klasach jest niezrównoważony – spójrz na poniższy rysunek:



Oryginalne postacie plików emajli zostały poddane przetwarzaniu wstępnemu tak, aby wyłuskać informacje dotyczące nagłówka wiadomości oraz samej treści. Następnie wobec tekstu zastosowano metody przetwarzania języka naturalnego, gdzie trzeba było podjąć decyzję między innymi co do: konfiguracji parametrów takich jak:

- analiza języka (lematyzacja,..),
- metoda rozdzielania poszczególnych wyrazów (tokenizacja),
- pomijanie znaczników języka HTML.

Doprowadziło to do powstania kilku wersji plików – w załączeniu otrzymujesz dwie możliwości (kiedy wszystkie parametry są włączone lub są wyłączone).

W dalszej kolejności zbudowano zbiór atrybutów – większość z nich to częstości wystąpienia termów – na ogół pojedynczych słów. Inne to atrybuty liczbowe dotyczące występowanie określonych wartości parametrów w nagłówkach emejli. W stosunku do atrybutów typu termy dokonano selekcji poprzez odrzucenie najmniej dyskryminujących klasy (najrzadziej lub najczęściej występujących w emejlach).
Możliwa jest dalsza selekcja atrybutów, jeśli doprowadzi to do skuteczniejszego klasyfikatora.

Uwagi do analizy:

Ogólnym celem jest skonstruowanie jak najskuteczniejszego systemu uczącego się przydzielania wiadomości do poszczególnych katalogów – klas.

W ogólności oczekujemy skuteczności na poziomie minimum 80%

Przy czym oprócz globalnej jak najwyższej zdolności predykcyjnej, istotna jest analiza skuteczności rozpoznawania poszczególnych kategorii. Nie jest akceptowalna sytuacja, gdzie pomimo wysokiej globalnej trafności, któraś z kategorii jest rozpoznawana poniżej 50%.

Dodatkowo należy uwzględniać rozmiar danych i czas związany z jego przetwarzaniem. Poszukiwanie rozwiązania w czasie większym niż kilkanaście minut jest niekorzystne.

Warto zwrócić uwagę na

(a) ocenę ważności poszczególnych atrybutów dla zdolności rozpoznawania klas, oraz ewentualną redukcję,

(b) konstrukcja tzw. klasyfikatora, czyli wskazań dla podejmowania końcowej decyzji – wymaga porównania wielu sposobów konstrukcji takich klasyfikatorów opartych na zróżnicowanych algorytmach uczących.

(c) ze względu na zadanie automatycznej klasyfikacji, forma symbolicznej prezentacji wyników nie jest konieczna.

(d) należy się zastanowić nad wyborem właściwych miar oceny zdolności predykcyjnej.

(e) przemyślenia sposobu przetwarzania tak dużych wolumenów danych.

Ponadto trzeba być świadomym ograniczeń wydajnościowo-pamięciowych związanych z wczytaniem i przetwarzaniem tak dużych zbiorów danych.
Na przykład w oprogramowaniu WEKA zamiast interfejsu graficznego można wykorzystać tryb uruchomienia poszczególnych składników z linii komend.

Inne zagadnienia przekaże prowadzący – konsultant.