MMDS Challenge – Projekt eksploracji danych

Raport I

Organizacja, wstępne przetwarzanie danych oraz wyszukiwanie najbliższych sąsiadów

Imię Nazwisko Imię Nazwisko

31 października 2016

1 Konkurs Roq.ad (10p)

1.1 Sformułowanie problemu

Krótki i zwięzły opis zadania konkursowego.
(maks. 1/2 strony)

1.2 Opis danych

Opis danych udostępnionych w materiałach do pierwszych zajęć wraz z podstawowymi statystykami: rozmiar, liczba atrybutów, liczba przykładów, liczba klas, dominująca klasa, procent niezerowych wartości cech.
(maks. 1/2 strony)

1.3 Reprezentacja danych w pamięci zewnętrznej i operacyjnej

Krótki opis (bardzo zwięzły i konkretny) z uwzględnieniem:

- opisu reprezentacji z jej zaletami i wadami, również z perspektywy eksperymentów, które należy przeprowadzić,
- dyskusji na temat czasu utworzenia i objętości reprezentacji danych,
- dyskusji na temat czasu wczytywania do pamięci, dostępu do danych i objętości danych,
- dyskusji na temat napotkanych trudności.

Można opisać więcej niż jedno rozwiązanie w celach porównawczych.
(maks. 3/4 strony)

1.4 Standaryzacja danych

Wyniki eksperymentów wraz z odpowiednimi komentarzami i wnioskami.

Podaj dla pierwszej cechy średnie, odchylenia standardowe, wartość najmniejszą i największą, oraz wartość najmniejszą i największą spośród wszystkich wartości cech, dla danych surowych, oraz dla obu wymienionych podczas drugich zajęć sposobów ich przetwarzania.

Przykładowa tabela w LaTeX-u została pokazana w Tabeli 1.

Nazwa	średnia	odchylenie	min.	max.	min. globalne	max. globalne
nazwa	0.00	1.00	-1	1	-1	1
nazwa	0.00	1.00	-1	1	-1	1
nazwa	0.00	1.00	-1	1	-1	1

Tablica 1: Wyniki eksperymentalne

(maks. 1/2 strony)

1.5 Podsumowanie

Na końcu jest zawsze miejsce na krótkie podsumowanie (maks. 1/4 strony).

2 Konkurs Data Ninja (35p)

2.1 Sformułowanie problemu

Krótki i zwięzły opis zadania konkursowego.

(maks. 1/2 strony)

2.2 Opis danych

Opis danych udostępnionych w ramach konkursu Data Ninja, wraz z podstawowymi informacjami o nich, takimi jak rozmiar, liczba przykładów, opis struktury danych.

Warto przedstawić przykładowe dane.

(maks. 1/2 strony)

2.3 Reprezentacja danych w pamięci zewnętrznej i operacyjnej

Krótki opis (bardzo zwięzły i konkretny) z uwzględnieniem:

- opisu reprezentacji z jej zaletami i wadami, również z perspektywy eksperymentów, które należy przeprowadzić,

- dyskusji na temat konfiguracji i administracji danego rozwiązania,
- dyskusji na temat czasu utworzenia i objętości reprezentacji danych,
- dyskusji na temat czasu wczytywania do pamięci, dostępu do danych i objętości danych,
- dyskusji na temat napotkanych trudności.

Można opisać więcej niż jedno rozwiązanie w celach porównawczych.

Jeżeli dane zostały zorganizowane w dedykowany sposób dla danego eksperymentu, to należy to odpowiednio zaznaczyć i opisać.

(maks. 1 strona)

2.4 Wstępne przetwarzanie danych

W dalszej części raportu będzie należało opisać wyniki eksperymentu przeprowadzonego na zbiorze danych przygotowanym z surowych danych training.zip. W tym rozdziale należy opisać sposób przeprowadzenia ekstrakcji cech, użytą reprezentację danych oraz napotkane problemy.

Tytuły ofert z surowych danych o ogłoszeniach należy przekształcić do reprezentacji *bag of words*. Przekształcenie należy wykonać używając 100 najczęściej występujących słów w tytułach. Można, ale nie jest to konieczne, przeprowadzić uprzednio stemming lub lematyzację. Eksperymenty należy przeprowadzić tylko na ogłoszeniach, w których tytułach występuje przynajmniej jedno słowo ze zbioru 100 najczęściej występujących słów – należy odrzucić pozostałe ogłoszenia ze zbioru danych na potrzeby tego eksperymentu.

(maks. 1 strona)

2.5 Eksperyment

W tym rozdziale należy umieścić wyniki eksperymentów wraz z odpowiednimi komentarzami i wnioskami. Należy wykonać następujące zadania:

1. Znalezienie najczęstszych kategorii ofert pogrupowanych po cechach i po parach cech,
2. Wyszukiwanie najbliższych sąsiadów dla ofert w otrzymanej reprezentacji ofert.

W treści sprawozdania można przedstawić przykładowe wyniki grupowania i wyszukiwania najbliższych sąsiadów.

Dla wszystkich eksperymentów należy podać czasy wykonania. Przy wyznaczaniu czasów nie należy brać pod uwagę czasu wyświetlania wyników. Cały eksperyment można powtórzyć parokrotnie i uśrednić wyniki (można podać błąd standardowy).

2.5.1 Grupowanie

Dla każdej cechy wyszukaj kategorie przypisane do ogłoszeń posiadających daną cechę oraz ich liczbę, a następnie dla każdej pary cech wyszukaj kategorie przypisane do ogłoszeń posiadających daną parę cech oraz ich liczbę. Dla każdej cechy oraz pary cech podaj trzy najczęściej występujące kategorie oraz ich prawdopodobieństwa. Sprawdź wyniki dla 30 najczęściej występujących cech, oraz dla 30 najczęściej występujących par cech. W raporcie można podać przykładowe rezultaty.

2.5.2 Wyszukiwanie najbliższych sąsiadów

Wyszukaj 10 najbliższych sąsiadów dla każdego ogłoszenia z pierwszych 1000 (dla chętnych 10 tyś. ogłoszeń). Szukanie odbywa się w całym zbiorze danych będącym wynikiem wstępnego przetwarzania danych, tzn. na zbiorze zawierającym 100 cech oraz tylko przykłady zawierające co najmniej jedną cechę. Dla każdego zbioru najbliższych sąsiadów sprawdź liczbę wystąpień kategorii, do których są przypisani, oraz sprawdź, czy zgadza się ona z kategorią ogłoszenia, dla którego zostali wyszukani ci najbliżsi sąsiedzi. W raporcie można podać przykładowe rezultaty

Jako miarę podobieństwa przy wyszukiwaniu najbliższych sąsiadów należy zastosować współczynnik Jaccarda. Mierzy on podobieństwo między dwoma zbiorami i jest zdefiniowany jako iloraz mocy części wspólnej zbiorów i mocy sumy tych zbiorów:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

gdzie A i B są zbiorami, które odpowiadają obiektom opisanym przez atrybuty binarne.

Dla poniższego przykładu zbiór $A = \{\text{bmw, seria, 1, nawigacja}\}$ odpowiada pierwszemu obiektowi, a zbiór $B = \{\text{volvo, nawigacja, opony}\}$ drugiemu. W postaci tabelarycznej możemy te zbiory zapisać następująco:

bmw	seria	1	nawigacja	volvo	opony	kabriolet
1	1	1	1	0	0	0
0	0	0	1	1	1	0

Dla tego przykładu współczynnik Jaccarda wynosi:

$$J(A,B) = \frac{1}{6}$$

(maks. 4 strony).

2.6 Podsumowanie

Na końcu jest zawsze miejsce na krótkie podsumowanie (maks. 1/2 strony).

Całość raportu nie może przekraczać 10 stron.