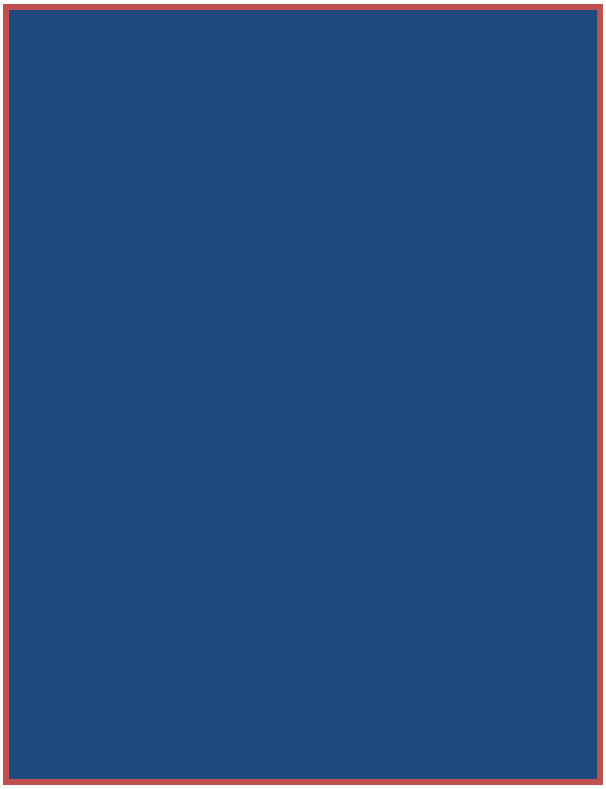

Komputerowa analiza danych

[Projekt]

Katedra Automatyki i Technik Informacyjnych

Autor: Szymon Łukasik

Wprowadzenie

Celem projektu jest zapoznanie Państwa z praktycznymi algorytmami i problemami współczesnej analizy danych oraz narzędziami służącymi do realizacji wybranych jej zadań.

Wybór tematów deklarujemy w systemie **ELF2** na odpowiedniej stronie Wiki od **24 marca 2014, godz. 22.00**. Część z tematów może być realizowana przez kilka grup jednocześnie – zostało to zaznaczone poniżej.

Tematy opisowo-programistyczne

W tej grupie tematów znalazły się projekty dotyczące zapoznania się z danym narzędziem analizy danych oraz demonstracji jego instalacji i zastosowania w formie 10-20 stronicowej instrukcji. W każdym przypadku należy pokazać – poza krótką informacją o instalacji i sposobie użytkowania narzędzia – import danych, klasteryzacja wybranym algorytmem, klasyfikacja wybranym algorytmem (z wcześniejszym podziałem zbioru na uczący i testujący). Plus za ciekawe przykłady, wizualizacje itp.

Temat 1 – zapoznaj się z biblioteką Apache Mahout

Temat 2 – zapoznaj się z biblioteką SciPy

Temat 3 – zapoznaj się z biblioteką scikit-learn

Temat 4 – zapoznaj się z biblioteką Orange

Temat 5 – zapoznaj się z pakietem R

Temat 6 – zapoznaj się z inną wybraną biblioteką programistyczną pozwalającą na realizację w/w zadań – temat do realizacji dla wielu grup, należy jednak wskazać narzędzie podlegające opracowaniu (może być ono przedmiotem rozważań **jednej grupy**)

Tematy programistyczne – wstępne przetwarzanie danych

Temat 1 – pobierz i przetwórz dane z bazy Instytucje Naukowe

Należy opracować parser pobierający i przetwarzający dane z serwisu OPI Instytucje Naukowe (http://www.nauka-polska.pl/dhtml/raportyWyszukiwanie/wyszukiwanieInstytucjeNaukowe.fs?lang=pl) . Parser powinien pobrać dane dla wybranego województwa i przetwarzać aktualnie istniejące jednostki wraz z ich wybranymi atrybutami. Proszę skoncentrować się

jedynie na cechach przydatnych do analizy np. dyscypliny, kierunki działalności, skład osobowy + np. pobranie podobnych atrybutów od poszczególnych członków zespołów. Koncentrujemy się na jednostkach podstawowych najniższego poziomu (Katedry, Instytuty) lub Średniego (Wydziały) pomijając najwyższy (Uczelnie).

W sprawozdaniu należy zamieścić kompletny opis opracowanego podejścia, przykładowe wyniki, komentarze dotyczące dalszego rozwoju zaproponowanego algorytmu. Należy załączyć także wszystkie pliki źródłowe. Oceniane będą: czytelność opisu, kreatywność zastosowanego podejścia, jakość przedstawienia wyników oraz ilość przetwarzanych atrybutów.

Temat 2 – pobierz i przetwórz dane z wybranego polskiego portalu newsowego – temat do realizacji dla wielu grup, należy jednak wskazać portal podlegający opracowaniu (może być ono przedmiotem rozważań **jednej grupy**)

Należy opracować parser pobierający i przetwarzający dane z wybranego portalu z newsami i dokonać analizy występujących słów w zbiorze dokumentów dotyczących danego zagadnienia/tematu.

Przykład 1: pobrać po 200 dokumentów w których w tytule pojawia się nazwisko 2 znanych polityków. Porównać występowanie słów próbując wysnuć na tej podstawie wniosek dotyczący nastawienia autorów do postaci występującej w artykule.

Przykład 2: pobrać godziny publikacji artykułów dla jednego miesiąca pracy i zidentyfikować ich rozkład

Tematy programistyczne – analiza danych

Temat – rozwiąż zadanie w konkursie FedCSIS - temat do realizacji dla wielu grup

Należy sformułować zespół, wziąć udział w zadaniu konkursowym organizowanym w ramach konferencji FedCSIS (https://fedcsis.org/2014/dm_competition) oraz uzyskać w nim jak najlepszy wynik. Problem dotyczy prognozowania obrażeń i ofiar podczas zdarzeń w których interweniowała Państwowa Straż Pożarna.

W sprawozdaniu należy zamieścić kompletny opis opracowanego podejścia, przykładowe wyniki, komentarze dotyczące dalszego rozwoju zaproponowanego algorytmu, a także uzyskany wynik z systemu konkursowego. Należy załączyć także wszystkie pliki źródłowe. Oceniane będą: czytelność opisu, kreatywność zastosowanego podejścia, jakość przedstawienia wyników.

Grupa tematów – rozwiąż zadanie w Kaggle

Należy zgłosić zespół w wybranym zadaniu konkursowym ze strony http://www.kaggle.com/ oraz uzyskać w nim jak najlepszy wynik. Wskazane jest zadanie aktualnie ogłoszone oraz nie występujące w sekcji Getting Started Competitions. **Niedozwolona jest realizacja zadania konkursowego podjętego przez inną grupę.** Wybór zadania deklarujecie Państwo w ELF2.

W sprawozdaniu należy zamieścić kompletny opis opracowanego podejścia, przykładowe wyniki, komentarze dotyczące dalszego rozwoju zaproponowanego algorytmu, a także uzyskany wynik z systemu konkursowego. Należy załączyć także wszystkie pliki źródłowe. Oceniane będą: czytelność opisu, kreatywność zastosowanego podejścia, jakość przedstawienia wyników.

Temat własny

Prowadzący jest otwarty na propozycje tematów własnych. Można je zgłaszać mejlowo do akceptacji, wraz z listą osób zaangażowanych do **31 marca 2014.** Każdy z projektów musi dotyczyć albo przetwarzania interesujących danych z serwisu internetowego i ich zapisania do postaci dogodnej do dalszej analizy (z demonstracją przykładowego procesu wnioskowania) lub/oraz bądź to analizy skupień bądź klasyfikacji danych które już istnieją w postaci ustrukturalizowanej.

Informacje praktyczne

Sprawozdanie wraz z ewentualnymi załącznikami należy złożyć w formie elektronicznej w systemie **elf2.pk.edu.pl** (odpowiednie instrukcje zawarto w sekcji projektu tamże)**.** Wgranie sprawozdania do systemu e-learningowego jest jednorazowe i ostateczne. Sprawozdanie nie powinno zawierać nadmiaru informacji teoretycznych – a jedynie najważniejsze aspekty praktyczne rozważanego zagadnienia. Szczególnie doceniane będą umiejętność krytycznej analizy i formułowania wartościowych wniosków o charakterze technicznym.

Zespoły mogą być max. 3 osobowe Proszę pamiętać o odpowiednim nazwaniu pliku np. *Nazwisko1_Nazwisko2_Nazwisko3.zip* oraz wymienić w sprawozdaniu wszystkich członków zespołu, max. rozmiar pliku: 20MB)

Końcowym terminem oddania sprawozdania jest **30 czerwca 2014**. Po tym terminie ocena ostateczna będzie uwzględniała spóźnienie w realizacji zadania. Daty wpisów zostaną wcześniej ogłoszone w systemie **elf2.pk.edu.pl**. Konsultacje poza spotkaniami projektowymi są możliwe również *online* via e-mail lub poprzez system **elf2.pk.edu.pl**.