Opis wymagań projektu

Celem projektu jest opracowanie, implementacja i opisanie systemu Big Data do przetwarzania danych. Wykonujący projekt powinien postawić cel biznesowy z użyciem systemu analizy danych, przeprowadzić analizę postawionej hipotezy biznesowej i opisanie jej wyniku.

Zestaw danych to otwarte pliki *Stack Exchange Data Dump* dostępne pod adresem https://archive.org/details/stackexchange. Pliki spakowane są programem 7-zip (http://www.7-zip.org/download.html). Opis zawartości plików znajduje się w https://archive.org/download/stackexchange/readme.txt. Nie ma wymogu co do wyboru zestawu danych, niemniej, przyczyny wyboru powinny zostać opisane. Niezależnie od wielkości wybranego zbioru danych, analiza powinna być wykonana za pomocą narzędzi do przetwarzania Big Data.

Zaprojektowany system może skupiać się zarówno na problemie analitycznym jak i technicznym przetwarzania danych. Należy pamiętać, że opis części rozwiązania w takim przypadku musi się skupiać bardziej na analitycznym bądź technicznym aspekcie rozwiązania.

Proces Data Engineering/Science można podzielić w ogólności na kroki:

1. Zbieranie surowych danych
2. Przetwarzanie danych
3. Eksploracja danych
4. Czyszczenie danych
5. Modelowanie (opcjonalne)
6. Produkt oparty na danych (czyste dane, analiza, wyniki itp.)
7. Komunikacja wyników

Ciekawe rozwinięcie powyższego opisu można przeczytać na stronie http://www.kdnuggets.com/2016/03/data-science-process.html.

Nie ma wymagań odnośnie technologii użytych w wykonaniu projektu, jednak muszą być to technologie Big Data. Niemniej, wybór technologii powinien być umotywowany. Oczywiście końcowy etap analizy danych zagregowanych czy wizualizacji, może być wykonany przy pomocą narzędzi do małych danych (np Python Pandas), ale znacząca część przetwarzania danych powinna być wykonana w technologii Big Data. Nie ma obowiązku ograniczania się do technologii poznanych na zajęciach, niemniej należy pamiętać, że prowadzący zajęcia mogą mieć ograniczoną wiedzę na temat technologii nie przedstawionych na zajęciach.

O ile wizualizacja wyników jest wskazana ze względu na łatwość ugruntowania tezy przy pomocy odpowiednich wykresów, nie jest to krok wymagany. Zatem jeżeli opis tekstowy jest wystarczający, nie ma obowiązku zamieszczania dodatkowych ilustracji i ich brak nie wpłynie na ocenę.

W trakcie wykonywania projektu zachęcamy do używania szerokiego wachlarza poznanych technik, niemniej, nie ma obowiązku wykorzystania jakiś konkretnych rozwiązań czy algorytmów analitycznych, jak modelowanie statystyczne czy uczenie maszynowe. Niemniej, analiza przedstawiona w końcowym opisie projektu musi być adekwatna do postawionej tezy. Przykładowo, o ile projekt mający na celu budowę procesu przetwarzania danych i dostarczenia ich do klienta końcowego może mieć proste techniki opisu ilościowego, to hipoteza zakładająca predykcyjny model zjawiska, powinna tenże model i jego sprawdzenie zawierać.

Projekt musi być złożony w formie pisemnej. Nie ma obowiązku używania konkretnego edytora tekstu czy pakietu biurowego, pod warunkiem, że końcowe dzieło będzie czytelne i będzie miał formę pracy opisowej.

Temat projektu powinien być omówiony z prowadzącym w trakcie pierwszych zajęć z projektu. W razie nieobecności na zajęciach można alternatywnie przedstawić zamysł prowadzącemu w innej formie, np. poprzez e-mail. Niemniej, pamiętać trzeba o potrzebie otrzymania informacji zwrotnej na temat zakresu projektu, w razie jakby informacja nie trafiła do prowadzącego. Konsultacje tą należy przeprowadzić przed kolejnymi zajęciami z projektu.