Eksploracja danych w Internecie

(Kod przedmiotu: 02 64 5054 00)

Laboratorium 1

Zadanie:

Zostałeś zatrudniony przez firmę Trends Sp. z o. o. zajmującą się prognozowaniem trendów w różnych dziedzinach życia i biznesu. Firma posiada modele zdolne z wysoką dokładnością przewidzieć zbliżające się trendy na rynku, jednak system dostarczający dane z Internetu, którym obecnie dysponują przysyła zaszumione dane (zawierające znaczniki HTML a każdy link trzeba podawać osobno). W związku z tym Twoim pierwszym zadaniem jest napisanie systemu, który w przeciwieństwie do istniejącego będzie zawierał poniższe funkcjonalności:

1. Dostarczy dane pozbawione znaczników HTML.
2. Automatycznie "odwiedzi" (przetworzy w analogiczny sposób) wszystkie linki z podanej strony.
3. Odnajdzie adresy email.

Dodatkowe wymagania do aplikacji:

1. Język programowania: Java lub Python.
2. Przejście przez strony, tak aby nie przepełnić stosu (bez rekurencji).
3. Wejście programu: link do strony internetowej.
4. Wynik pracy programu: Szereg par plików csv. Pierwszy plik z pary zawiera treść stony bez tagów HTML. Drugi plik zawiera listę adresów email.
5. Należy zaprojektować wyrażenie regularne dzięki, któremu znajdziemy wszystkie adresy email.

Ocena zadania uwzględnia (czas realizacji: 1 tydzień. Po tygodniu ocena jest obniżana):

- poprawność mechanizmów weryfikacji wejścia (czy podano link i czy link jest właściwy),
- poprawność mechanizmów akwizycji danych z Internetu (usuwanie tagów, ekstrakcja adresów email),
- poprawność wyjścia (liczba pozyskanych adresów email musi zgadzać się z liczbą na stronie, treść musi zgadzać się z treścią na stronie),
- obsługę błędów i wyjątków (np. Jeżeli adres nie istnieje, albo przekroczono czas odpowiedzi na żądanie),

- Jakość napisanego kodu (formatowanie, nazwy zmiennych, …).