МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ В.Н.КАРАЗІНА
ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК

Лабораторна робота №3

з дисципліни «крос-платформне програмування»

на тему «JAVA & XML»

Виконав:
студент групи КС-23
Терещенко Є.Ю.

Перевірив:
Канд. фіз.-мат. наук
Споров О. Є.

ХАРКІВ – 2023

Завдання №0

Із сайту з відкритими даними (https://catalog.data.gov/dataset/popular-baby-names) було отримано свіжий (від 3 березня, 2023), великий за розміром датасет в XML форматі з інформацією про популярні імена дітей у місті Нью-Йорк. Цей датасет складений за офіційною інформацією із служби реєстрації актів цивільного стану міста Нью-Йорка. Архів з цим датасетом має назву Popular_Baby_Names_NY.zip та розміщений в лекційному Гугл-класі в розділі Методичні вказівки з виконання лабораторних робіт. Кожен запис цього датасету представляє інформацію про дитину: вказано дату народження, гендер, етнічну приналежність мами, власне ім'я дитини, кількість (count) дітей з цим іменем та рейтинг (rating) імені у відповідній групі. Потрібно провести попередній аналіз цих даних та вибрати з них лише потрібну для подальшої роботи інформацію. Виконати наступні завдання:

1. Написати програму для виведення на екран частини XML документу за допомогою SAX парсеру без валідації для вивчення його структури та вмісту; програмно отримати перелік всіх тегів, імена яких присутні в документі.

2. За невеликим характерним фрагментом скласти xsd схему документу, створити валідатор та перевірити, чи правильно було зрозуміло структуру документу.

3. Написати програмне рішення, що за допомогою SAX парсеру без валідації отримає назви всіх національних груп, що представлені в документі.

4. Написати додаток, що з всього XML документу вибирає задану кількість найбільш популярних імен в заданій етнічній групі із зберіганням інформації про: ім'я, гендер, кількість імен та рейтинг імен, а також створює відповідні Java об'єкти для зберігання цієї інформації та сортує інформацію по збільшенню номеру в рейтингу. Зберегти — 42 — вибрану та відсортовану інформацію до нового XML файлу за допомогою DOM парсеру.

5. Прочитати цей новий документ за допомогою DOM парсеру та вивести інформацію, що в ньому зберігається, на екран.

Успішно виконано всі завдання:

1. Вивів на екран частину XML документу за допомогою SAX парсеру без валідації для вивчення структури та вмісту.
2. Отримав перелік всіх тегів, присутніх у документі.
3. Склав xsd схему документу на основі характерного фрагменту та перевірив її за допомогою валідатора.
4. Використовуючи SAX парсер без валідації, отримав назви всіх національних груп у документі.
5. Написав додаток, що вибирає задану кількість найпопулярніших імен в

заданій етнічній групі, зберігаючи інформацію про ім'я, гендер, кількість та рейтинг імен. Сортував цю інформацію за збільшенням рейтингу.
6. Зберіг вибрану та відсортовану інформацію до нового XML файлу за допомогою DOM парсеру.
7. Знову прочитав цей новий документ за допомогою DOM парсеру та вивів збережену в ньому інформацію на екран.

1. DataHandler.java - Цей клас відповідає за обробку XML даних, зокрема за виведення частини XML документу та отримання переліку тегі
2. EthnicityHandler.java - Цей клас використовує SAX парсер без валідації для отримання назв національних груп, що представлені в документі.
3. Main.java - Основний клас програми, який містить метод main() і відповідає за запуск програми і координацію роботи інших класів.
4. XMLValidator.java - Цей клас використовується для валідації XML документу. Він перевіряє правильність структури документу на основі xsd схеми, що була складена.
5. NameRankingParser.java - Цей клас використовує SAX парсер без валідації для обробки XML документу і вибору заданої кількості найпопулярніших імен в заданій етнічній групі. Він також зберігає інформацію про ім'я, гендер, кількість та рейтинг імен і сортує цю інформацію за збільшенням рейтингу.

```
response [] row [] row [_id=row-v4f5~xz3v-vr86,_uuid=00000
000-0000-0000-E0ED-52E8592E0A7A,_position=0,_address=https
://data.cityofnewyork.us/resource/_25th-nujf/row-v4f5~xz3v
-vr86,] brth_yr []  = 2011
gndr []  = FEMALE
ethcty []  = HISPANIC
nm []  = GERALDINE
cnt []  = 13
rnk []  = 75

row [_id=row-gdep~mr7x-dj3u,_uuid=00000000-0000-0000-68AD-
7B741D1DF31B,_position=0,_address=https://data.cityofnewyo
rk.us/resource/_25th-nujf/row-gdep~mr7x-dj3u,] brth_yr []
 = 2011
gndr []  = FEMALE
ethcty []  = HISPANIC
nm []  = GIA
cnt []  = 21
rnk []  = 67

row [_id=row-4f22~xggt-ah4b,_uuid=00000000-0000-0000-B276-
36740F71A706,_position=0,_address=https://data.cityofnewyo
rk.us/resource/_25th-nujf/row-4f22~xggt-ah4b,] brth_yr []
 = 2011
gndr []  = FEMALE
ethcty []  = HISPANIC
nm []  = GIANNA
cnt []  = 49
rnk []  = 42
```

Малюнок 1 – Отримання тегів, атрибутів та значень за допомогою SAX парсеру

[] - атрибути

```
--- --- ---
Congrats! The XML file is valid!
--- --- ---
```

Малюнок 2 – Лог про валідність документу

```
--- --- ---
Ethnicity: ASIAN AND PACI, Count: 1767
Ethnicity: HISPANIC, Count: 14530
Ethnicity: BLACK NON HISP, Count: 1740
Ethnicity: ASIAN AND PACIFIC ISLANDER, Count: 6799
Ethnicity: BLACK NON HISPANIC, Count: 7213
Ethnicity: WHITE NON HISP, Count: 3441
Ethnicity: WHITE NON HISPANIC, Count: 14019
```

Малюнок 3 – Результат парсингу всіх всіх національних груп

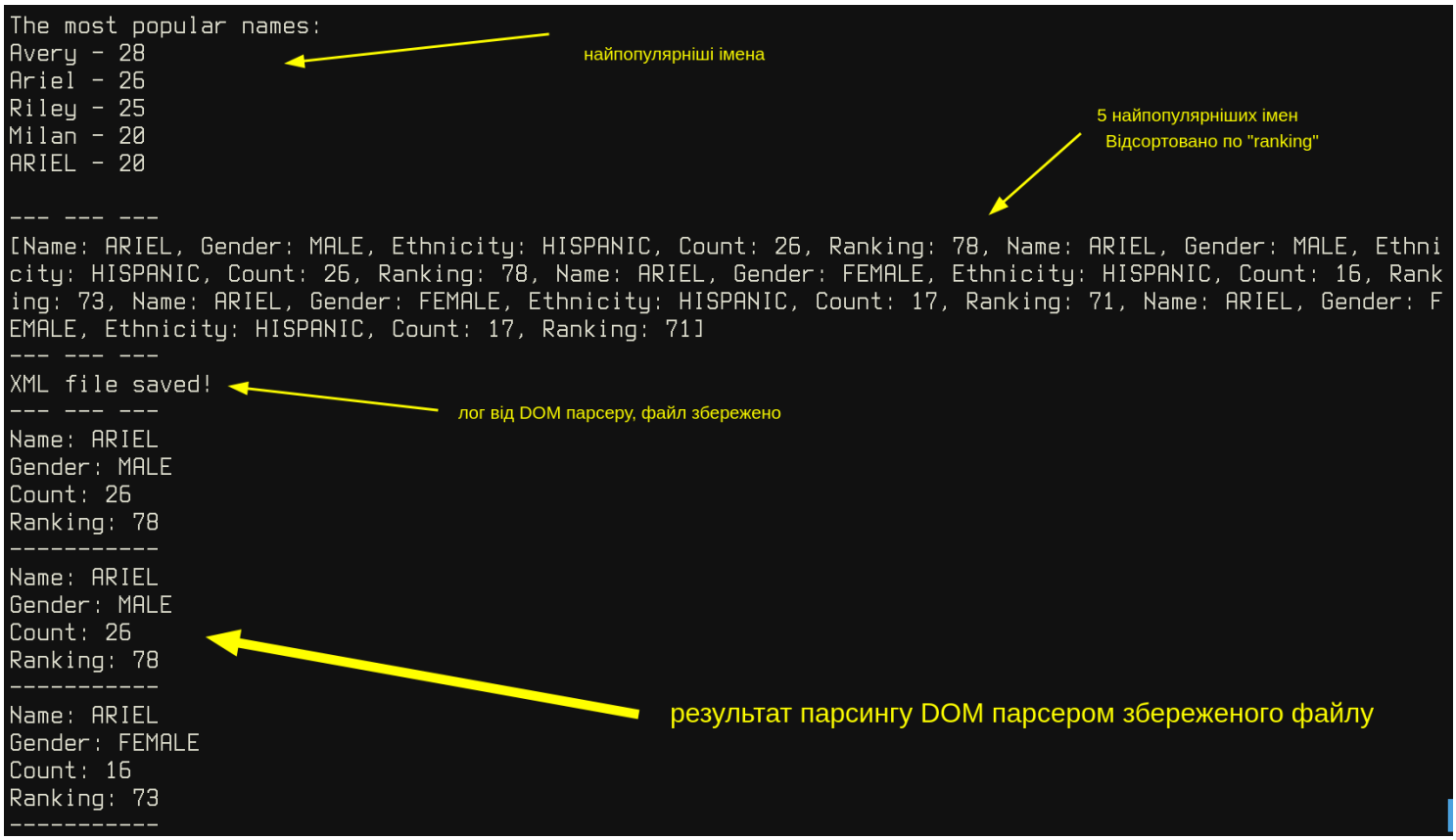

Рисунок 4 – Результат виконання додатка про найпопулярніші імена

Посилання на GidHub: https://github.com/zellii1/KPP.git