Оленчикова Т.Ю.

**Методические указания и задачи к выполнению лабораторной работы по дисциплине «**Программирование на языке Java для анализа данных**»**

**Лабораторная работа 7 ЛР 7 Распределенные вычисления с применением MapReduce**

Цель: Ознакомиться с фреймворком Apache Hadoop для распределенного хранения и обработки данных, освоить модель программирования MapReduce для обработки больших данных на кластерных вычислительных системах

**Теоретические материалы**

# 1) Методы и модели исследования сложных систем и обработки больших данных : монография / И. Ю. Парамонов, В. А. Смагин, Н. Е. Косых, А. Д. Хомоненко ;. — Санкт-Петербург : Лань, 2020. глава 6, С,203-211. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/126938 (дата обращения: 23.08.2023). – Режим доступа: для авториз. пользователей.

# 2) Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. раздел 2, С.20-30. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/182452 (дата обращения: 23.08.2023). — Режим доступа: для авториз. пользователей.

3) Учебник по Hadoop [Электронный ресурс]. – URL: <https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/uchitsia-hadoop/uchebnik-po-hadoop> Доступ свободный

Задание.

Необходимо решить задачу формирования списка рекомендованных товаров для пользователей интернет-магазина с применением алгоритма кросс-корреляции (имея множество кортежей объектов, для каждой возможной пары объектов посчитать число кортежей, в которых они встречаются вместе).

1. Реализовать два алгоритма на MapReduce:

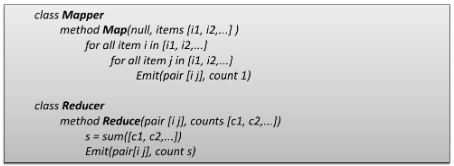
* Cross Correlation Pairs (листинг 1);
* Cross Correlation Striples (листинг 2).

2. Написать генератор базы данных интернет-заказов. Учесть, что заказ состоит из произвольного количества позиций (товаров).

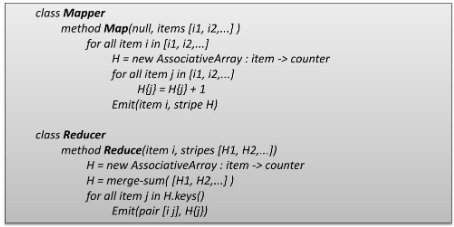
3. Обработать алгоритмом кросс-корреляции базу данных заказов (подсчитать количество вхождений каждой пары товаров).

4. Реализовать компонент советника, не применяя паттерн MapReduce. Вводится название товара. Выводятся 10 товаров, которые чаще всего покупают с заданным товаром. Чтение результатов работы алгоритма кросс-корреляции из HDFS реализовать вручную с помощью FileSystemIP.

5. Напишите ответ на контрольный вопрос в соответствии с вашим вариантом.



Листинг 1 – Псевдокод алгоритма Cross Correlation Pairs



Листинг 2 – Псевдокод алгоритма Cross Correlation Pairs

**Индивидуальные задания для лабораторной работы.**

*Лабораторная работа содержит только 1 вариант*

**Контрольные вопросы**

1. Что такое большие данные?
2. Объясните пять признаков больших данных
3. Сколько форматов ввода существует в Hadoop?
4. Что такое Thread? Кто использует Hadoop?
5. Каковы основные функции Hadoop?
6. В каких трех режимах может работать Hadoop?
7. В каких операционных системах работает Hadoop?
8. Каковы компоненты HDFS?
9. Что такое Hadoop Map Reduce?
10. Как работает Hadoop MapReduce?
11. Объясните, что такое спекулятивное исполнение?
12. Объясните, в чем разница между входным разделением и блоком HDFS?
13. Укажите, в чем разница между СУБД и Hadoop?
14. Укажите, какие компоненты данных использует Hadoop?
15. Укажите, какие наиболее распространенные форматы ввода определены в Hadoop?
16. Перечислите три файла конфигурации Hadoop.

**Критерии оценивания.**

За решение задачи вы можете получить до 2 баллов: задание выполнено полностью и правильно – 2 балла; имеются незначительные ошибки – 1 балл; приложение не работоспособно – 0 баллов;

Ответ на контрольный вопрос должен продемонстрировать понимание архитектуры Hadoop, за ответ вы можете получить оценку: ответ полный и правильный – 1 балл; ответ не удовлетворительный – 0 баллов.

Итого, максимальная оценка – 3 балла

**Внимание!** Полученная оценка автоматически снижается на 2% за каждую полную неделю задержки сдачи отчета по работе, но не более, чем на 40%

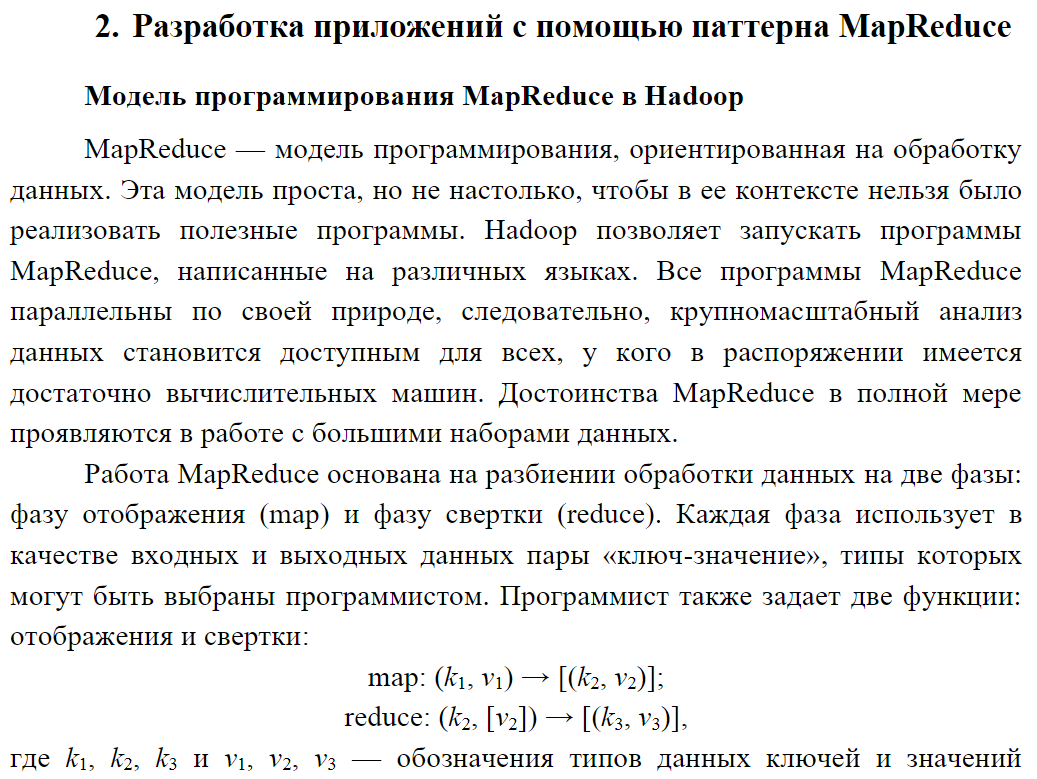
**Библиографический список**

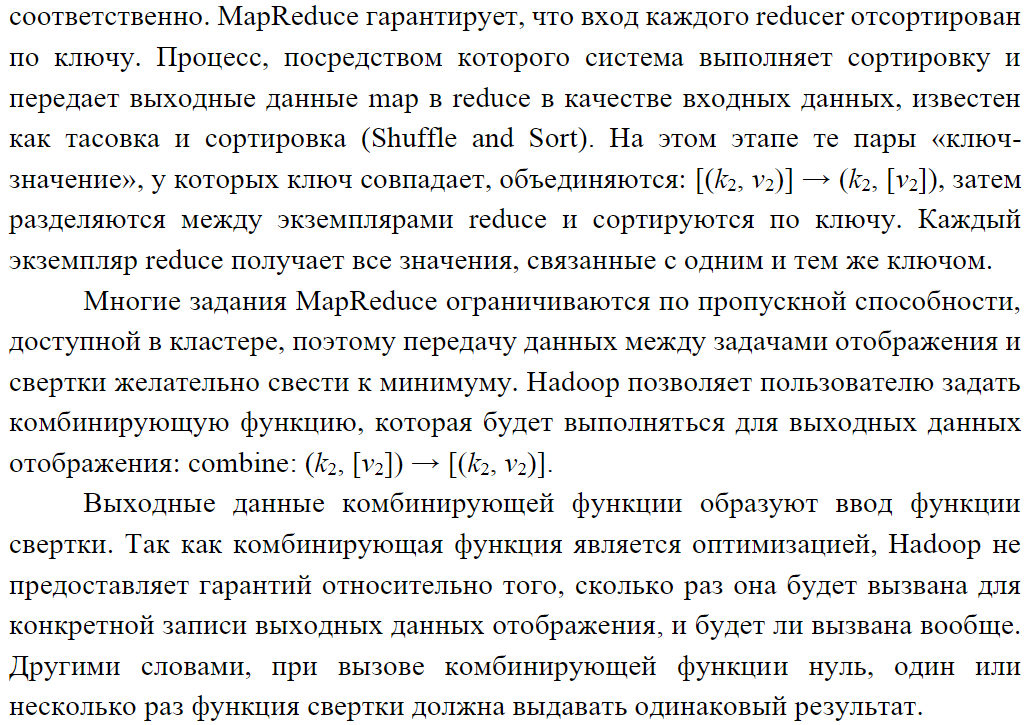
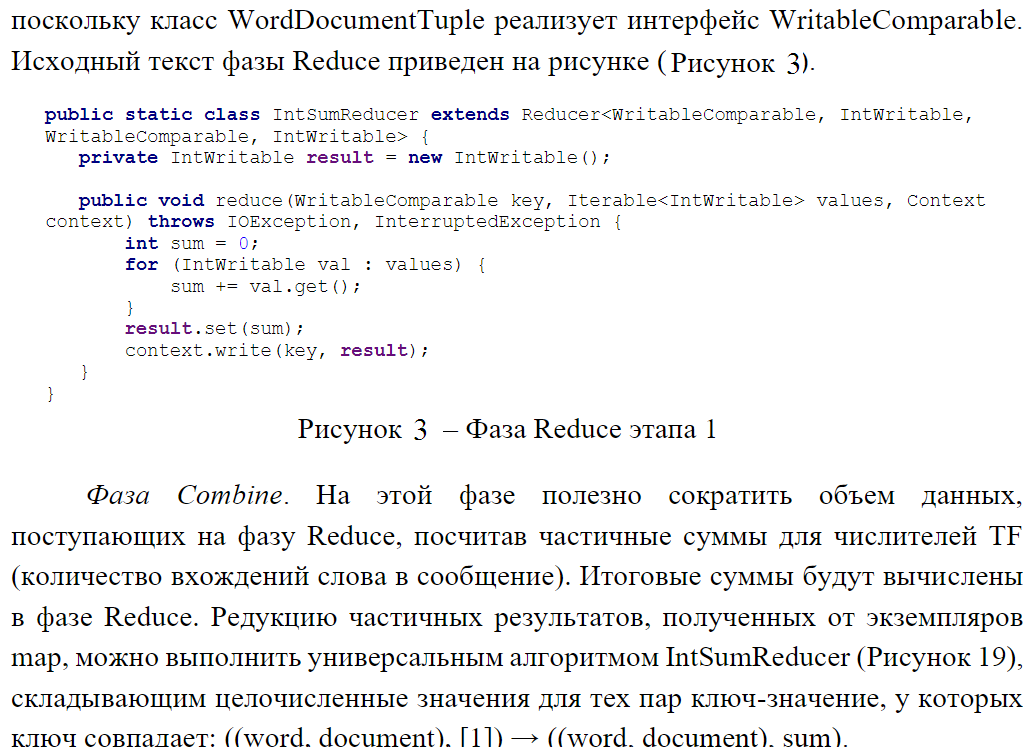
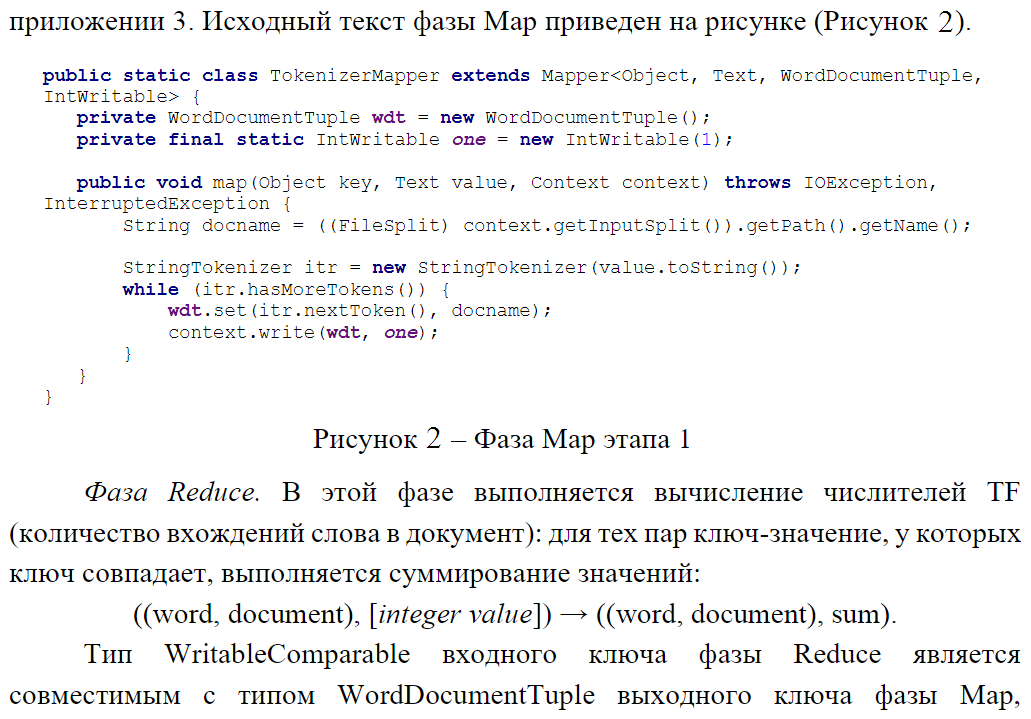
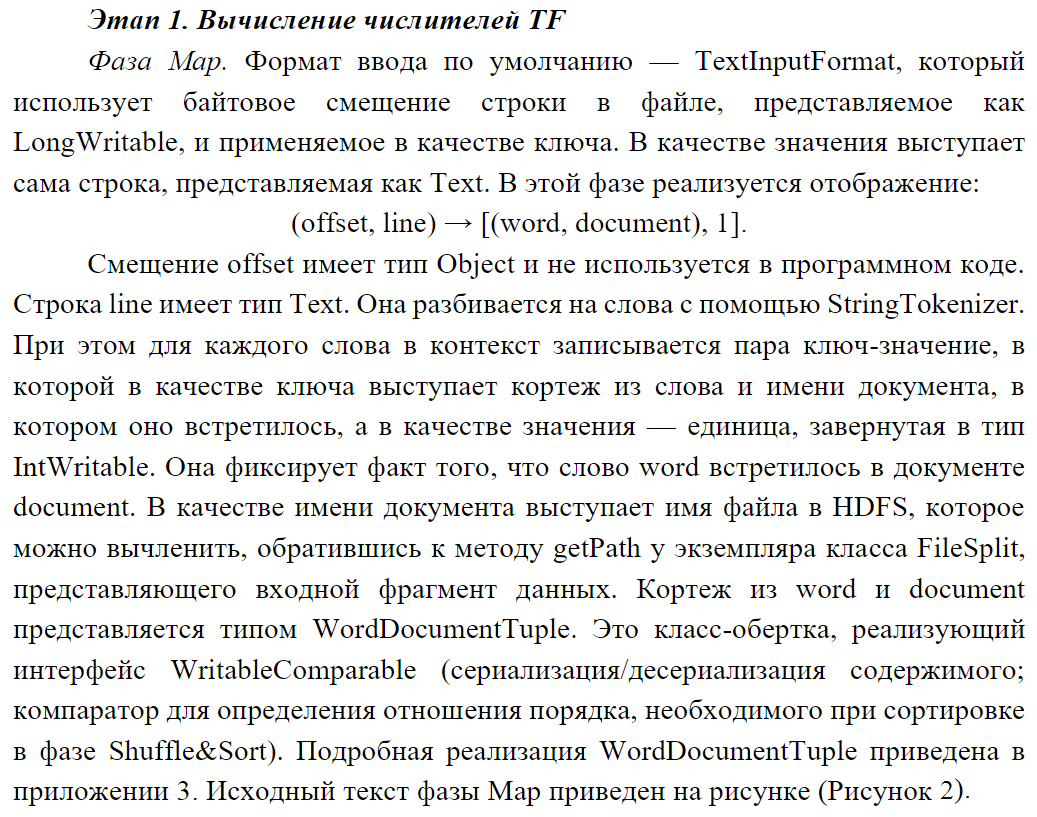
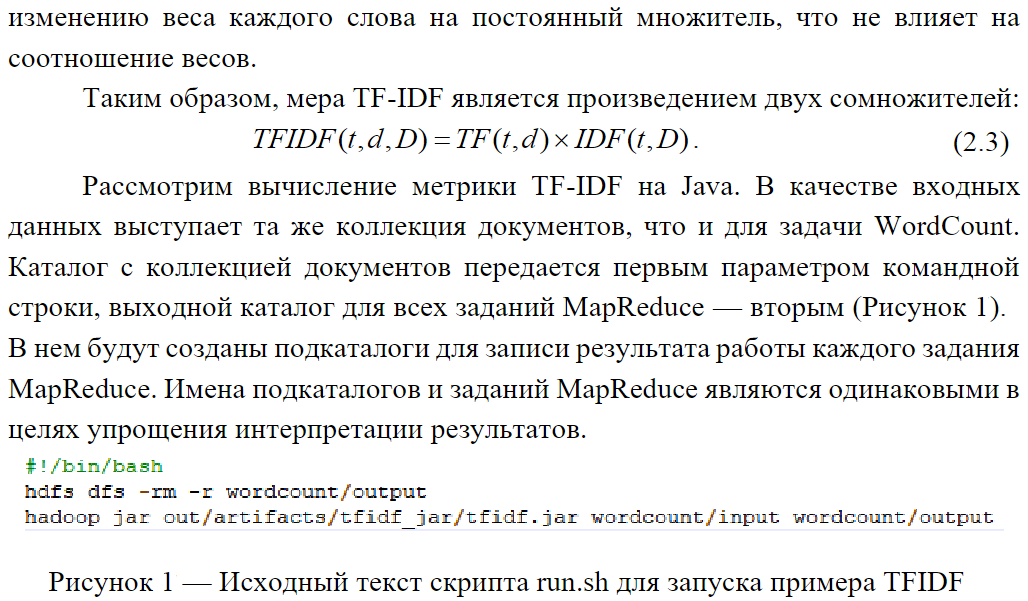
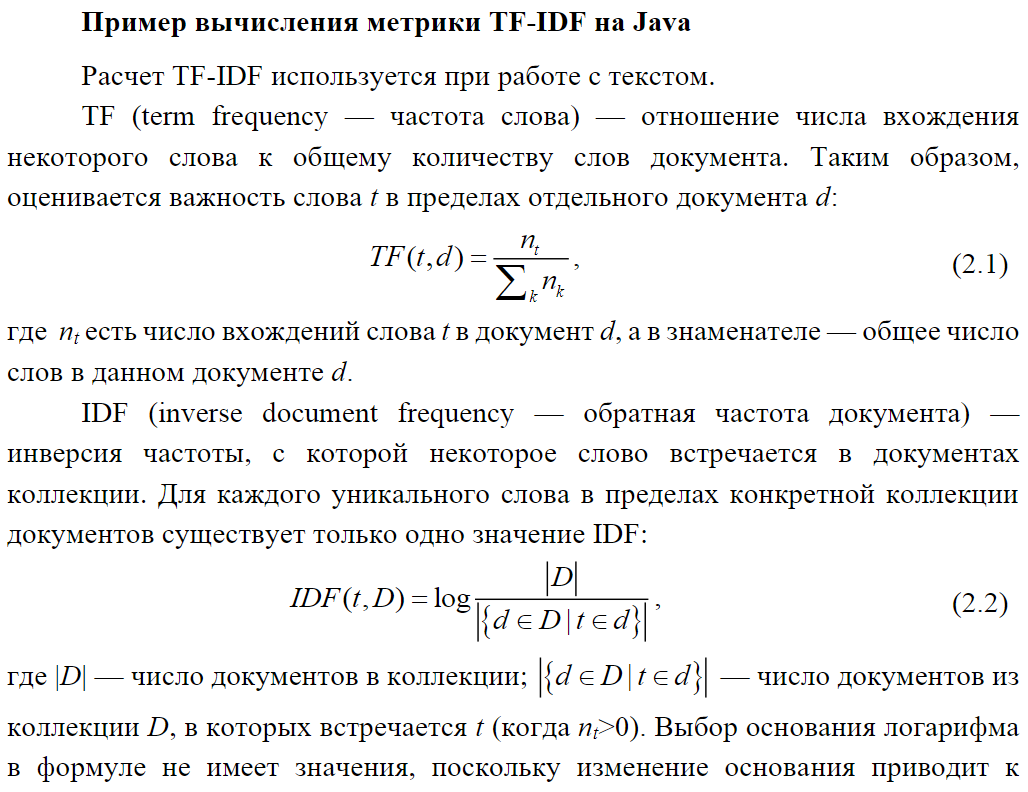
1. Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. – Москва : РТУ МИРЭА, 2021. – 91 с. // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/182452 (дата обращения: 24.08.2023). – Режим доступа: для авториз. пользователей.

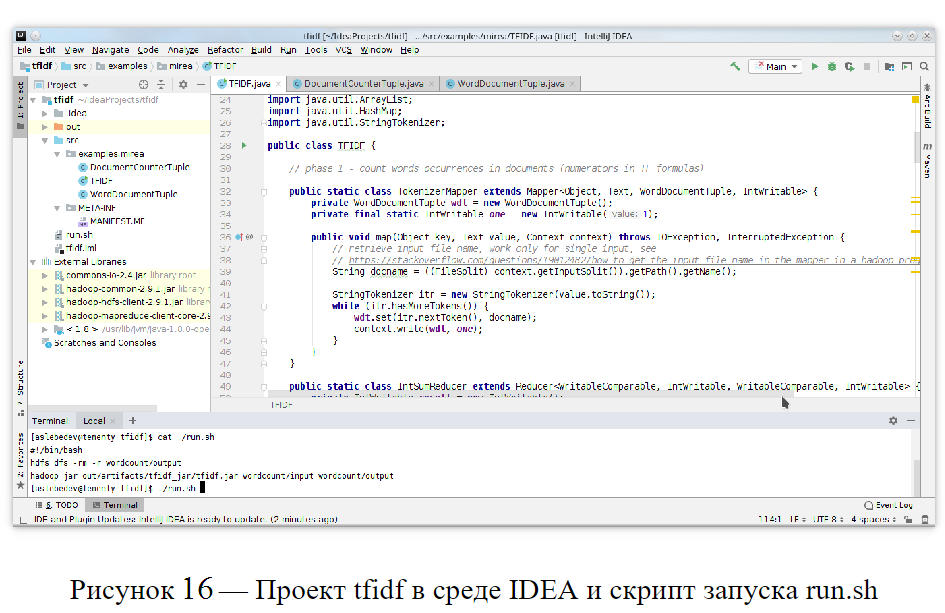
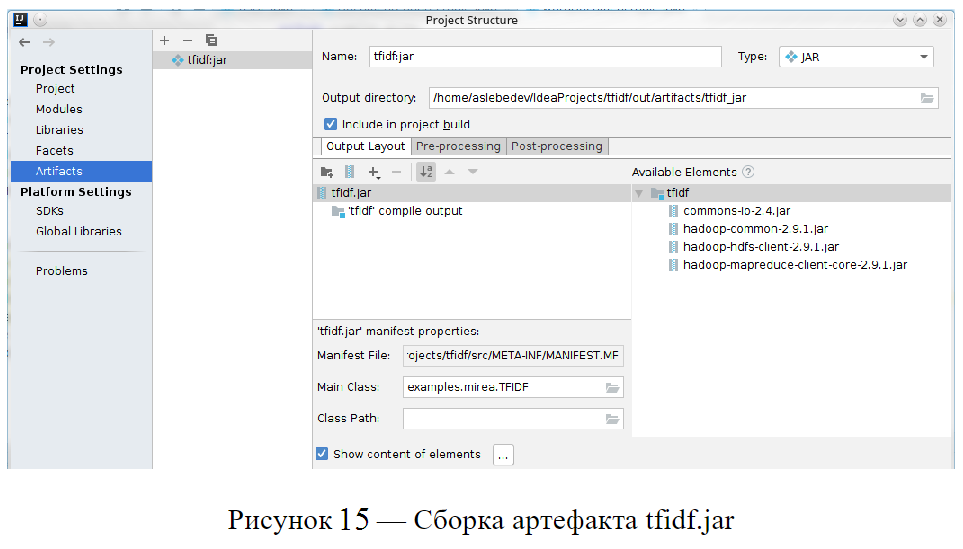
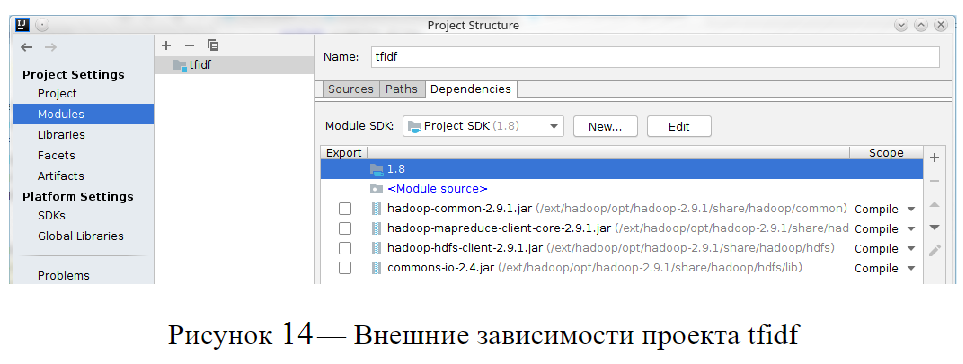
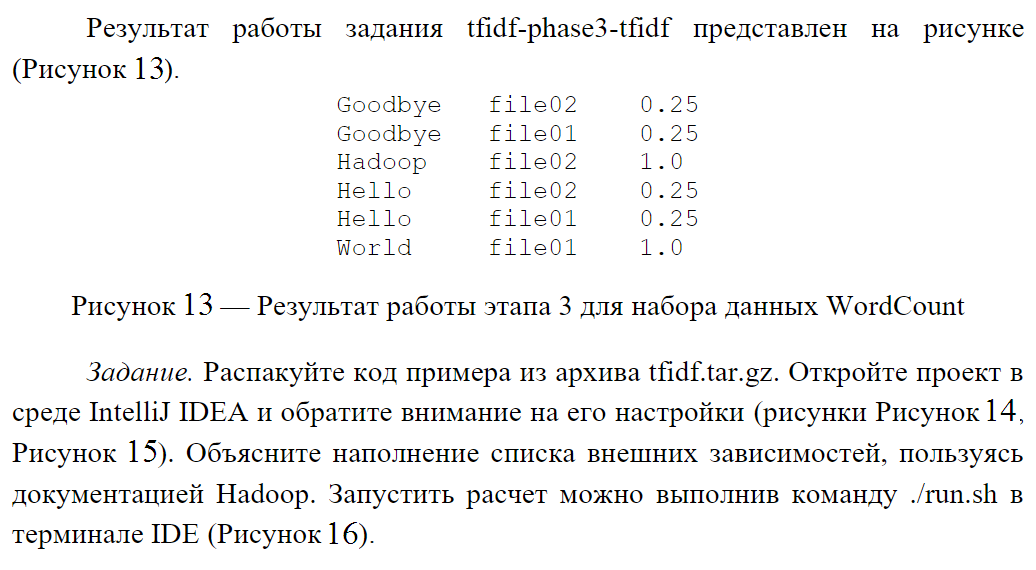
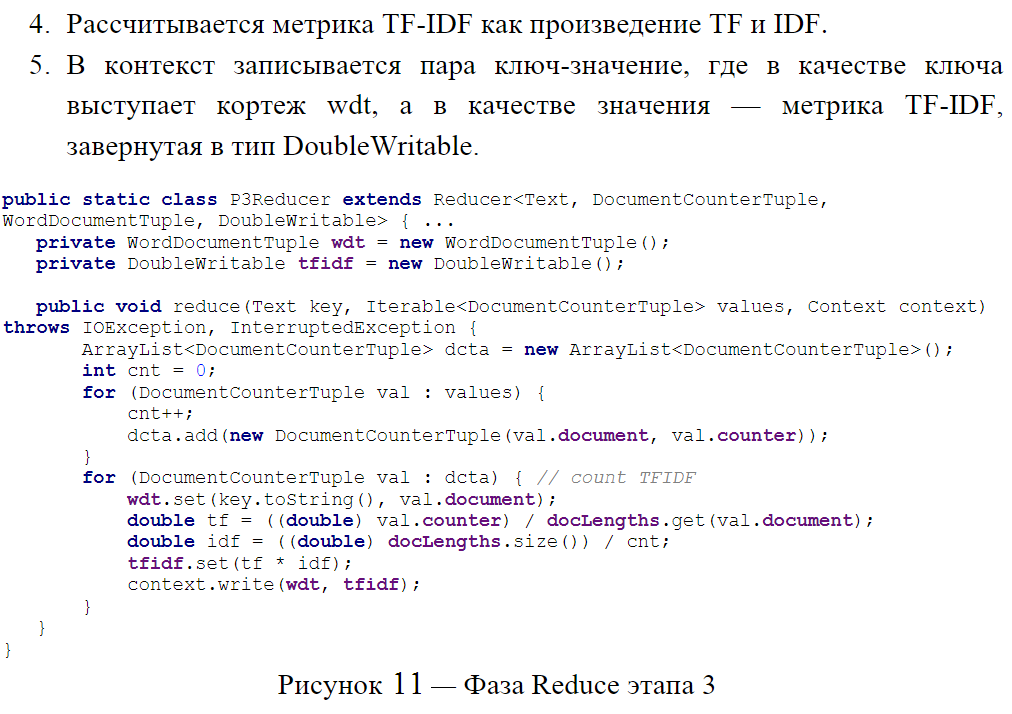
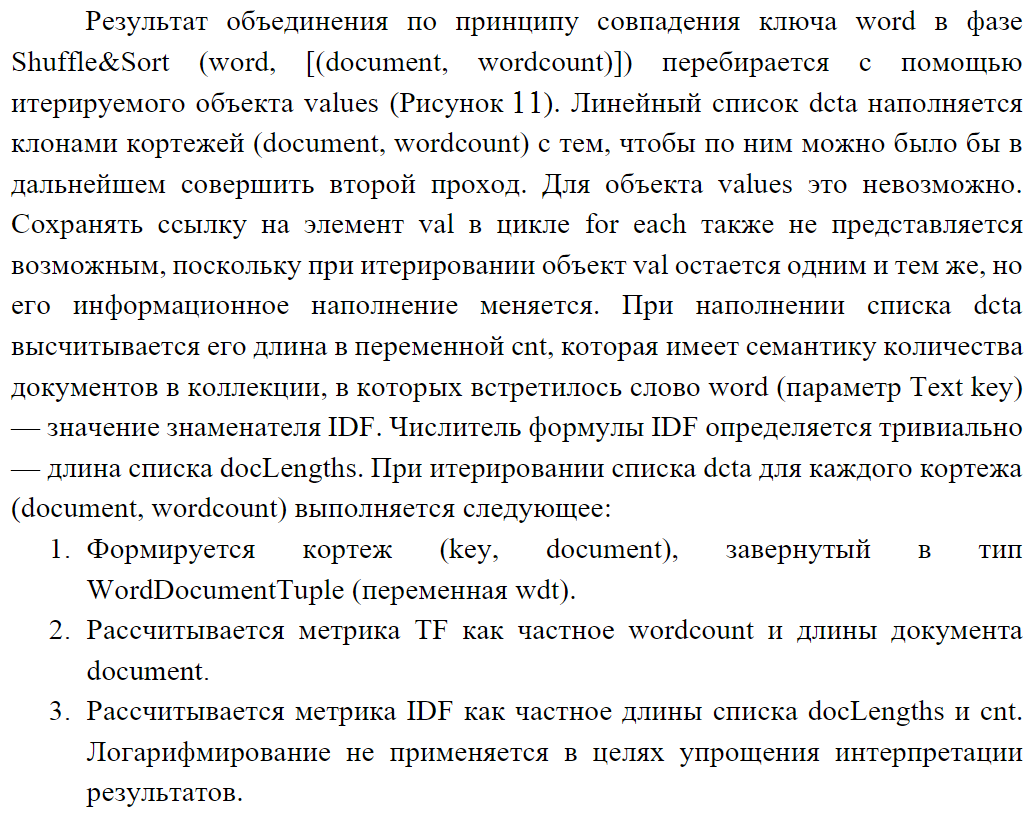
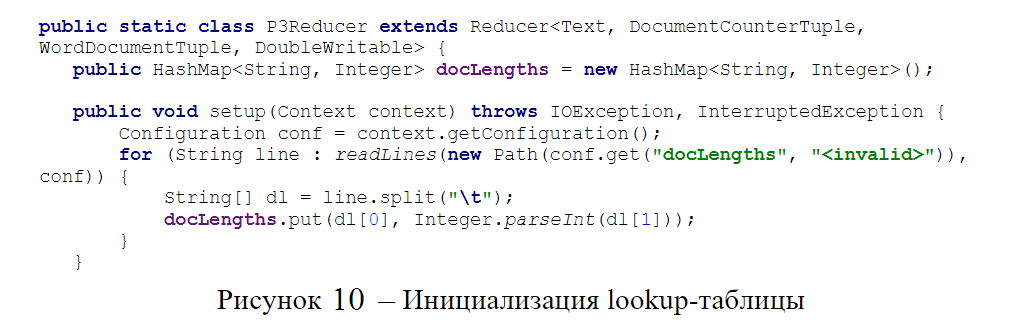
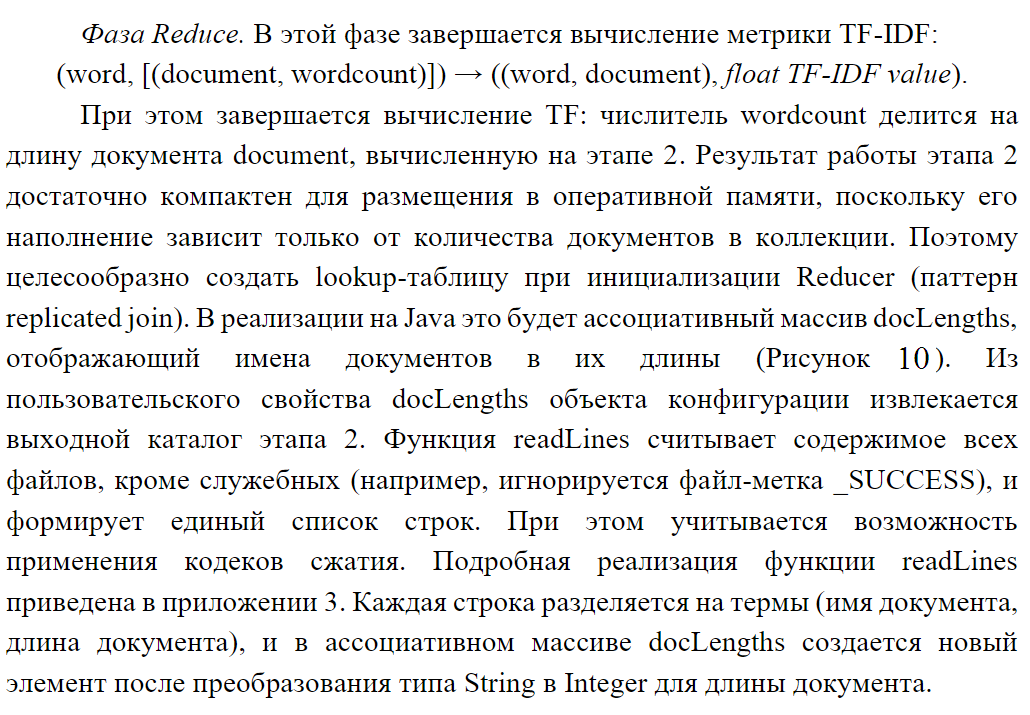
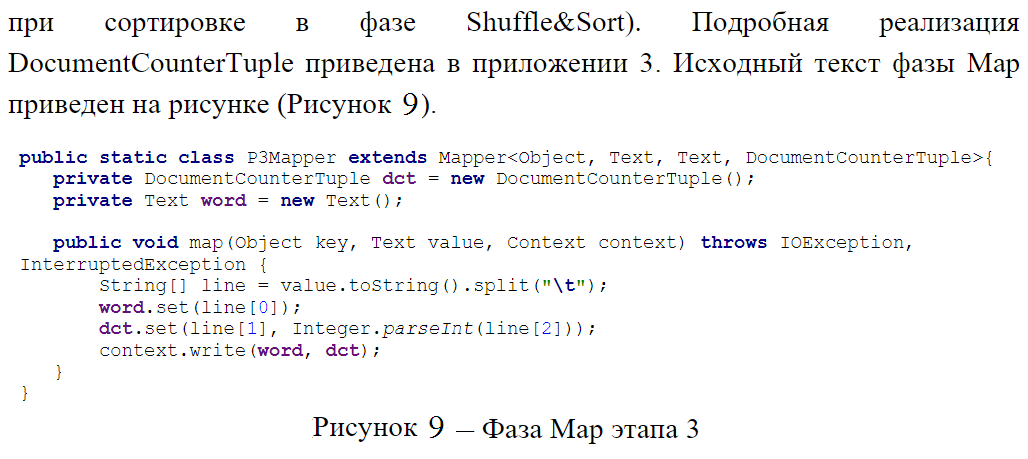
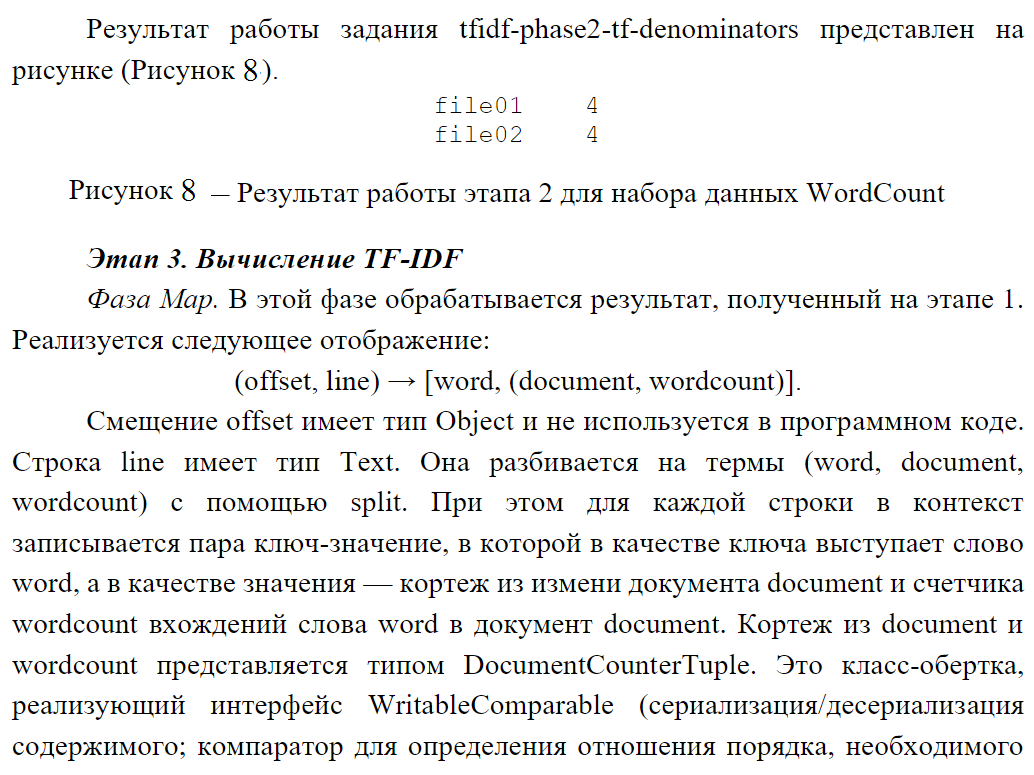
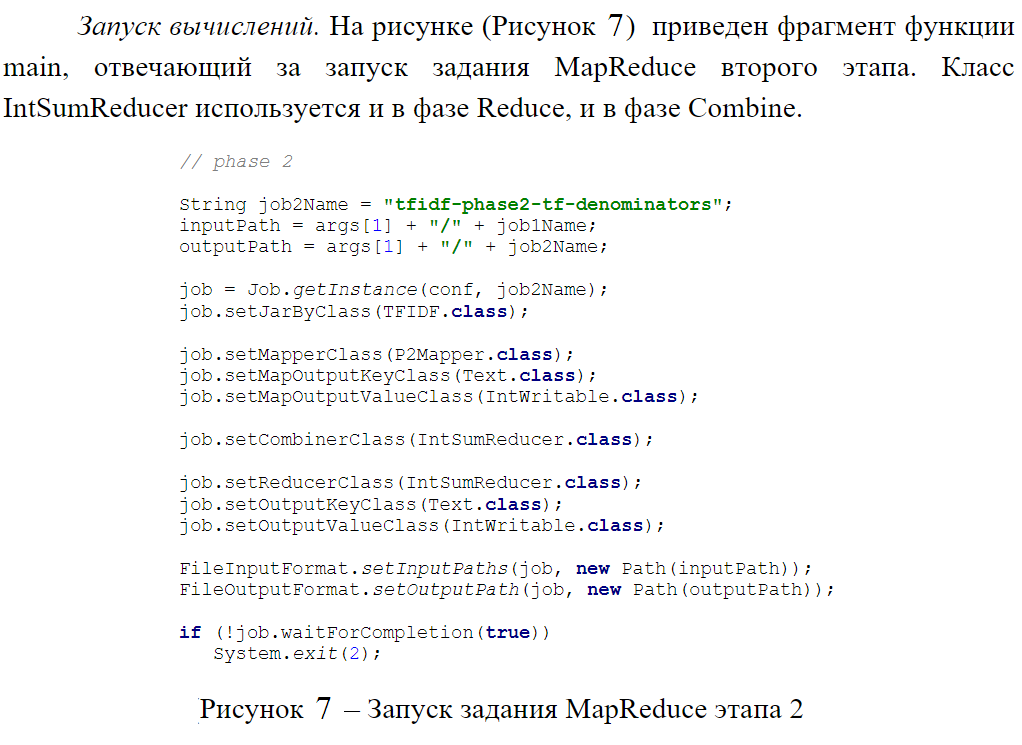
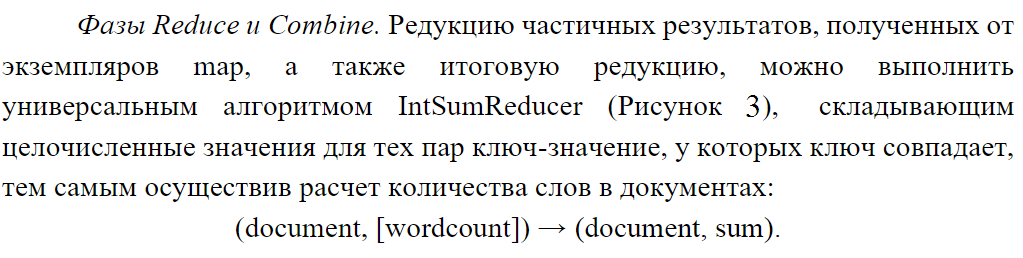
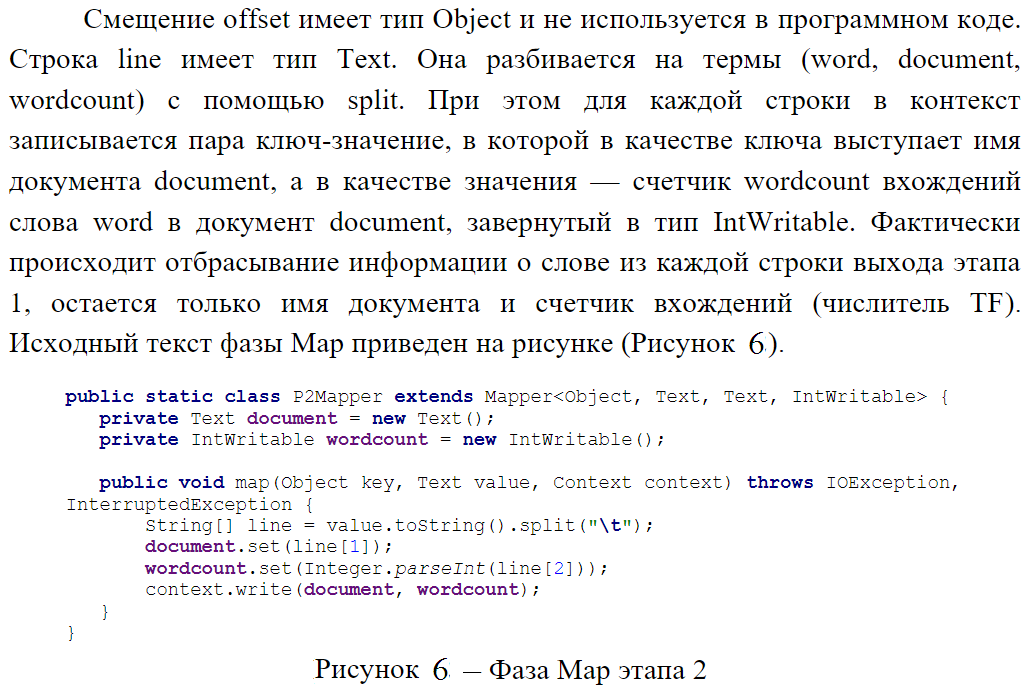
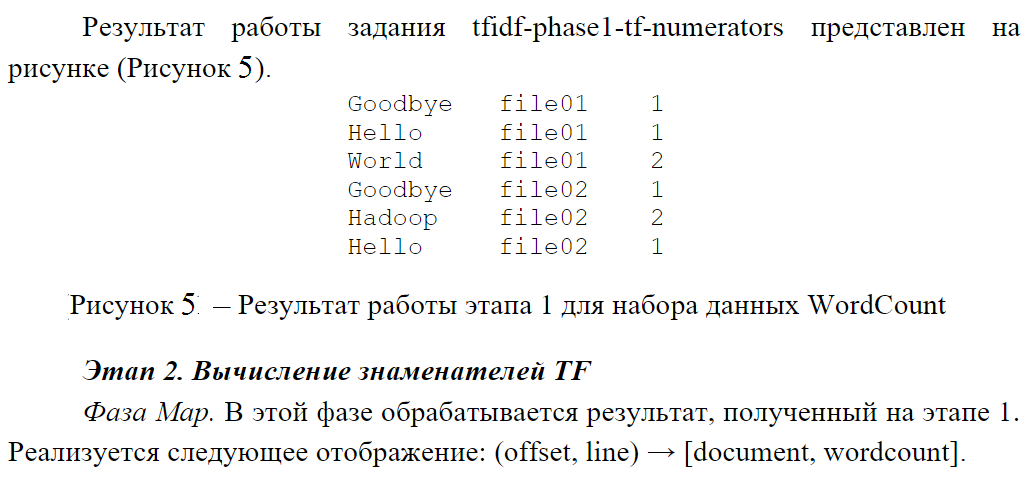
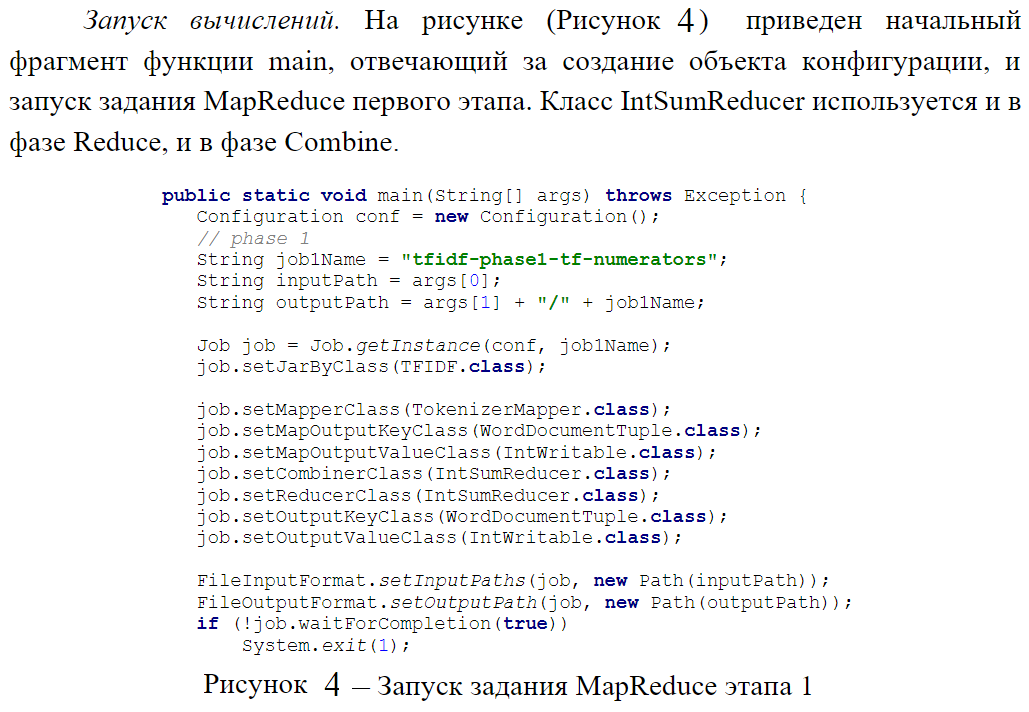
**Электронные ресурсы**

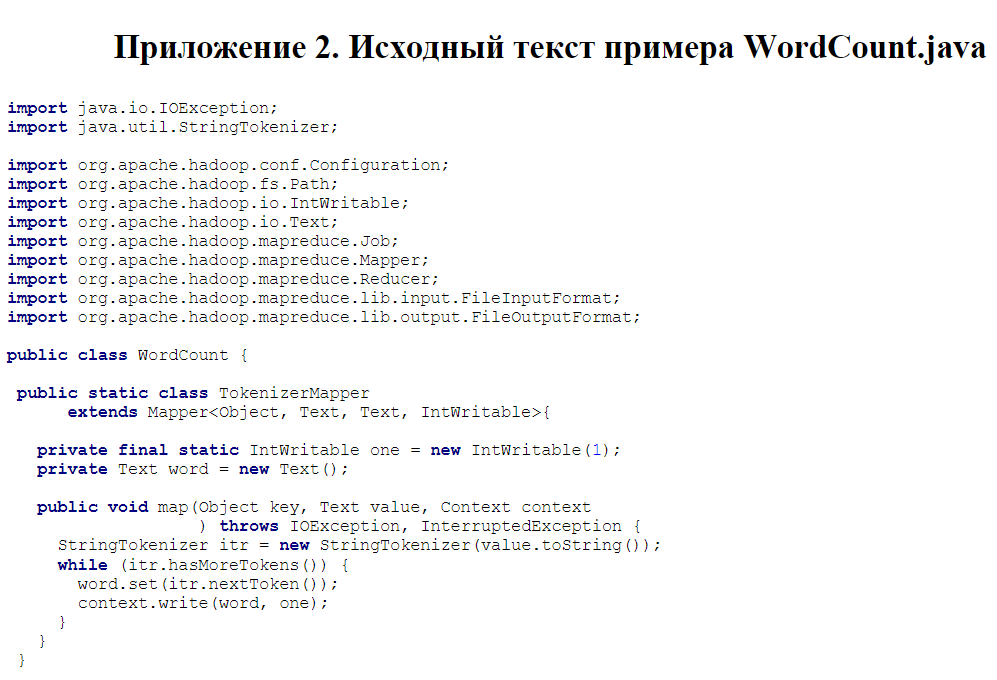
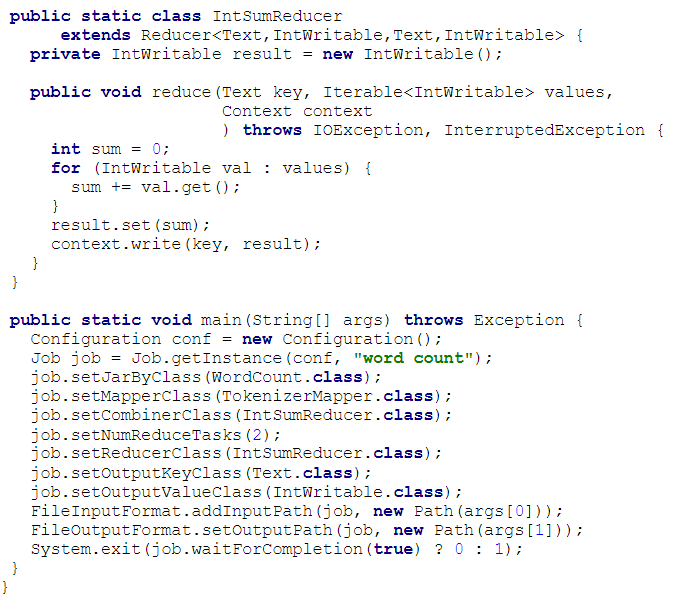
# Hadoop – MapReduce [Электроный ресурс]. – URL: <https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/uchitsia-hadoop/hadoop-mapreduce> Доступ свободный

**Приложение 1. Пример разработки программы с помощью паттерна MapReduce**

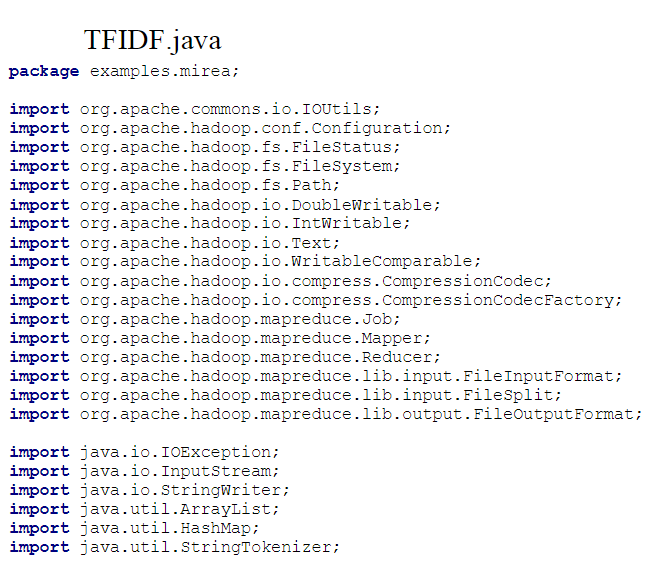
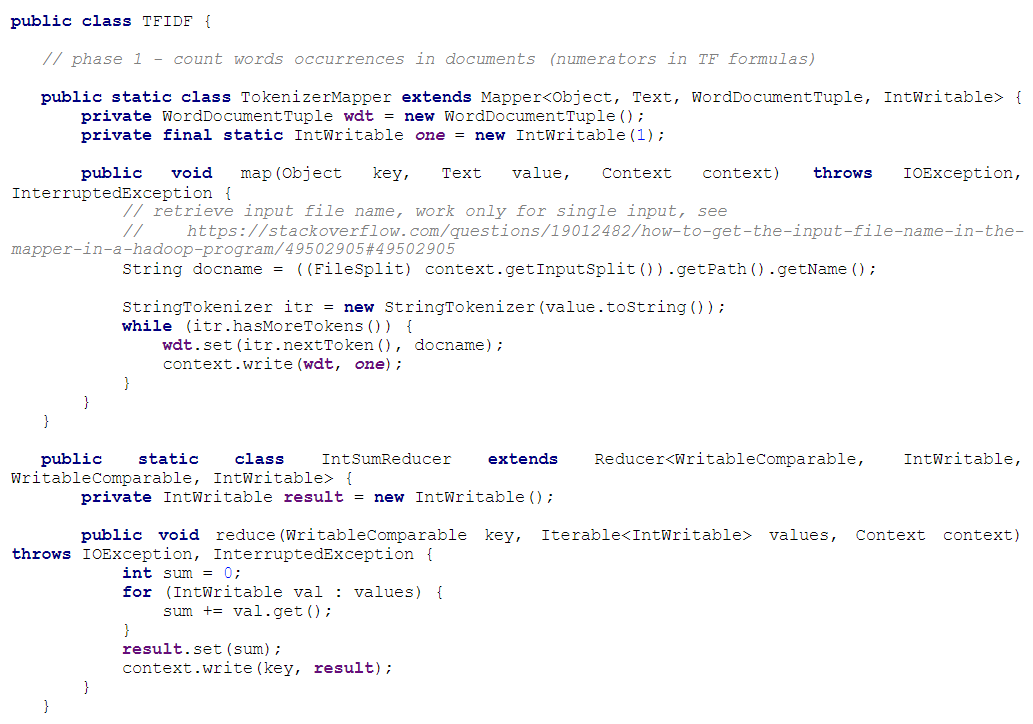
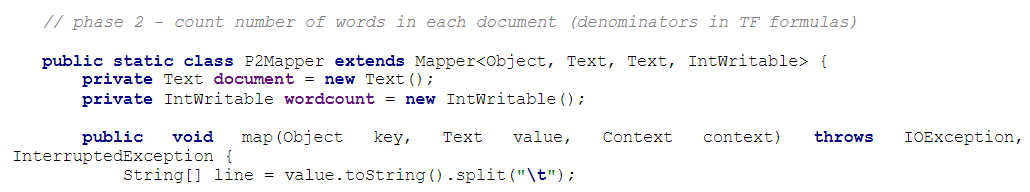
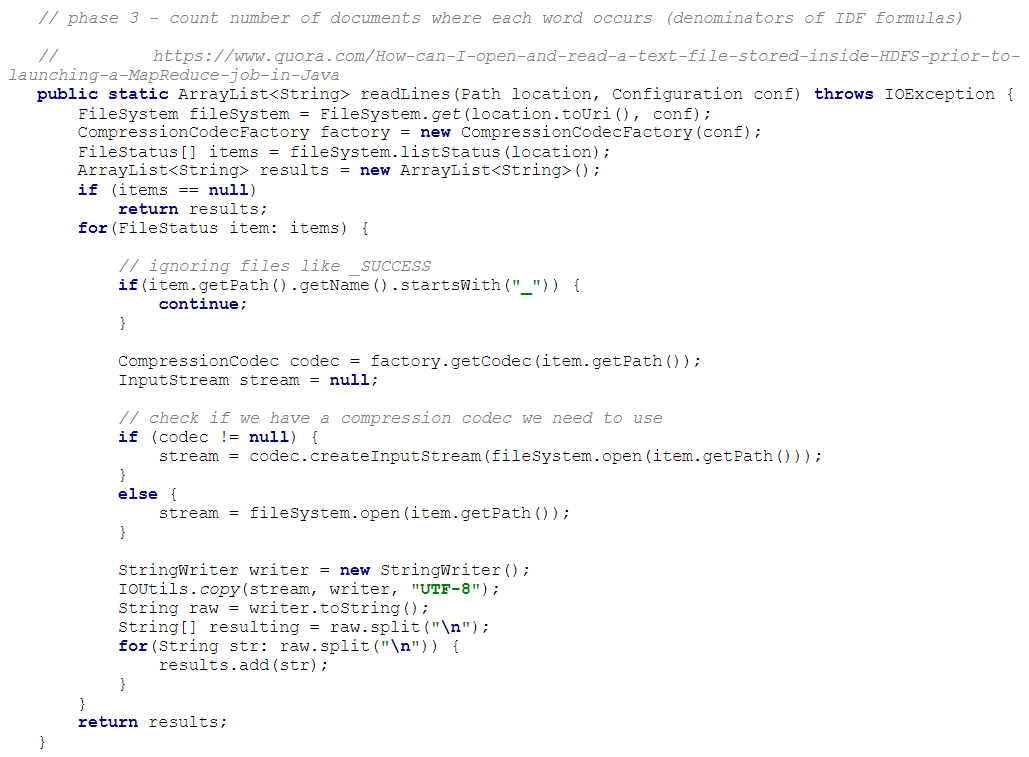
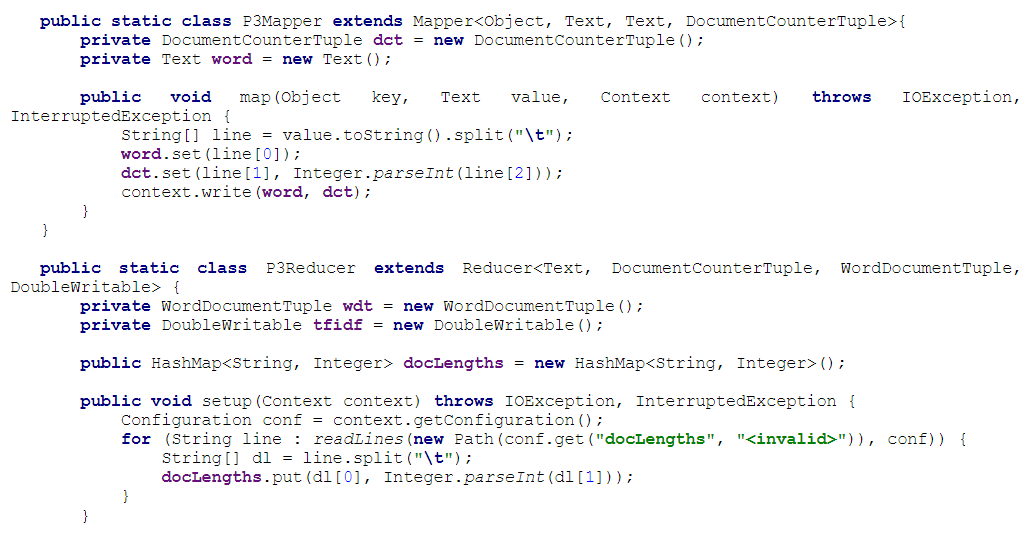
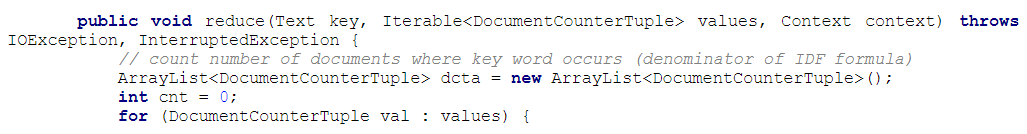
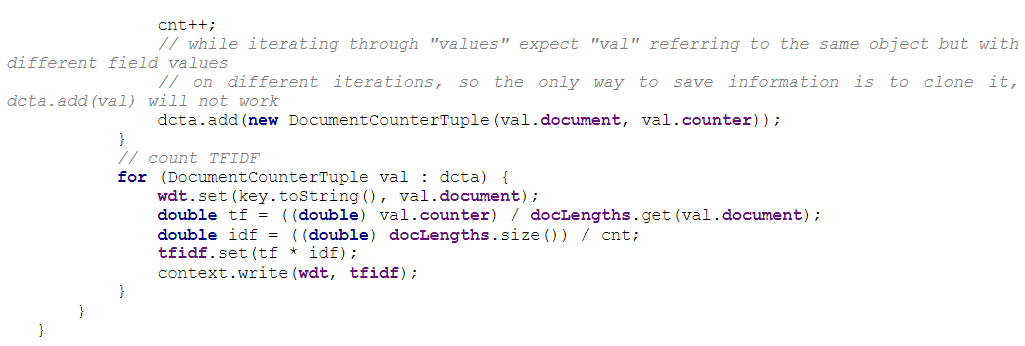


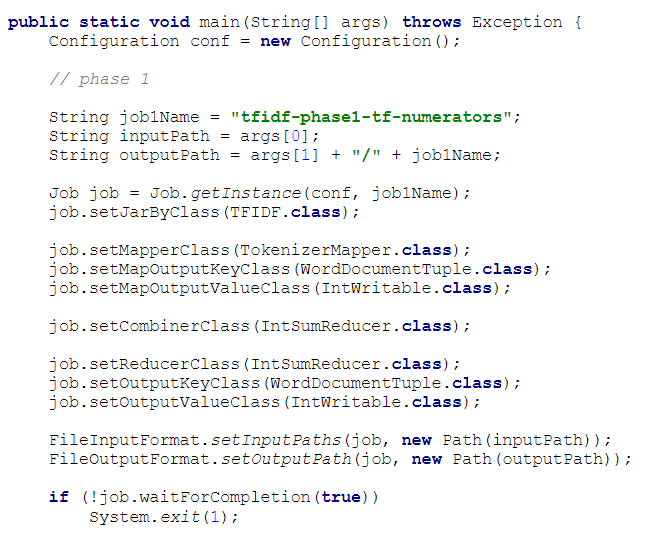
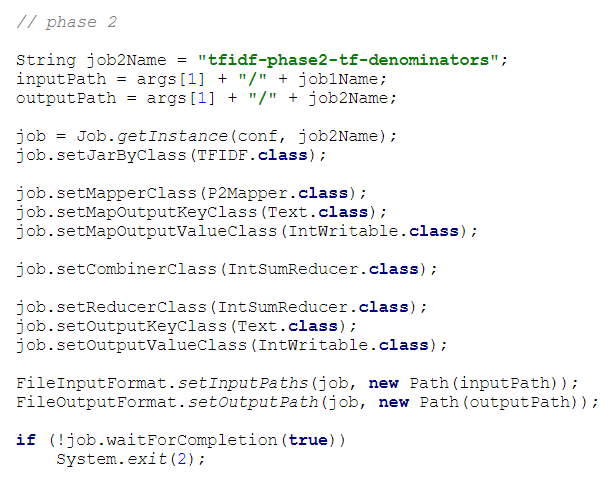
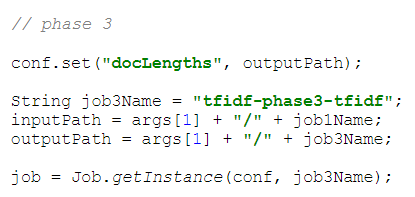
 

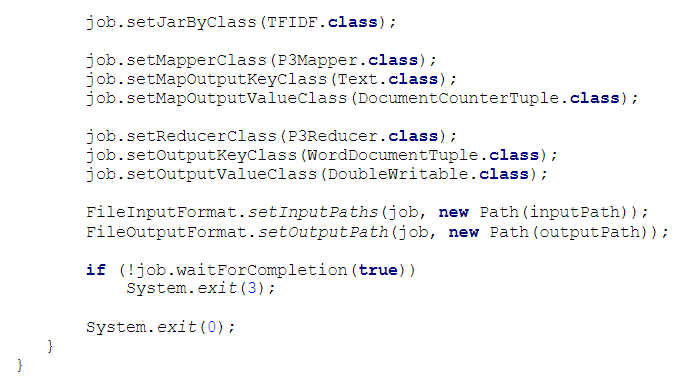
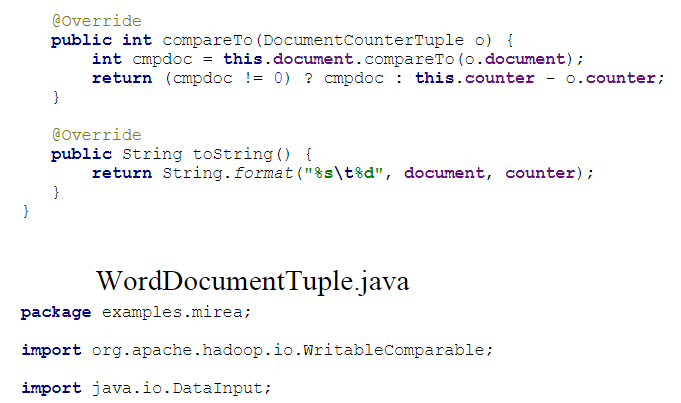


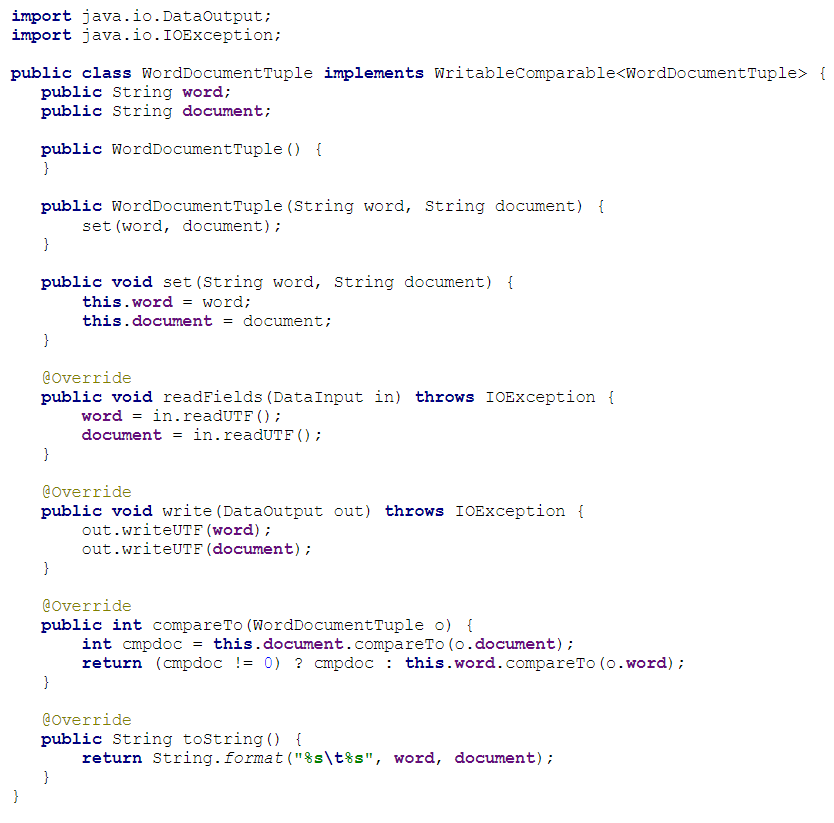
**Приложение 2 Исходный текст примера WorldCount.java **

**Приложение 3. Исходный текст примера TFIDF**

**      **

**  **

**  **

****