Федеральное агентство связи

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Сибирский государственный университет телекоммуникаций и информатики»

(СибГУТИ)

Кафедра вычислительных систем

**Лабораторная работа №4**

«Инвертированный индекс Wikipedia»

Выполнил:

студент гр. МГ-211 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / Бурдуковский И.А./

подпись

Проверил:

Профессор

кафедры ВС \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /Курносов М.Г./

ОЦЕНКА, подпись

Оглавление

[Задание 3](#_Toc133933349)

[Описание работы 4](#_Toc133933350)

# Задание

Требуется построить инвертированный индекс (inverted index) для русской Википедии (и английской — по желанию).

Дампы Википедии в формате XML загружены на кластер Jet: (<http://dumps.wikimedia.org>):

/home/pub/hadoop/lab34/enwiki-\*

/home/pub/hadoop/lab34/ruwiki-\*

Входные данные map:

(docid, content)

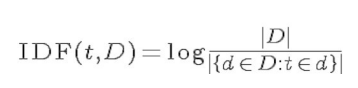
Результирующий инвертированный индекс должен иметь следующую структуру:

(word, [<docid1, TF-IDF1>, <docid2, TF-IDF2>, …])

* Статьи должны быть отсортированы в порядке убывания TF-IDF (Term Frequency – Inverse Document Frequency)
* Для каждого слова ограничить список статей N наиболее релевантными
* Определить и исключить из индекса Top20 высокочастотных слов

При вычислении TF-IDF считаем, что:

* TF(t, d) — это число вхождений слова t в документ d (Wiki-статью)
* IDF(t, D) — обратная частота, с которой слово t встречается во множестве документов D (Wiki-статьях):



* TF-IDF = TF \* IDF

Программы должны быть написаны на языке Java (Apache Hadoop Java API).

# Описание работы

Для работы с XML-дампом была взята специальная версия класса InputFormat – XmlInputFormat и пример реализации подсчета слов в файле xml.

Файлы размещены на сервере:

/home/ipa/pub/hadoop/lab34/xmlinput/wordcount-xml/

Для построения инвертированного индекса программа WordCount.java предварительно подсчитывает количество слов в файле дампа, формируя промежуточный результат. В дальнейшем он используется для определения высокочастотных слов программой TopWordCount.java, которые необходимо исключить при построении инвертированного индекса.

Также программой CountDocs.java из дампа заранее считается общее количество документов, для его дальнейшего использования в формуле при расчете IDF.

На вход программы InvertedIndex.java подаётся файл XML-дампа, список исключаемых слов. общее количество документов и число N для ограничения списка статей в результате.

1. WordCount.Java

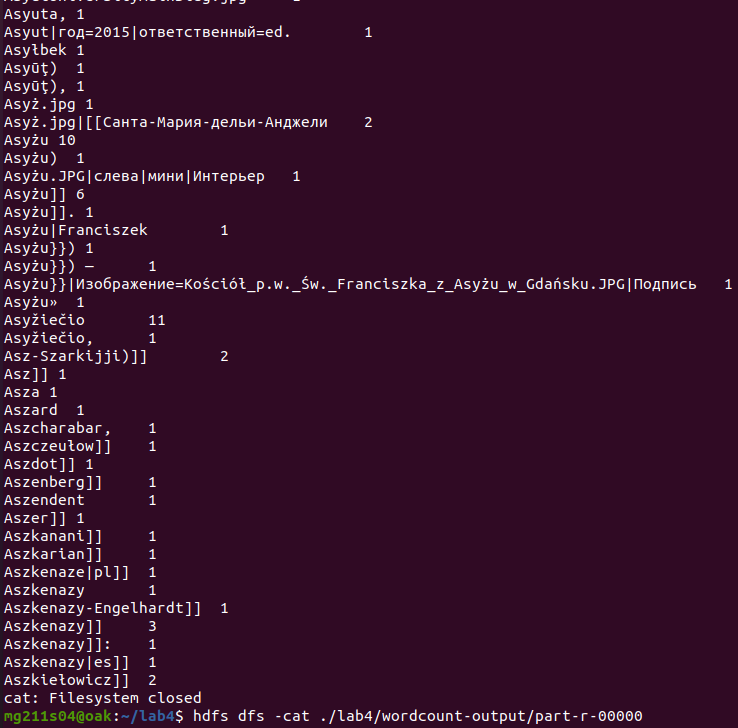
Запуск сборки:



Запуск приложения:



Результат:



1. TopWordCount.java

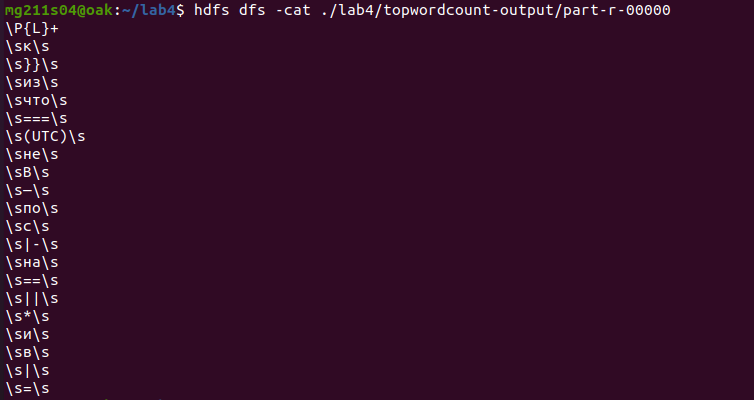
Запуск сборки:



Запуск приложения:



Результат:



1. CountDocs.Java

Запуск сборки:



Запуск приложения:



Результат:



1. InvertedIndex.Java

Запуск сборки:



Запуск приложения:



Результат:

