МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский ядерный университет «МИФИ»

**Обнинский институт атомной энергетики –**

филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ»

**(ИАТЭ НИЯУ МИФИ)**

Отделение интеллектуальных кибернетических систем

ЛАБОРАТОРНАЯ РАБОТА №2

«Изучение возможностей Scala»

по дисциплине

«Анализ литературной работы»

Выполнил студент 1 курса

группы ИВТ-М20

Архипов Д.А.

Проверил:

кандидат технических наук

Грицюк С. В.

Обнинск, 2020

# Цель работы

Разработать приложение для Apache Spark в автономном режиме. Проанализировать литературное произведение (Федор Михайлович Достоевский – Братья Карамазовы).

# Краткая теория

Apache Spark — фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop.

В отличие от классического обработчика из ядра Hadoop, реализующего двухуровневую концепцию MapReduce с хранением промежуточных данных на накопителях, Spark работает в парадигме резидентных вычислений — обрабатывает данные в оперативной памяти, благодаря чему позволяет получать значительный выигрыш в скорости работы для некоторых классов задач, в частности, возможность многократного доступа к загруженным в память пользовательским данным делает библиотеку привлекательной для алгоритмов машинного обучения.

## Структуры данных RDD

RDD – это распределенная коллекция данных, расположенных по нескольким узлам кластера, набор объектов Java или Scala, представляющих данные. RDD работает со структурированными и с неструктурированные данными. Также, как DataFrame и DataSet, RDD не выводит схему загруженных данных и требует от пользователя ее указания.

RDD-коллекция сериализуется каждый раз, когда Spark требуется распределить данные внутри кластера или записать информацию на диск. Затраты на сериализацию отдельных объектов Java и Scala являются дорогостоящими, т.к. выполняется отправка данных и структур между узлами.

# Ход работы

В начале работы происходит инициализация конфигурации состояния Spark, что позволяет в дальнейшем подключить контекст работы для прочтения файла с литературным произведением и дальнейшим взаимодействием с ним.

Для анализа подготовлен исходный текст с помощью устранения пунктуационных знаков: точка, запятая и так далее. В дополнение, слова текста переводятся в нижний регистр для увеличения числа показателей значений.

В процессе обработки текста участвует специальный список стоп-слов, хранящийся в файле stop.txt. При сравнении исходного текста произведения со словами, которые представлены в этом файле, происходит устранение союзов, предлогов, глаголов и других лишних слов.

Затем происходит подсчет всех слов и их reduce. По результату проводится поиск 50 самых популярных и 50 самых редко используемых слов. Результат:

**Top50 most common words:**

(алеша,867)

(митя,587)

(иван,512)

(федорович,501)

(пред,342)

(ибо,310)

(господа,246)

(смердяков,234)

(дмитрий,229)

(отца,224)

(отец,220)

(деньги,214)

(грушенька,200)

(убил,197)

(вовсе,188)

(сердце,188)

(брат,188)

(тысячи,182)

(что‑то,180)

(павлович,167)

(федор,167)

(коля,161)

(глаза,161)

(ужасно,158)

(столь,157)

(ко,154)

(все‑таки,151)

(бог,151)

(начал,148)

(пусть,147)

(алексей,146)

(жизни,142)

(завтра,140)

(главное,138)

(старик,136)

(своим,135)

(напротив,133)

(старец,131)

(как‑то,129)

(денег,129)

(правда,128)

(самое,126)

(давеча,125)

(рублей,122)

(воскликнул,122)

(мальчик,121)

(григорий,121)

(люблю,119)

(петр,118)

(своем,117)

**Top50 least common words:**

(неестественная,1)

(порази,1)

(услужить,1)

(поляку,1)

(собралось,1)

(мучают,1)

(пунктах,1)

(висевшим,1)

(одеяльцем,1)

(уступать,1)

(рыбной,1)

(приобресть,1)

(гильотину,1)

(глубокою,1)

(плеч,1)

(грудным,1)

(маленькими),1)

(старенькой,1)

(подымаясь,1)

(скатертью,1)

(промолчали,1)

(энергично,1)

(мерзким,1)

(провозгласили,1)

(купалось,1)

(раннего,1)

(вязалось:,1)

(жандармы,1)

(европеец,1)

(геологический‑то,1)

(барыням‑то,1)

(схваток,1)

(возраста,1)

(l’inventer,1)

(уничтожается:,1)

(поддельные…,1)

(напророчествовал,1)

(мотивировать,1)

(рыбой;,1)

(слов‑то,1)

(таилось,1)

(дрожащие,1)

(губительницей,1)

(оригинальных,1)

(потерянным,1)

(наткнувшись,1)

(ветчины,1)

(жила‑была,1)

(самому‑то,1)

(радовались,1)

Для семантического поиска однокоренных слов используется stemmer, который позволяет выделять общей корень у родственных (однокоренных) слов. Его использование происходит в методе mapPartitions, который работает с разделами RDD, вместо каждого элемента, что позволяет повысить эффективность.

В процессе лабораторной работы был произведён поиск для 50 самых популярных и редко встречающихся слов с использованием стемминга. Результат:

**Top50 most common stems:**

((алеш,List(алеша, алешей, алеши, алешу, алеше)),1166)

((мит,List(мити, митей, митю, мите, митя)),843)

((федорович,List(федоровича, федорович, федоровичем, федоровичу, федоровиче)),681)

((ива,List(иван, ивана)),617)

((сам,List(самым, самая, самую, самое, самыми, самою, самый, самых, самые)),528)

((мо,List(моей, моего, моим, моему, мою, моими, мое, моем, моих, моею)),520)

((сво,List(своя, свой, своем, своею, своим, своему, своими)),509)

((так,List(таким, таков, тако, таком, таки, такими, такою, таких, такую, таку, такому, такого)),459)

((брат,List(братиею, братий, брату, братия, братьев, брате, братья, братьями, братом, братии, браться, братию, брата, братьях, братией, брат, брать)),414)

((вид,List(видов, видом, видим, видит, вида, видимся, вид, видишь, виду, видят, виде, видите, видя, видах)),398)

((люб,List(любившими, любишь, любуется, любит, любим, любить, люби, любившего, любом, любившая, любят, любили, любя, любите, любивший, любил, любившие, любила, любую)),394)

((отц,List(отца, отцам, отцом, отцами, отцов, отцы, отце, отцу)),385)

((господ,List(господе, господу, господах, господ, господам, господь, господа, господи, господом)),374)

((дмитр,List(дмитрии, дмитрию, дмитрия, дмитрий, дмитрием)),372)

((пред,List(пред, предо)),364)

((кол,List(коле, коля, колю, колена, коли, колей, колен, колено, коль)),353)

((грушеньк,List(грушенька, грушеньки, грушеньке, грушеньку, грушенькой)),345)

((друг,List(другими, другою, другом, друга, друг, други, другим, другому, другую, другого, другу, друге)),345)

((уб,List(убей, убью, убить, убит, убив, убившим, убивший, убил, убили, убейте)),338)

((бог,List(богам, бога, богом, богов, боги, бог, боге, богу)),327)

((федор,List(федоре, федору, федором, федора, федор)),315)

((глаз,List(глазу, глазах, глаза, глазами, глазом, глаз, глазам)),312)

((говор,List(говорят, говоря, говоривший, говорила, говорили, говорите, говорило, говорилось, говори, говором, говорившего, говорено, говорится, говорим, говоришь)),311)

((иб,List(ибо)),310)

((павлович,List(павловичем, павловиче, павловичу, павловича, павлович)),307)

((сердц,List(сердца, сердцами, сердцам, сердцах, сердце, сердцу, сердцем)),303)

((вер,List(верой, веря, верую, вере, верили, верующие, веры, веруют, веруйте, верующий, верите, верующих, верят, верил, веру, верила, верь, вера, веруй, верен, веришь, верующими, верующего, верившего, верит, верить, верю, верующим, верует)),298)

((стол,List(стол, стола, столь, столом, столам, столах, столу, столы, столе)),297)

((котор,List(которою, котором, которое, которым, которому, которую, которыми)),297)

((дом,List(дома, домам, дом, домов, домой, доме, домах, дому)),290)

((душ,List(душах, душу, душили, душе, душой, душ, души, душою, душить, душа)),270)

((деньг,List(деньгу, деньгах, деньги, деньгами, деньгам)),266)

((ваш,List(вашего, вашему, вашим, вашими, вашей, вашу, ваших, вашем, вашею)),263)

((дел,List(делиться, делся, дели, дела, делами, делам, делом, делить, делах, деле, делая, делись, делу)),260)

((слов,List(словам, слов, слово, словами, словом, словах, слове, слову)),260)

((мальчик,List(мальчиками, мальчика, мальчики, мальчиков, мальчиком, мальчикам, мальчик, мальчику, мальчике)),256)

((одн,List(одних, одному, одну, одном, одною, одними, одни)),255)

((тысяч,List(тысяча, тысячи, тысячами, тысячам, тысячах, тысячу, тысяче)),248)

((как,List(каков, каким, како, какими, каком, какою, какому, каку, какое, какие, каких, какого, какую)),247)

((нача,List(началась, начав, начаться, началось, начать, начались, начало, начали, начавшийся, начавшая, начался, начал, начавший)),244)

((наш,List(нашем, наших, нашею, нашу, нашим, нашего, нашими, нашему, нашей)),238)

((смердяк,List(смердяков)),234)

((сдела,List(сделал, сделано, сделала, сделаемся, сделаешься, сделаем, сделаешь, сделался, сделан, сделает, сделали, сделало, сделаюсь, сделанное, сделай, сделана, сделайте, сделав, сделать, сделалось, сделаете, сделают, сделаться, сделались, сделанный, сделаю, сделавшее, сделалась)),231)

((дума,List(думала, думали, думаешь, думай, думает, думать, думавши, думаю, думают, думаете, думайте)),230)

((старик,List(старику, старике, старикам, стариков, старик, старика, старики, стариком)),230)

((мест,List(местам, местами, мест, места, мести, место, месте, месту, местах)),227)

((страшн,List(страшною, страшно, страшных, страшном, страшный, страшнее, страшные, страшную, страшным, страшной, страшная, страшное, страшного, страшному, страшнейшая, страшными)),222)

((час,List(часов, часа, часом, часам, часами, час, часах, часы, часу, часе)),221)

((ивановн,List(ивановною, ивановна, ивановны, ивановну, ивановне, ивановной)),220)

((отец,List(отец)),220)

**Top50 least common stems:**

((графин,List(графина)),1)

((пойти…,List(пойти…)),1)

((заболет,List(заболеть)),1)

((подростк,List(подростка)),1)

((пренеприличн,List(пренеприлично)),1)

((истинную:,List(истинную:)),1)

((расчетл,List(расчетлив)),1)

((сюртуком…,List(сюртуком…)),1)

((маленькими),List(маленькими))),1)

((широт,List(широты)),1)

((вязалось:,List(вязалось:)),1)

((биллиардн,List(биллиардной)),1)

((утверждаете…,List(утверждаете…)),1)

((европеец,List(европеец)),1)

((святош,List(святоши)),1)

((схваток,List(схваток)),1)

((возраста,List(возрастало)),1)

((l’inventer,List(l’inventer)),1)

((высокооткровен,List(высокооткровенного)),1)

((уничтожается:,List(уничтожается:)),1)

((поддельные…,List(поддельные…)),1)

((рыбой;,List(рыбой;)),1)

((самообладан,List(самообладания)),1)

((снег‑т,List(снег‑то)),1)

((помыка,List(помыкал)),1)

((протерш,List(протершеюся)),1)

((разыгрыва,List(разыгрывать)),1)

((сопоставлен,List(сопоставление)),1)

((грушеньке;,List(грушеньке;)),1)

((помещич,List(помещичьем)),1)

((блажные‑с,List(блажные‑с)),1)

((вытянеш,List(вытянешь)),1)

((разгорож,List(разгорожена)),1)

((предосадн,List(предосадное)),1)

((заглянув,List(заглянув)),1)

((новшеству…,List(новшеству…)),1)

((приходили‑с,List(приходили‑с)),1)

((regierender,List(regierender)),1)

((прелюбод,List(прелюбодей)),1)

((неотступн,List(неотступным)),1)

((детектив,List(детективом)),1)

((была‑с,List(была‑с)),1)

((невозмутим,List(невозмутимо)),1)

((легкомыслия;,List(легкомыслия;)),1)

((тронут…,List(тронут…)),1)

((объяден,List(объядении)),1)

((охватыва,List(охватывала)),1)

((раз‑с,List(раз‑с)),1)

((выйдут‑с…,List(выйдут‑с…)),1)

((сентябрьск,List(сентябрьскую)),1)

# Вывод

В ходе лабораторной работы была настроена сборка проекта с помощью Maven для версии языка Scala 2.11.0. Подключена основная библиотека для фреймворка Spark и написан исходный код программы для анализа романа Ф. М. Достоевского «Братья Карамазовы».

В процессе были выявлены 50 самых популярных и 50 самых редко используемых слов. Полученные результаты были дополнены процессом стемминга, что позволило глубже изучить семантический анализ текста. Благодаря расширяемому списку стоп-слов проект может быть расширен в рамках более детального рассмотрения.

# Листинг кода



