По условию задачи на анализ будет подаваться файл и java-кодом, следовательно, такой текстовый файл почти наверняка будет соответствовать следующим условностям:

Ус1) В файле должны быть строки с кодом

Ус2) В файле помимо исключительно кода могут быть пустые строки, однострочные и многострочные комментарии.

Ус3) В строках с кодом используются отступы.

Ус4) Отступы могут состоять из пробелов, из табуляций, либо из того и другого вперемешку.

Ус5) Одна табуляция соответствует какому-то целому числу пробелов (если и табуляции, и пробелы присутствуют)

Ус6) Один отступ соответствует какому-то целому числу пробелов (если пробелы присутствуют)

Ус7) Один отступ соответствует какому-то целому числу табуляций (если табуляции присутствуют)

На основании условия задачи и всего вышесказанного, программе поставлена цель:

Ц) На вход программы подается файл с java-кодом, на выходе программа должна указать, сколько пробелов приходится на один отступ (если пробелы есть), сколько табуляций приходится на один отступ (если табуляции есть), сколько пробелов приходится на одну табуляцию (если и пробелы, и табуляции есть).

Так как это java-код, то: (предположение)

П1) В файле нем обязательно будет хотя бы одна скобка ‘{‘, после которой следующая строка с кодом будет начинаться с отступа. Если это не так, то в файле либо проблемы с кодом, либо проблемы с форматированием.

Если стиль форматирования более-менее соблюдался (самим программистом или с помощью редактора), то можно предположить, что:

П2) Скорее всего все строки кода одного блока будут начинаться с одного и того же отступа

П3) Разность отступа между первой строкой в блоке кода и строкой, открывающей этот блок (т.е. строка с последним символом ‘{‘) будет примерно одинакова для блоков кода любого уровня вложенности.

Заметим следующее:

Замеч1) Предположения (П1 и П3) по аналогии справедливы и для закрывающей скобки.

Замеч2) По предположению (П2) делаем вывод, что если была совершена неточность в первом отступе открытого блока кода, то и все остальные строки этого уровня повторят его неточность.

Исходя из всего вышесказанного, принято решение принимать во внимание только отступы кода после открытия очередного блока кода. А на основании полученных данных сделать все необходимые выводы.

Алгоритм решения:

шаг1) Обнаружить все добавочные отступы в первой строке каждого открытого блока кода (как это описано в (П3) ).

шаг2) Проверить наличие пробелов и табуляций в найденных отступах.

шаг3) При наличии и пробелов, и табуляций определить, сколько пробелов приходится на одну табуляцию

шаг4) При наличии пробелов выяснить, сколько пробелов приходится на один отступ. Аналогично для табуляций.

Методика (шаг1):

По строкам посимвольно будет проходить «курсор», который будет обнаруживать все необходимые параметры для текущей строки: отступ, наличие кода, наличие комментария, наличие открывающей скобки ‘{‘.

При попадании «курсора» в конец строки, если строка содержала код, то запоминается, ее все ее параметры. Если предыдущая строка кода содержала открывающую скобку ‘{‘, то вычисляется добавочный отступ и заносится в коллекцию.

Если «курсор» попадает в комбинацию “/*”, то есть открывается многострочный комментарий, то «курсор» входит в состояние

«многострочного комментария» и ходит далее по строкам до тех пор, пока не найдет конец этого комментария. В состоянии «многострочного комментария» никаких действий, кроме поиска комбинации “*/” не производится.

Методика (шаг2):

Выполнить обход коллекции найденных добавочных отступов и найти значения пробелов и значения табуляций, большие нуля.

Методика(шаг3):

Выполняется перебором: предполагается, что кол-во пробелов, приходящихся на одну табуляцию, равно X. Очевидно, что X должен быть больше 0. Скорее всего, X будет меньше либо равно некоторого значения Xmax. По (Ус5) X – целое. Значит будем перебирать X от 1 до Xmax. Далее выражаем все добавочные отступы через пробелы. В идеале, если X попал в значение, которое использовалось в редакторе, в котором писали идеальный код, то все полученные отступы будут равны. А если X не попал в это значение, то появятся сильные различия. В реальном случае, где X попал в нужное значение, отступы будут более-менее равны, а если не попал, то различия будут сильные.

Меру различия решено определять значением среднеквадратического отклонения выборки полученных отступов для очередного X. Искомым X выбирается тот X, при котором среднеквадратическое отклонение будет наименьшим. (тот самый X, где отступы «более-менее равны»).

Методика (шаг 4):

Чтобы сравнивать отступы, которые состоят из пробелов и табуляций, выражаем их через пробелы, используя результат шага (шаг3).

Если формат более-менее выдержан, то «правильные» отступы будут встречаться в нем чаще, чем любой из «неправильных». Поэтому за правильный отступ выбирается тот, который использовался чаще всего.