Министерство образования и науки Российской Федерации

САНКТ–ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И ПРОГРАММИРОВАНИЯ

КАФЕДРА ИНФОРМАЦИОННЫХ СИСТЕМ

Лабораторная работа №1-2

по дисциплине: Представление знаний в информационных системах

Суммаризация текстового форума

Выполнили:

Виноградов П.Д.

Трофимов В.А.

Преподаватель: Гусарова Н.Ф.

Санкт–Петербург

2015

# Описание исследуемых форумов

Для анализа были выбраны следующие текстовые форумы на русском языке:

* <http://forum.velomania.ru/showthread.php?t=74626> Велофорум, топик посвящен обсуждению рядовой проблемы велосипедистов – возврату поршня в калипер дисковой тормозной системы велосипеда. Форум содержит большое количество оффтопа, т.к. в процессе обсуждения основной темы форумчане перешли на обсуждение сторонних нетривиальных физических явлений. На форуме явно выделяются несколько самых активных авторов. Топик содержит чуть больше 500 сообщений, из которых все были взяты для анализа. Ключевые слова: 'калипер', 'порш', 'тормоз', 'возвращ', 'диск' (некоторые слова были умышленно обрезаны для корректного нахождения в составе видоизмененных слов, а также слов с измененной формой суффикса и окончания)
* <http://forum.guitarplayer.ru/index.php?topic=219514> Музыкальный форум, топик посвящен настройке гитары. Содержит малое количество оффтопа в виде неинформативных сообщений не по теме. Специфика данного топика такова, что ключевая полезная информация по настройке гитары заключена в первом сообщении топика (топикстартере), а также в нескольких других сообщениях. Большая часть оставшихся сообщений связана с вопросами форумчан по решению проблем с настройкой их гитар, а также ответами на них. Данные сообщения не несут полезной информации в общем случае для настройки гитары, вопросы очень узко направленные и связаны с конкретной моделью гитары, конкретными струнами и т.п. Всвязи с этим при разметке *Utility* для таких сообщений было выбрано значение 1 по шкале от 0 до 5 (не является оффтопом, не несет полезной информации, но связано с темой топика). В качестве эксперимента было принято решение увеличить значение *Utility* для данных сообщений с 1 на 2 и повторить исследование. Топик содержит около 1500 сообщений, из которых первые 500 были взяты для анализа. Ключевые слова: 'гитар', 'струн', 'строй', 'гриф', 'анкер', 'лад', 'мензур', 'тремол'

# Автоматизация процесса анализа форумов

* 1. Автоматизация процесса перевода топика в табличный формат

Для ускорения процесса переноса форума в табличный формат с автоматической заменой картинок, видео, а также ссылок на них на специальные метки для последующего сбора статистики были написаны скрипты на языке Python 3 для обоих форумов с учетом специфики их HTML-разметки.



Данные скрипты генерируют файлы с табуляцией в качестве разделителей (tsv). Данный формат файла можно открыть с помощью пакета Microsoft Excel. Для удобства просмотра и модификации к файлам tsv приложены файлы в формате Excel 2007.

Примечание: в прикрепленных файлах добавлены столбцы *Sentiment* и *Utility*, которые были размечены вручную посредством анализа контекста и структуры сообщений форумов.



* 1. Автоматизация сбора статистики по размеченным файлам и формирования социальных графов

Для ускорения процесса сбора статистики по сообщениям при последующей переразметке форума были написаны скрипты на языке Python 3.



В описании к лабораторной работе предложен метод определения цитирования сообщений – если в сообщении упоминается имя (никнейм) одного из авторов на форуме, то это значит, что данное сообщение цитирует последнее сообщение упомянутого автора. Данная методика определения цитирования показалась очень наивной, поэтому было принято решение реализовать более сложную методику определения цитирования.

При переносе данных в табличный формат сохраняется информация о том, что часть сообщения автора является цитатой сообщения некоторого другого автора (эта информация сохраняется в разметке, однако какое именно сообщение цитируется, не указывается). Поэтому, разбив текст цитаты на слова, производится поиск данной последовательности слов в предыдущих сообщениях цитируемого автора. Первое (хронологически) сообщение с найденной последовательностью слов считается цитируемым.

Для получения более комплексного анализа, анализу будут подвержены обе статистики по сообщениям – с наивным цитированием и с предложенным.

Приложенные скрипты принимают на вход файлы, сгенерированные предыдущими скриптами, и генерируют tsv файлы со статистикой по сообщениям.



Также эти скрипты генерируют tsv файлы с социальными графами для каждого из двух вариантов расчета цитирования:

* Социальный граф с количеством цитирований (clean raw) – матрица смежности, в узлах которой содержится число цитирований одного автора другим



* Социальный граф c количеством цитирований нормализованный (clean normalized) – матрица смежности, в узлах которой находится обратное число цитирований одного автора другим (если число цитирований 0, указывается некоторое большое число порядка )



* Социальный граф с количеством цитирований с учетом настроения (sentim raw) – матрица смежности, в узлах которой содержится сумма значений *Sentiment* для каждого цитирования одного автора другим (значения проградуированы от 1 до 5)



* Социальный граф с количеством цитирований с учетом настроения нормализованный (sentim normalized) – матрица смежности, в узлах которой содержится сумма обратных значений *Sentiment* для каждого цитирования одного автора другим (значения проградуированы от до 1, если число цитирований 0, указывается некоторое большое число порядка )



* 1. Автоматизация анализа социальных графов и сбора статистики по авторам

Для ускорения расчета *Betweenness* и *BetweennessSent* на основе анализа социальных графов были написаны программы на Java 8.





Для расчета этих характеристик был выбран вариант расчета по определению – через количество кратчайших путей в социальном графе, проходящих через рассматриваемую вершину. Для этого был произведен подсчет кратчайших путей в социальных графах между всеми вершинами. Для этого и производилась нормализация графов – кратчайшим маршрутом является маршрут с наименьшим весом. Для расчета кратчайших путей между всеми вершинами во взвешенном ориентированном графе без циклов отрицательного веса был выбран алгоритм [Floyd–Warshall](https://en.wikipedia.org/wiki/Floyd%E2%80%93Warshall_algorithm). Асимптотическая сложность данного алгоритма составляет , где – размер уникального множества авторов в выборке.

Пример расчета *Betweenness* и *BetweennessSent*:

Рассмотрим два кратчайших пути, проходящих через вершину 5: . При этом вес ребер: . Тогда величина *Betweenness* для вершины 5*,* исходя из определения, равна двум. Величина *BetweennessSent* для вершины 5 равна . При этом (при отсутствии других вершин и связей между ними) для вершин 1, 2, 3, 9 величина *Betweenness* будет равна 1, а *BetweennessSent* для .

На выходе программа выдает tsv файлы с характеристиками *Betweenness* и *BetweennessSent* для каждого варианта цитирования.





Сбор статистики по авторам осуществляется с помощью скрипта на языке Python 3.



На вход подаются все tsv файлы, полученные ранее. Скрипт генерирует файл со статистикой по авторам в формате tsv для каждого из используемых вариантов цитирования.



* 1. Сравнение результатов по двум алгоритмам цитирования

Исследуемые алгоритмы цитирования дают сильно разные результаты – это видно и по матрицам смежностей невзвешенных социальных графов – в особенности для музыкального форума прослеживается, что большая часть информационного потока при использовании алгоритма наивного цитирования скапливается около двух авторов, имена которых употребляют в большом количестве постов без какого-либо намека на цитирование. При использовании наивного цитирования статистика сильно портится также потому, что процитированным считается последнее сообщение автора, что не всегда является правдой.

Альтернативный вариант цитирования, который был предложен, пытается восстановить исходное цитирование, которое было применено средствами разметки форума. Для этого в каждом сообщении ищется терм, сигнализирующий о явном цитировании автора средствами разметки форума (Сообщение от, Цитата и т.п.). Далее извлекается имя процитированного автора, среди всех сообщений автора, следовавших до момента цитирования, ищется последовательность процитированных слов. Первое (хронологически) сообщение, в котором найдена процитированная последовательность слов, обязана принадлежать исходно процитированному посту. Данный метод был реализован и отлично восстановил исходную цепочку цитирований для обоих форумов. Недостатком данного метода является то, что он не учитывает ответы на сообщения, идущие друг за другом без использования средств цитирования. Оптимизацией данного алгоритма будет являться добавление некоторых эвристик из наивного метода цитирования для учета сообщений, идущих друг за другом без явного цитирования.

Ввиду различности подходов результаты получились также сильно разными – величины *Betweenness* и, соответственно, *BetweennessSent* стали таковы, что для каких-то авторов поток информации при одном алгоритме нулевой, а при другом является очень весомым, и наоборот. Также сильно изменилось распределение авторов с большим значением *Betweenness* – при использовании обоих методов два-три самых значимых автора в любом случае имели самое большое значение данной характеристики, однако в случае авторов со средним потоком информации на форуме (выявлено экспертным методом), наивный метод приравнял их к авторам со слабым потоком информации, а метод явного цитирования наоборот, заметно «приподнял» их над малозначащими авторами. Величины *BetweennessSent*, *Degree* и *DegreeSent* изменились аналогичным образом ввиду той же причинно-следственной связи – более точному выделению значимых цитирований с помощью алгоритма явного цитирования.

* 1. Автоматизация процедуры машинного обучения и сбора результатов

Моделирование осуществлялось в пакете Weka для языка программирования Java.



Для сравнительного анализа были выбраны следующие классификаторы:

* LibSVM
* AdaBoostM1
* SimpleLogistic
* RandomTree
* NaiveBayes
* PART
* Grading
* Dagging
* Stacking
* IB1
* KStar
* ClassificationViaRegression
* RacedIncrementalLogitBoost
* RBFNetwork
* IBk
* RotationForest
* ND
* ZeroR
* MultiClassClassifier
* FT
* RandomSubSpace
* MultiScheme
* DecisionStump
* ConjunctiveRule
* J48
* END
* ClassBalancedND
* J48graft
* SMO
* HyperPipes
* StackingC
* JRip
* NNge
* MultilayerPerceptron
* DecisionTable
* FilteredClassifier
* Bagging
* Logistic
* LogitBoost
* MultiBoostAB
* Vote
* VFI
* AttributeSelectedClassifier
* OneR
* CVParameterSelection
* DTNB
* BFTree
* LWL
* DataNearBalancedND
* LMT
* Ridor
* Decorate
* NBTree
* REPTree
* LADTree
* RandomCommittee
* ClassificationViaClustering

Для каждого из представленных классификаторов была проведена операция бутстрепа на 100 итераций, датасетом являлась статистика по постам, полученная в результате работы скриптов, представленных в п.2.2.

Датасеты были сконвертированы из формата tsv в формат arff с помощью скрипта на языке Python 3:

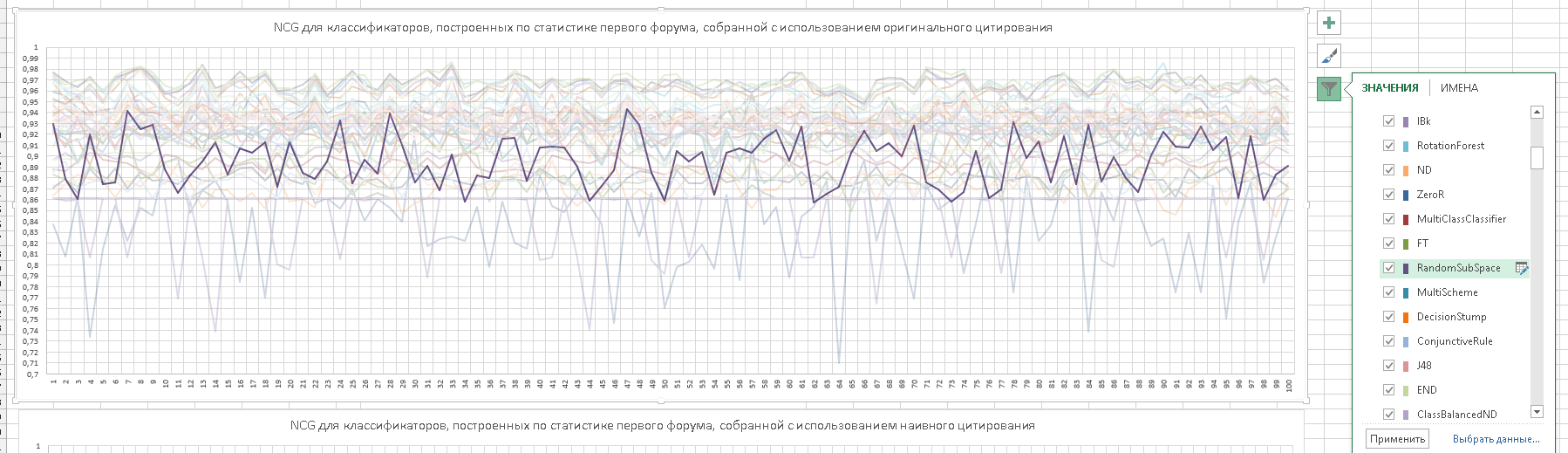


Для музыкального форума, для каждого из двух датасетов, была добавлена копия датасета, в которой значение было увеличено с 1 на 2.

Информация, сгенерированная в результате написанной программы, была сведена в один Excel файл – по отдельному листу на каждый проанализированный датасет (в виде двух таблиц – исходной и, ниже, транспонированной, для удобства построения графиков и восприятия информации), а также отдельный лист (первый), в котором предоставлены графики NCG для каждого из 6 датасетов.

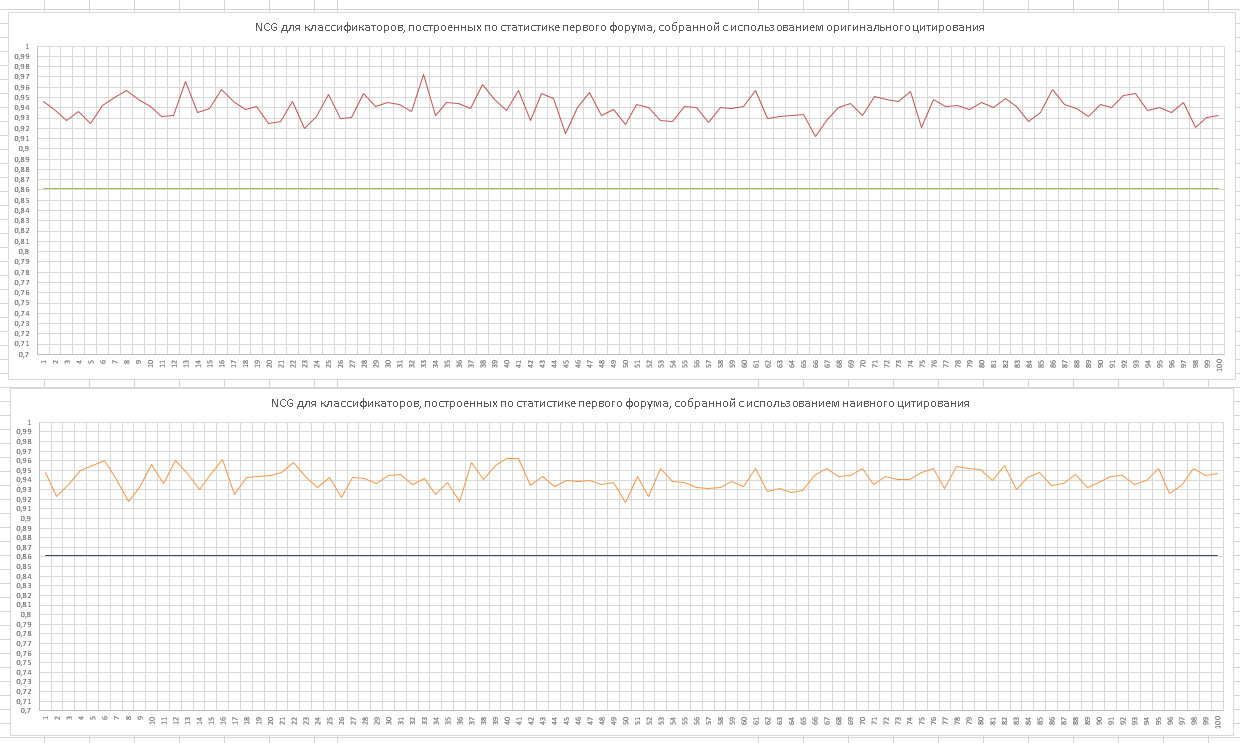


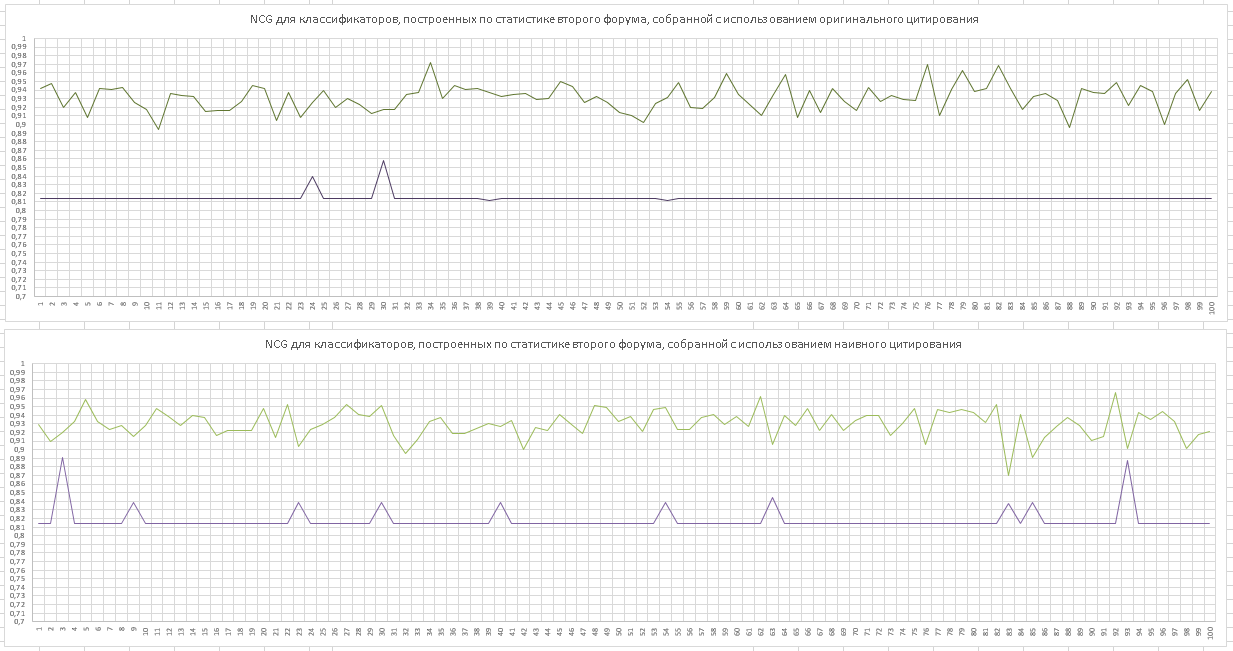
Ввиду большого количества классификаторов (большого количества графиков на одной диаграмме), пришлось отказаться от построения доверительного интервала для NCG – ввиду ограничений Excel, это необходимо делать вручную для каждого графика на диаграмме в отдельности. Несмотря на большое нагромождение графиков на диаграммах, средствами Excel можно удобно и интерактивно смотреть и сравнивать любые подмножества графиков:

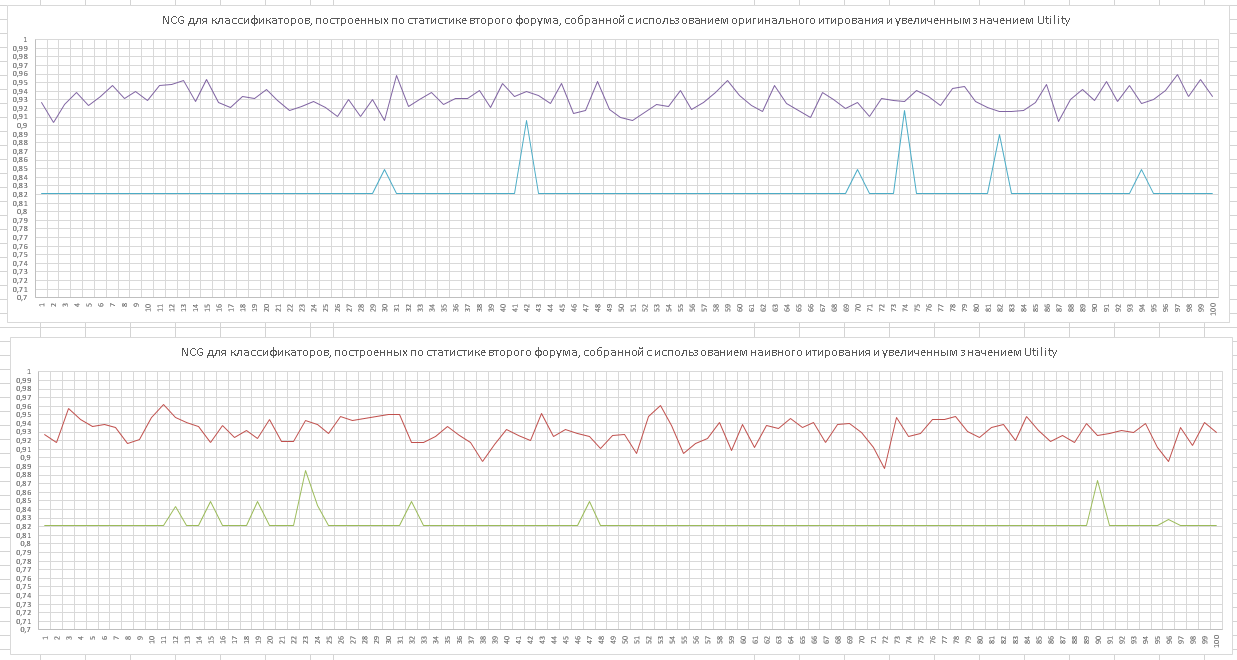


Если говорить о классификаторе, который предложен к рассмотрению в лабораторной работе – СГБ, то он проявил себя крайне плохо – на велофоруме, содержащем большое количество оффтопа, он предсказывал, что любое сообщение является оффтопом, на каждой итерации бутстрепа. Это можно увидеть на графике NCG – линия горизонтальная и монотонная. Схожая картина наблюдается и в случае музыкального форума, однако для него на графиках наблюдаются периодические скачки.

Что касается предложенного к сравнению классификатора SVM, то он показал лучше по сравнению с СГБ результаты, верно определял посты с большим значением , одна по сравнению со всеми классификаторами, показал очень средние результаты. Ниже представлены графики NCG для СГБ и SVM (график, расположенный выше –SVM).







Если рассматривать весь набор классификаторов и оценивать их по значению NCG для каждого из проанализированных датасетов, то лучшими классификаторами показали себя:

* RandomTree
* IB1
* KStar
* IBk
* NNGE
* RandomCommittee

Характеристики данных классификаторов лучше, по сравнению с другими классификаторами, также по следующим параметрам:

* Лучшие значения микро и макро f-мер с относительно малым значением дисперсии каждой величины
* Лучшие показатели по процентам правильно предсказанных

В качестве пояснения к последнему пункту стоит добавить, что велосипедный форум содержит примерно 72% сообщений, являющихся оффтопом (), а музыкальный форум – примерно 90% сообщений, которые являются слабо информативными (, или 2, в случае увеличения).

Всвязи с этим, можно ввести оценку, основанную на том, что процент верно отгаданных должен быть строго больше 72 (90) процентов в зависимости от рассматриваемого датасета. При это чем большее значение будет наблюдаться, тем лучше.

Классификаторы, выбранные лучшими, также являются лучшими при оценивании с помощью предложенной метрики – для велосипедного форума процент правильно отгаданных значений превосходит 95% с дисперсией менее одного процента. Для музыкального форума это 90% с дисперсией менее одного процента. Другие классификаторы имеют худшие показатели.

Также есть классификаторы, которые показали себя хорошо на датасетах одного форума, и плохо на датасетах другого – данные классификаторы не вошли в состав лучших, т.к. хорошо предсказывали только с учетом специфики одного форума:

* HyperPipes
* JRip
* VFI