ПРИКЛАДНОЙ АНАЛИЗ ДАННЫХ: ОТ ТЕОРИИ К ПРАКТИКЕ

Д.И. Пиршту́к
Руководитель группы анализа данных в компании InData Labs

InData Labs, Республика Беларусь
E-mil: d_pirshtuk@indatalabs.com

Abstract. In this paper it is investigated the problem of internships for students in applied data analysis. We generalize our experience into training and research process at our data science R&D laboratory in collaboration with Institute of Applied problems of Mathematics and Informatics in Belarusian State University. We describe our working processes, research projects, which we have proposed to solve, features and problems, achievements. In conclusion, we express their suggestions for improving the training of specialists in the field of data analysis in Belarus.

В 2015 году компания InData Labs проводила второй отборочный конкурс для студентов и магистрантов Белорусского государственного университета в совместную научно-исследовательскую лабораторию, организованную компанией совместно с НИИ прикладных проблем математики и информатики БГУ. По результатам отбора 5 талантливых студентов на протяжении одного семестра под наставничеством сотрудников компании решали прикладные задачи по анализу контента в социальных сетях.

В настоящем докладе мы хотим рассказать про особенности задач, которые решали участники лаборатории, с какими проблемами мы столкнулись, каких успехов достигли.

1 Инфраструктура студенческой лаборатории. Одной из первых серьезных проблем, с которой мы сразу столкнулись с первым составом студенческой лаборатории, был недостаток вычислительных ресурсов: для рациональных экспериментов по анализу данных в социальных сетях объем оперативной памяти на ноутбуках слишком мал. Например, загрузка предобученных моделей Word2Vec уже требует порядка 5 Gb RAM. А раздать каждому студенту сотни гигабайт намайненных в социальных сетях данных вообще невозможно. Кроме того, большинство студентов не были продвинутыми Linux-пользователями, многие пользовались ОС Windows и были знакомы с языком программирования R, но не Python, знакомство и настройка Python Data Science-окружения занимала слишком много времени, помочь каждому с настройкой достаточно трудно.

Библиотека БГУИР

При втором наборе в лабораторию все студенты оказались уже в большей или меньшей степени были знакомы с языком программирования Python, Jupiter Notebook-ами (при конкурсном отборе мы получили лишь одно решение на языке R, остальные были именно в формате *.ipynb) и большинство были пользователями Ubuntu Linux. Однако мы решили проблему радикальным способом: предоставили участникам лаборатории в коллективное пользование один общий выделенный изолированный сервер-“песочницу” в датацентре (6-ядерный процессор Intel Core i7, 64 Gb RAM, 3 Tb HDD (RAID 1), интернет-подключение 200 Mb/s) с уже настроенным типичным для Data Science-окружением, аналогичном тому, что используем сами в компании, и доступом по ssh/sshfs. Преднастроенное окружение включало в себя

- CentOS 7 с установленным Development Tools
- Anaconda Python (Python Data Science Toolbox), Java 8
- Jupiter Notebook 4
- Dataiku Data Science Studio Community Edition

- Много наборов данных в общей папке в /usr/share/indatalabs/…
- Сервис Github.com и git lfs для хранения исходного кода и обученных моделей.

Другое программное обеспечение устанавливалось по мере необходимости, но сразу после установки становилось доступным всем студентам.

Дополнительным преимуществом такого решения было то, что

- сотрудники компании имели постоянный доступ к коду участников лаборатории на сервере и всегда быстро проконсультировать по возникающим техническим вопросам и проблемам;
- удобно было запускать ресурсоемкие эксперименты "на ночь";
- размещение сервера в дата-центре с качественным интернет-соединением упрощало эксперименты по майнингу данных.

2 Критерии отбора задач для лаборатории. При выборе задач мы руководствовались следующими принципами:

1 Задачи (мини-проекты) должны быть именно в области Data Science, то есть максимально соответствовать диаграмме Венна на рис. 1. Хорошего знания только теоретической информатики, программирования и статистики должно было быть недостаточно, т.к. целью лаборатории было дать реальный практический опыт.

Библиотека БГУИР

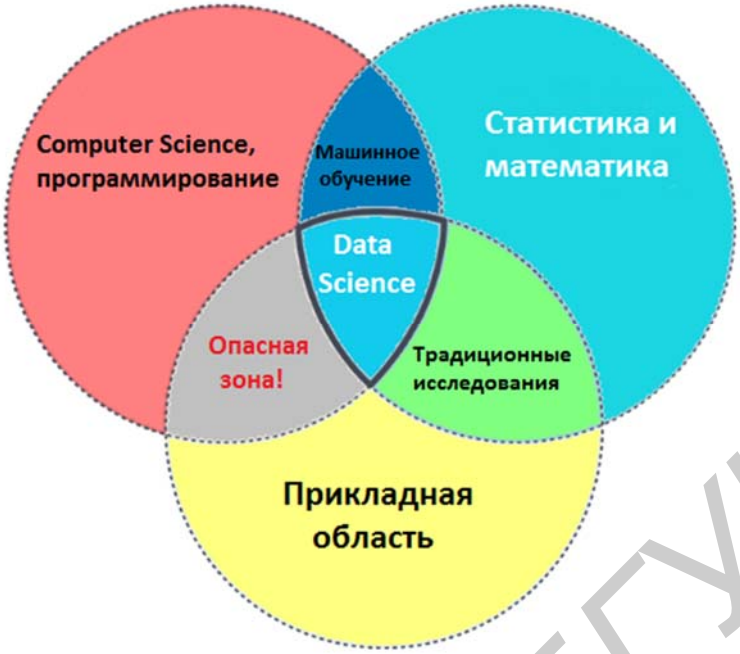

Рис. 1.

2 Мини-проекты должны быть актуальными как с практическом, так и с научно-исследовательской точки зрения.

3 По всем предлагаемым темам должно быть ненулевое количество свежих публикаций в реферируемых журналах. Способности к контролируемому самостоятельному изучению современной научной литературы, пониманию и реализации предлагаемым в ней алгоритмов, анализу результатов и дальнейшим улучшениям мы рассматривали как первичный навык специалиста в области анализа данных.

4 Общая предметная область с продуктами компании.

5 Нечеткая, достаточно общая поставка задач. С одной стороны мы не хотели ограничивать исследовательский потенциал, а с другой стороны хотели помочь сформировать навыки самостоятельного анализа задач, сбора требований и формулировки технического задания и разбивки больших задач на итерации.

3 Примеры мини-проектов. В рамках лаборатории в течении семестра мы предложили несколько задач по машинному обучению на неструктурированных или слабоструктированных данных. Таковыми выступали тексты и профили пользователей в социальных сетях. Требовалось научиться проводить различные предобработки текстов, разведочный анализ данных, самостоятельно выделять и отбирать признаки, строить модели, используя различные библиотеки машинного обучения и компьютерной лингвистики. Со всеми участниками лаборатории мы старались вести общение именно как со своими коллегами. Организация процесса в рамках мини-проектов велась итерациями, аналогичными agile-методологиям.

3.1 Лексическая нормализация текста в социальных сетях. Особенностью сообщений в социальных сетях являются большое количество опечаток,

Библиотека БГУИР

нарушение правил орфографии и пунктуации, специфичных сокращений. Это создает шум и затрудняет использование стандартных алгоритмов обработки текстов на естественных языках, извлечение из сообщений различных признаков для машинно-обученных моделей.

Одна из актуальных для сообщений в социальных сетях проблем — выделение из текста имен собственных. Требуется либо нормализация текстов, либо разработка адаптированных для работы в социальным контентом алгоритмов компьютерной лингвистики, ибо в противном случае традиционное использование готовых моделей приведет к тому, что почти любое слово, написанное с прописной буквы (или прописными буквами), будет ложно-положительным образом определяться именем собственным.

Мы предложили ребятам разобраться в методах, предложенных в конкурсе-воркшопе ACL 2015 Workshop on Noisy User-generated Text [1].

Ключевой особенностью этой задачи, как и большинства задач компьютерной лингвистики, является то, что во входных данных признаков нет вообще, на входе есть только текст. Это еще одна актуальная исследовательская задача из области Natural Language Processing.

Использованные библиотеки: scikit-learn (метрики качества), pycrfsuite (conditional random fields), fuzzy (soundex — алгоритм установления одинакового индекса для строк, имеющих схожее звучание в английском языке), enchant (библиотека проверки орфографии), pyxdameraulevenshtein.

Достигнутый результат: удалось добиться качества, аналогичного решению-победителя международного соревнования.

3.2 Online Reputation Management in Social Media. Был предложен набор задач на основе материалов конкурса в рамках конференции RepLab 2014 [2]:

1 Категоризация авторов сообщений в социальных сетях;

2 Бинарная классификация того, является ли автор influencer-ом в социальной сети;

3 Классификация контекстов упоминания брендов;

4 Категоризация сообщений (выделение тем).

Основными прикладными областями применениями таких задач являются различные системы интеллектуального мониторинга сообщений в социальных сетях и так называемый Influencer Marketing.

Главным отличием этой исследовательской задачи от реальной было ограничение размеченного в академических целях набора данных авто- и финансовой тематиками (использовали открытые данные). Обученные модели получались тематико-зависимыми, но позволяли судить об эффективности разработанных алгоритмов и целесообразности использования выбранных методов.

В данных задачах доступен для потенциального использования как сам текст сообщения, так и профиль автора, заполненный им самим, а также информация про количество ретвитов и т.п., однако большая часть информации по-прежнему является неструктурированной. Участникам лаборатории

Библиотека БГУИР

приходилось изучать материалы конференции, реализовывать лучшие из предложенных там алгоритмов и дополнять их собственными идеями.

В отличие от "синтетических" лабораторных заданий по машинному обучению нужно было мало заниматься настройкой классификаторов и много думать над тем, что может быть источником сигналов для классификаторов. Таковыми могли быть количества подписчиков, количество подписок, их соотношение, качество заполненности профиля, наличие ссылок на веб-сайты, стиль написания сообщений, наличие сообщений, похожих на спам, использовать как признак оценку сентимента сообщения (а для этого сентимент тоже нужно было и предсказать), наборы каких-то ключевых слов.

Такие задачи нельзя легко загрузить в какой-либо готовый программный пакет типа Weka, RapidMiner, Orange или т.п., сформировать в графическом интерфейсе pipeline трасформаций данных, нажать кнопку и получить на выходе какую-то модель. Как и трудно написать код на языке R При решении задач приходилось планировать

Использованные библиотеки: twitter-text-python, twokenize (токенизация сообщений в Twitter), nltk (стемминг), spacy (предсказание частей речи), gensim (word2vec), scikit-learn (TF-IDF-векторизация текстов, логистическая регрессия, случайный лес, метрики качества, кросс-валидиция), xgboost (эффективная реализация градиентного бустинга над решающими деревьями), pandas, numpy, ujson.

Достигнутый результат: также удалось добиться качества, аналогичного решению-победителя международного соревнования. Дополнительная апробация полученной модели на корпусе twitter-сообщений в Гонконге подтвердила потенциальную возможность внедрения подобных решений в бизнес-решения.

Среди других мини-проектов, которые мы предлагали лаборатории были:

– предсказание дохода человека по его аккаунту в социальной сети (на данных из [3], формально результат смогли улучшить, однако среднее абсолютное отклонение в отличие от [3] мы не считаем адекватной метрикой);

– предсказание заработной платы по описаниям вакансий на русском языке;

– предсказание популярности контента.

4 Заключение. Мы провели большое количество собеседований, почти все кандидаты имели желание работать именно в области анализа данных, все обладали какими-то теоретическими знаниями разного уровня, полученными в ВУЗе или в рамках самообразования. Однако уровень академической подготовки был, все-таки, ниже требуемого сегодня IT-индустрии. Была заметна высокая фрагментация знаний. В Беларуси подготовка специалистов по прикладному анализу данных ведется в основном в рамках второй ступени высшего образования (магистратуры) с углубленной подготовкой специалиста по специальностям 1-31 81 12 "Прикладной компьютерный анализ данных" и 1-31 81 09 "Алгоритмы и системы обработки больших объемов информации". Однако

Библиотека БГУИР

остается неудовлетворенный спрос на знания по анализу данных на этапе I ступени высшего образования, которые одинаково полезны и в научно-исследовательском, и в прикладном плане, когда студенты стоят перед выбором своей основной специализации.

Мы видим большую необходимость и нереализованный потенциал в формировании IT-сообщества специалистов в области анализа данных и стимулирования привлечения талантливой молодежи.

Мы высоко оцениваем достигнутыми в рамках лаборатории результатами и очень благодарны всем ее участникам, без каждодневных усилий которых этот результат был бы невозможен.

Литература

[1]. ACL 2015 Workshop on Noisy User-generated Text. Lexical Normalisation for English Tweets. [Электронный ресурс]. — Режим доступа: https://noisy-text.github.io/norm-shared-task.html. — Дата доступа: 01.06.2016.

[2]. RepLab 2013. Track for Online Reputation Management. [Электронный ресурс]. — Режим доступа: http://nlp.uned.es/replab2014/. — Дата доступа: 01.06.2016.

[3]. D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, N. Aletras. Studying User Income through Language, Behaviour and Affect in Social Media. [Электронный ресурс]. — Режим доступа: http://dx.doi.org/10.1371/journal.pone.0138717. — Дата доступа: 01.06.2016.

Библиотека БГУИР