# Изучаем платформу расширенной аналитики: Часть 3. Анализ неструктурированного текста с использованием шаблонов

## Изучаем шаблоны проектирования для анализа неструктурированных текстов и связанных задач

Крис Лоза
ИТ-специалист
IBM

Арвинд Сатхи
главный системный архитектор
IBM

Мэтью Томас
ведущий архитектор
IBM

26.09.2014

Из этой статьи вы узнаете, как использовать шаблоны для анализа неструктурированного текста в контексте больших данных. Поскольку для решения проблем бизнеса в этой области, как правило, нужно решать множество задач, авторы описывают простые шаблоны для проектирования решений, использующие информацию, содержащуюся в неструктурированных документах.

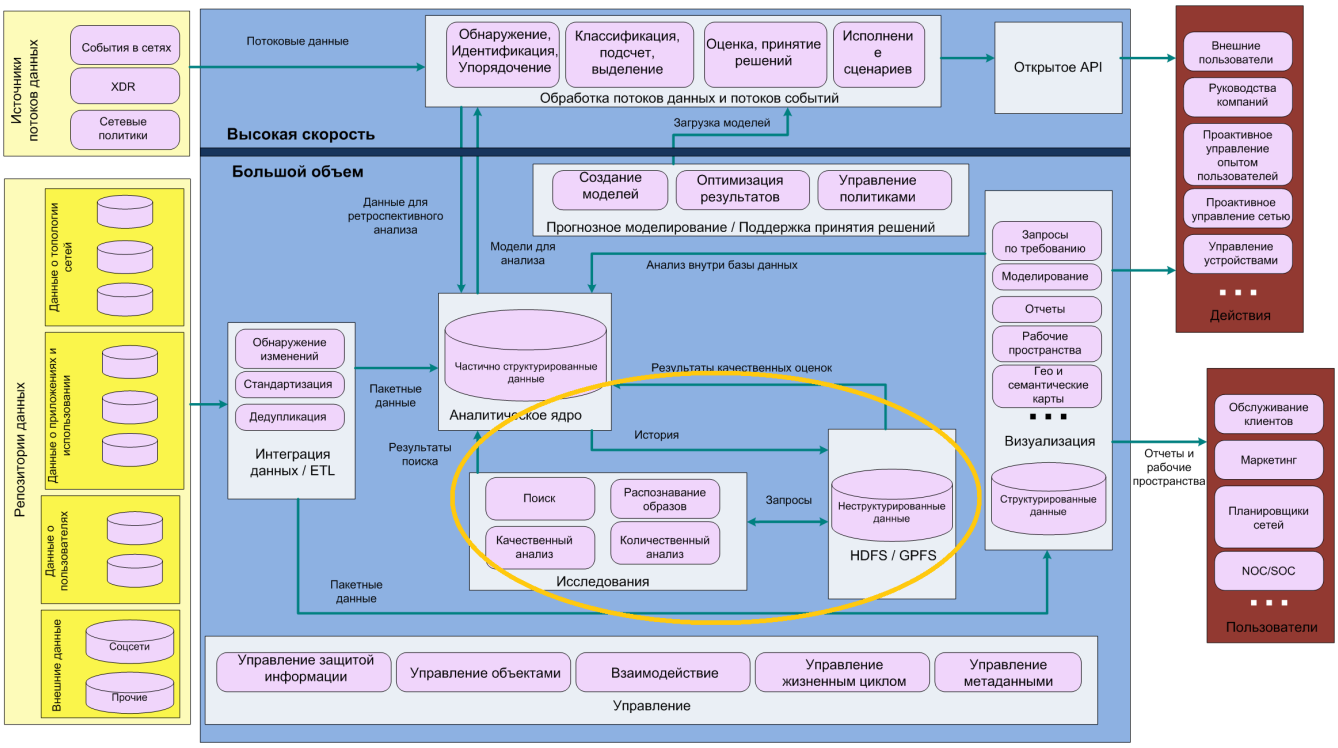
Больше статей из этой серии

## Введение

Предыдущие статьи этой серии содержали описание платформы расширенной аналитики и некоторых ключевых примеров, которые можно реализовать, используя эту платформу. На примере одного потока информации было проиллюстрировано, как различные компоненты платформы работают вместе. Ключевым аспектом любой аналитической платформы является ее способность анализировать неструктурированные данные. До недавнего времени решения позволяли работать только со структурированными данными, в то время как большая часть данных является неструктурированной. Изучение неструктурированных данных имеет ряд специфических проблем, связанных с их природой. В этой статье мы

обсудим задачи, которые входят в анализ текстов, шаблоны проектирования, используемые этими задачами, и продукты IBM, которые вы можете использовать для анализа, обработки и извлечения знаний из неструктурированного текста.

## Рисунок 1. Платформа расширенной аналитики. Неструктурированные данные




На рисунке 1 показана высокоуровневая архитектура взаимодействия компонентов в рамках платформы расширенной аналитики. В этом случае платформа получает на вход неструктурированные текстовые данные из различных источников как в виде непрерывного потока, так и в обычном пакетном режиме. Потоковые данные могут включать информацию о местоположении, финансовых операциях и другую потоковую информацию, специфичную для рассматриваемой области. Статические данные могут включать внешние хранилища информации из социальных сетей, информацию о клиентах, информацию об использовании приложений, другую отраслевую информацию. В этой статье мы сосредоточимся на поиске, количественном анализе, распознавании, качественном анализе, текстовой аналитике, неструктурированных хранилищах и компонентах аналитической платформы, приведенных на рисунке 1. (За подробной информацией об архитектуре и компонентах платформы расширенной аналитики рекомендуем обратиться к первой статье серии Платформа расширенной аналитики. Часть 1: архитектура и компоненты.)

## Цель

В этой статье мы обсуждаем шаблоны, предназначенные для множественного использования, которые можно применять в контексте задач, включающих обработку текстов на естественных языках (Natural Language Processing - NLP). В частности, мы хотим поделиться с вами информацией о том, как использовать различные инструменты компании IBM для анализа, обработки и извлечении знаний из неструктурированных текстов. В центре внимания будет рассмотрение общих инфраструктур, включающих множество задач по анализу текстов, а также реализация и применение этих инфраструктур для решения сложных проблем. В связи с тем, что для решения каждой из задач существует множество подходов, мы сконцентрируем внимание на подмножестве инструментов, которые мы использовали для реализации различных решений в контексте выбранных отраслей и больших данных.

## Определения

Ключевые определения, которые приведены ниже, помогут вам лучше понять оставшуюся часть статьи.

### Социальные медиа

Социальные медиа представляют собой совокупность взаимодействий и коммуникаций людей между собой с помощью Интернет-сообществ, таких как Facebook, Твиттер и Orkut. Социальные медиа являются одним из самых заметных новшеств в нашем обществе за последние годы и, с точки зрения данных, представляют собой наиболее доступный и широко используемый текстовый ресурс. В связи с тем, что анализ информации из социальных медиа улучшает наше представление о тенденциях, мнениях и предпочтениях, такой анализа является сегодня насущной задачей для большинства компаний.

### Обработка сообщений на естественных языках

Обработка сообщений на естественных языках – это область искусственного интеллекта, которая изучает человеческий язык и различные подходы к анализу, систематической обработке и пониманию языка. Естественный язык имеет неочевидную внутреннюю структуру, и частью задачи автоматической обработки является выявление этой структуры.

Процесс анализа может сильно различаться от языка к языку, поскольку язык сильно зависит от мировоззрения и общей культуры его носителей. Некоторые языки имеют общие корни и, развивались, влияют друг на друга (например, испанский и португальский). В то же время другие могут быть независимы друг от друга (например, арабский, русский и японский).

Текстовой аналитикой и углубленным анализом текстов обычно называют задачи, включающие анализ неструктурированных текстов и извлечение или формирование структурированных данных, а также достижение определённого уровня интерпретации этих данных. Операции, которые могут быть необходимы при анализе неструктурированных текстов, описаны ниже.

### Неструктурированные данные

Неструктурированными называют данные, которые не имеют описанной внутренней структуры или определения, соответствующего задаче, которую предполагается решать. Когда мы говорим о структуре данных, мы имеем в виду способ их организации, представленный в метаданных, которые сопровождают данные. Примером может служить определение столбцов таблицы в реляционной базе данных. В таких случаях добавление неструктурированных данных в аналитику больших данных может начинаться с операции упорядочивания данных в соответствии с их неявной структурой. В большинстве случаев такая обработка выполняется для того, чтобы иметь возможность агрегировать данные, формировать отчёты и производить какие-либо действия на основе информации, содержащейся в неструктурированных данных.

Массив данных может рассматриваться как структурированный или неструктурированный в зависимости от того, для решения какой задачи вы его собираетесь использовать. В сыром виде неструктурированный текст выглядит для аналитической системы как большой массив символов, текстов, чисел, знаков с определённой степенью упорядоченности. Эти данные требуют ряда преобразований, чтобы с ними могла работать аналитическая система.

Таким образом, термин "неструктурированные" можно интерпретировать как уровень организации данных по отношению к решаемой задаче. После того как данные упорядочены в структуру, пригодную для анализа, эти данные могут рассматриваться как структурированные.

Например, бинарный файл, содержащий изображение, может считаться структурированными данными при его визуализации программным обеспечением для работы с цифровыми изображениями. В то же самое время этот же файл может рассматриваться как неструктурированные данные при решении задачи распознавания контура изображения.

Зачастую бинарные файлы, такие как звуковые файлы или изображения в PDF-формате, требуют предварительной обработки для извлечения текста в формат, который далее может быть обработан с помощью подходов, которые описаны ниже.

## Простейшие задачи текстовой аналитики

Для удобства мы ввели простую классификацию задач на основе их относительной сложности. Эта классификация является произвольной в смысле того, что сложность реализации задач зависит от множества факторов, включая уровень точности, ограничения реализации, язык, с которым мы работаем, тип входных данных. В нашем случае мы рассматриваем контекст применения в типовом бизнес-сценарии и используем в качестве целевого языка английский.

На рисунке 2 изображен типовой пример преобразования текста из простого страхового заявления. Разбираем простой текст: Джон вывихнул лодыжку на лестнице. Вначале извлекаются части речи (существительное, глагол, наречие, дополнение). Затем факты

(персона, травма, часть тела, расположение) и концепции (заключение: травма мягких тканей) связываются друг с другом.

## Рисунок 2. Пример преобразования текста


Некоторые из ключевых элементарных задач, которые требуются для текстовой аналитики:

## Нормализация

Нормализация текста – это низкоуровневая задача по преобразованию текста в простую каноническую форму. Поскольку разные языки имеют разные машинные представления, нормализация текстов также определяет кодировку, формат и кодовую страницу. Эту задачу можно решать различными способами в зависимости от языка и инструментов, с которыми планируете работать.

Нормализация бывает необходима часто в силу характера социальных медиа. Платформа должна поддерживать множество языков и множество кодировок, поскольку она преобразует входную информацию к единому формату, который будет использоваться другими задачами более высокого уровня. Этот шаблон требует высокой скорости выполнения операций, объем которых пропорционален объёму входных данных.

## Идентификация языка

Первым шагом при анализе любого документа является определение содержания документа и языка, на котором написан документ. Идентификация языка – это процесс определения исходного языка любого документа. Работая с реальными данными, можно легко увидеть, что большинство документов, особенно в социальных медиа, зачастую написаны более чем на одном языке. Это обстоятельство усложняет задачу определения языка, на котором написан документ. Задача определения языка документа имеет решающее значение, поскольку обрабатывать и анализировать можно только документы с идентифицированным языком.

## Удаление служебных слов

Задача удаления служебных слов состоит в идентификации слов, которые не добавляют информации к семантическому контенту данной задачи и не убавляют её. Например, индексы поиска по ключевым словам не индексируют стандартные артикли английского языка, такие как "the".

Концепция служебных слов является нечеткой, и слово, которое в одном контексте может рассматриваться как служебное, в другом контексте может им не быть. Более того, служебные слова обычно имеют своё назначение в большинстве коммуникаций. В то время

как их удаление помогает подходам на основе "мешка слов", они могут быть полезны или даже необходимы при реализации других подходов.

### Разбиение на лексемы

Разбиение на лексемы – это выделение различных составляющих из заданного текста. В общем случае эта функция выделяет числа, знаки пунктуации, слова. В некоторых языках эта функция также включает добавление разделителя между словами в тех случаях, когда в языке такой разделитель отсутствует. Разбиение на лексемы является дополнительным шагом для упрощения задачи автоматической обработки письменного языка. Хотя большинство инструментов для разбиения на лексемы работают по детерминированным алгоритмам, они могут использовать эвристики, зависящие от языка. В некоторых языках, разбиение на лексемы может представлять сложную задачу, требующую отдельного решения.

### Выделение частей речи

Выделение частей речи представляет собой задачу привязывания признака "часть речи" каждому слову в предложении. Часто эта задача усложняется природой языка: для некоторых слов часть речи определяется контекстом, для некоторых – неоднозначна.

Выделение частей речи является обязательным условием для перехода к другим более высокоуровневым задачам, поскольку идентифицированные части речи являются ключом к повышению точности других видов анализа текста. Например, предварительное выделение частей речи помогает лучше определить эмоции, описываемые текстом, поскольку в эмоционально окрашенном тексте слова чаще используются в качестве прилагательных, чем существительных.

### Извлечение признаков

Извлечение признаков в контексте распознавания образов представляет собой задачу, направленную на поиск, извлечение и классификацию входной информации в соответствии с предопределенными извлекаемыми классами. Например, номера телефонов в США можно распознать по набору из трёх цифр в круглых скобках, трёх дополнительных цифр, тире и заключительных четырёх цифр. Простейшим способом извлечения признаков является использование регулярных выражений.

### Распознавание именованных сущностей

Целью распознавания именованных сущностей является идентификация специфических сущностей в тексте. Предопределённые категории или группы элементов имеют специфические семантические связи, соответствующие анализируемой теме. Эта задача предполагает некоторый предварительный анализ текста. Например, разметка частей речи помогает устранить неоднозначность в случае, когда некоторая лексема может иметь несколько разных значений.

## Более сложные задачи анализа текстов

Для большинства современных практических применений требуется более глубокий уровень анализа текстов, включающий множество шагов с преобразованиями, машинным

обучением, статистическим анализом. Такие подходы зачастую требуют учебных данных в случае использования подходов с обучением или доступа к общей базе знаний для подходов без обучения.

## Анализ настроений

Для большинства современных бизнес–применений требуется более глубокий семантический анализ естественного языка, чтобы получить большую отдачу от анализа социальных медиа. Одним из первых представленных применений было определение настроения.

Анализ настроений в последние годы привлекает много внимания по мере того, как все больше компаний анализируют коммуникации в социальных сетях, чтобы получить информацию о предпочтениях пользователей. Анализ настроений является очень тонкой задачей в том смысле, что подход, работающий для одной тематики, может не сработать для других. В ряд продуктов IBM включены алгоритмы анализа настроений общего назначения, с различными возможностями тонкой настройки и ввода дополнительных словарей в систему.

Анализ настроений является задачей высокого уровня, которая востребована во множестве бизнес–сценариев. Большинство задач сегодня использует агрегирование настроений по отношению к различным темам или продуктам, но становится все важнее иметь привязку индивидуальных мнений к конкретным тематикам.

## Извлечение информации

Извлечение информации представляет собой задачу поиска информации в коллекции документов. Первоначально такие задачи были связаны с поиском информации в Интернете, целью которых было проиндексировать и извлечь информацию в коллекции документов.

Запрос информации может быть представлен как вопрос, как шаблон поиска или во многих других формах. Результатом выполнения запроса могут быть: прямой ответ, коллекция ссылок, список людей, которые могут знать ответ или даже уточняющий вопрос.

В этой области различают ряд задач в зависимости от подхода и конкретных характеристик вопроса и ответа. Например, если запрос сформулирован в форме вопроса, а ожидаемый ответ является единственно правильным, мы говорим, что решается задача "Ответы на вопросы" (Question Answering).

## Автоматическое аннотирование текстов

Автоматическое аннотирование – это процесс сжатия документа в выходной документ, который называется аннотацией и содержит наиболее важные тезисы исходного документа.

## Классификация текстов

Классификация текстов представляет собой задачу по связыванию меток с документами. Этот общий подход может быть использован для многих целей. В последние несколько лет

все большее внимание уделяется социальным сетям и микроблогам. Более важно то, что классификация текстов может быть использована для автоматической группировки сходных мнений, относящихся к разным сегментам пользователей .

Другим примером классификации текстов является фильтрация спама. В контексте данных из социальных сетей фильтрация спама определяется как удаление сообщений, созданных скриптами или автоматическими ботами. Изучение спама в социальных сетях приобрело большую значимость в последние годы, так как спам негативно влияет на большинство применений аналитики к данным социальных сетей.

## Извлечение связей

Создание полного портрета клиента является сегодня предметом интереса большинства компаний. Знание о ваших клиентах критически важно для предоставления наилучших сервисов и также может быть важным для снижения рисков ведения бизнеса с конкретными клиентами.

Сервисы социальных сетей предоставляют доступ не ко всем индивидуальным характеристикам или свойствам каждого человека, но анализ взаимодействий между ними может дать значительно больше информации. Например, анализируя интерактивные сообщения в социальной сети (сообщения от пользователей, которые комментируют или отвечают другим пользователям), вы можете получить информацию о сетевых связях между людьми. Анализ содержания этих взаимодействий может определить природу связей между людьми.

Эта задача может быть разбита на несколько подзадач. В самом начале вам необходимо создать профили сущностей, которые необходимо проанализировать. Эти сущности имеют разные измерения или свойства, на основании анализа которых анализируются их внутренние связи. Например, задача анализа связей между людьми отличается от задачи анализа связей между людьми и компаниями.

Вы также можете определить, какие связи вы хотите извлечь. Например, в случае анализа мошенничества, вы, вероятно, захотите проанализировать семейные связи. При анализе потери клиентов представление о связях внутри социальной сети важно для определения влиятельности мнений пользователей сети.

## Ответы на вопросы

Задача ответа на вопросы является подзадачей извлечения информации, когда ответ представляется не как коллекция документов, а как единая совокупность данных, содержащая прямой ответ на поставленный вопрос.

Эта задача разбивается на две основные подзадачи. Начальная часть задачи состоит в понимании вопроса, сформулированного на естественном языке. Частью первой подзадачи является определение типа данных, которому соответствует ответ на вопрос, и тема, к которой относится вопрос. Вторая часть задачи состоит в том, чтобы найти информацию, которая непосредственно отвечает на поставленный вопрос.

# Шаблоны проектирования для анализа текстов

Теперь рассмотрим применение шаблонов для анализа неструктурированных данных.

## Шаблоны агрегирования

Шаблоны агрегирования данных могут быть описаны как "чертежи" для создания решений, требующих агрегации индексов, статистик, объединения индивидуальных вкладов и трендов на основе данных, содержащихся в текстовых входных данных. Используемое обычно для описательной аналитики, агрегирование помогает вам получить представление о том, как эти данные могут влиять на существующие продукты, услуги, восприятие имиджа, шаблоны покупательского поведения и т.д. С практической точки зрения задача агрегации данных имеет дело с относительно большим количеством данных, которые необходимо обрабатывать с высокой скоростью. Оба признака – большой объём и высокая скорость – характеризуют эту задачу как задачу работы с большими данными.

### Входные данные для агрегирования

Входные данные для задачи агрегирования могут включать:

- Социальные данные: задача анализа текстов социальных сетей в этом случае включает информацию социальных медиа, источниками которых могут быть сайты микроблогов, группы в социальных сетях, Интернет-конференции и другие площадки коммуникаций в сети.
- Записи журналов, сгенерированные компьютерами: такие записи генерируются и используются платформой для различных целей, включая оптимизацию сетевого трафика, оценку обслуживания клиентов, идентификацию проблем и т.д.
- Демографические данные: традиционные демографические данные.

### Общие шаги по агрегированию данных

Общепринятыми шагами внутри этого шаблона являются:

a. Нормализация данных
b. Разбиение на лексемы
c. Идентификация языка
d. Классификация текстов для устранения спама
e. Классификация текстов для автоматического отнесения к некоторой категории содержимого
f. Анализ впечатлений или извлечение мнений.

Некоторые из этих шагов, такие как нормализация данных и идентификация языка, актуальны и для любой другой задачи, которую вы захотите выполнить. Другие зависят от задачи, которую вам необходимо решить.

Вы можете выполнить эти задачи внутри одного простого сценария. Предположим, например, что мы хотим проанализировать мнения в сети относительно некоторой компании из индустрии медиа и развлечений.

### Пример агрегирования

Выполнение этого анализа включает следующие шаги:

1. Инфраструктура: необходимые программные средства для решения задачи:
   - BigInsights® 2.1
   - shell (Bash)
   - curl
2. Данные: BigInsights поставляется с некоторым количеством тестовых данных, которые можно использовать для тестирования работы платформы. Программа называется “Загрузка данных” и входит в состав стандартных программ, поставляемых вместе с BigInsights 2.1. Другой подход состоит в том, чтобы загрузить данные непосредственно из Твиттера с помощью следующих шагов:
   a. Ознакомьтесь с политикой конфиденциальности Твиттера.
   b. Создайте учетную запись разработчика в Твиттере.
   c. Создайте приложение и запросите OAuth-подпись для этого приложения.
   d. Скопируйте и вставьте команду `curl` со страницы OAuth и выполните ее в командной оболочке. Команда `OAuth` содержит заголовок авторизации для вашего приложения. Перенаправьте выход в файл и остановите процесс сбора информации, когда посчитаете, что получили достаточное для своих целей количество данных.
   e. Удалите последнюю строку файла. Обычно после отключения от сервиса последняя строка оказывается неполной, и ее стоит удалить.
   f. Загрузите полученный файл в BigInsights командой `hadoop fs –copyFromLocal`
   g. В BigInsights запустите сеанс конфигуратора управления брендами области в медиа и развлечений (**Brand Management Media and Entertainment Configuration**). Выберите опции конфигурации по умолчанию, отметив специально “Данные в пакетах” (“Packaged Files”). Используйте имя сценария в качестве идентификатора для вашей задачи.
   h. Запустите приложение “Локальный анализ для медиа и развлечений” (**Media and Entertainment Local Analysis**) для файла полученного из Твиттера, используя имя сценария и каталог, в который вы загрузили файл для локального анализа, как показано на рисунке 3.

**Рисунок 3. Запуск приложения “Локальный анализ для медиа и развлечений”**

i. Запустите приложение “Глобальный анализ для медиа и развлечений” (**Media and Entertainment Global Analysis**), которое создает социальные профили на основе ваших данных. Используйте то же самое имя сценария и направьте приложение на каталог или каталоги, содержащие результаты локального анализа (разделив их запятыми).
j. Проанализируйте результаты. В вашей HDFS-системе перейдите в каталог **accelerators / SDA / BrandManagement / ME / profiles** и экспортируйте результаты в BigSheets.
k. Используя BigSheets, вы можете теперь графически отобразить социальное настроение относительно конкретных брендов из области масс-медиа и развлечений.

## Пример: Audience Insights - анализ социальных данных из Твиттера

Решение Audience Insights разработано, чтобы ответить на вопрос о том, как различные аудитории относятся к различным телепередачам и как интегрировать аудитории разных каналов. Источниками данных для этого примера были выбраны социальные сети, отраслевые базы данных и базы данных CRM-систем.

В этом примере мы сопоставляем множество источников информации, чтобы понять, как доступ к цифровому контенту и социальные сети соотносятся с используемыми в настоящее время линейными метриками просмотров. Мы также анализируем вовлечение сегмента зрительской аудитории и различных телепередач в сети. В этом случае мы создаём хранилище данных с информацией о зрителях, их предпочтениях, их социальном

влиянии и их вовлеченности в сетевые телепередачи. Особенно нас будут интересовать измерение их публичного мнения и действий в связи с телепередачами.

### Пример: Прогнозирование продаж на основе сообщений в Твиттере, намерений и настроений

Это решение основывается на анализе, выполняемом на основе информационной активности, намерений и настроений, связанных с некоторым продуктом. Это решение использует корреляцию информации о социальных взаимодействиях как показатель для оценки объёма будущих продаж. В частности, вы можете использовать этот пример для соотнесения текущих продаж, запусков новых продуктов, настроений и отдачи от мероприятия и маркетинговых компаний. Проблемы при внедрении этого примера состоят в необходимости создания индивидуальных профилей на основе больших данных, глубокого анализа неструктурированных текстов и диалогов и формирования адаптированной текстовой аналитики для изучения намерений и настроений.

## Шаблоны маркировки и профилирования

Шаблоны маркировки представляют собой схемы для решения задач, требующих маркировки или классификации поступающей информации.

### Входная информация для маркировки и профилирования

- Социальная информация: в этом варианте для создания и маркирования индивидуальных сущностей или профилей используются социальные данные.
- Данные CRM-систем и персональные данные: хранилища, которые обычно содержат персональную информацию о клиентах.

### Маркировка и профилирование, как правило включает следующие шаги:

1. Нормализацию данных
2. Разбиение на лексемы
3. Идентификацию языка
4. Классификацию текстов для удаления спама
5. Классификацию текстов для автоматической классификации контента
6. Распознавание именованных сущностей
7. Извлечение связей

### Пример маркировки и профилирования

Для выполнения такого анализа мы использовали следующие шаги для обработки данных социальных сетей с помощью платформы расширенной аналитики.

1. Инфраструктура: необходимы следующие программные компоненты:
    - BigInsights 2.1
    - shell (Bash)
    - curl

- SPSS® Modeler или другой статистический пакет

2. Данные: BigInsights поставляется с некоторым количеством тестовых данных, которые можно использовать для тестирования работы платформы. Программа называется “Загрузка данных” и входит в состав стандартных программ, поставляемых вместе с BigInsights 2.1. Другой подход состоит в том, чтобы загрузить данные непосредственно из Твиттера с помощью следующих шагов:
   a. Ознакомьтесь с политикой конфиденциальности Твиттера и всех участвующих сторонних лиц.
   b. Получите сторонний доступ к социальной сети. В предыдущем примере мы описали, как использовать демонстрационный канал Твиттера. В этом сценарии мы воспользуемся платной подпиской на эти данные.
   c. В BigInsights выберите соответствующего стороннего провайдера социальной информации и введите свои учетные данные и URL для получения доступа к данным. Эта общая процедура работает для большинства сторонних провайдеров данных социальных сетей, поддерживаемых в BigInsights
   d. В BigInsights запустите **Brand Management Retail Configuration**. Выберите конфигурацию по умолчанию, отметив опцию Packaged Files. Используйте имя сценария в качестве идентификатора своей задачи.
   e. Запустите **Local and Global Analysis** для Твиттера, используя имя сценария, каталог, в который вы загрузили данные из Твиттера, и выходной каталог, как показано на рисунке 3. Вы можете настроить словари для анализа настроений и указать компании, которые вы хотите проанализировать, если их нет в конфигурации по умолчанию.
   f. После завершения глобального анализа вы найдёте созданные социальные профили в каталоге SMA hadoop.
   g. Изучите результаты. Если вы используете HDFS, воспользуйтесь профилями **ускорителей / SDA / BrandManagement / ME** и экспортируйте результаты в BigSheets.
   h. Вы можете использовать приложение BigInsights для экспорта результатов обработки в BigSheets в CSV-файл, базу данных, некоторые хранилища данных. Приложения могут быть не видны по умолчанию. Для их активации выберите пункт **Manage** в меню BigInsights Applications.
   i. После того, как вы сохраните профили в реляционной базе данных, вы можете работать с этими данными через SPSS Modeler.
   j. SPSS Modeler содержит разные модели для анализа рынка. В частности, для приведенного типа данных вы можете использовать модели кластеризации и анализа корзины.

## Пример: Платформа расширенной аналитики для целевого маркетинга

Платформа расширенной аналитики была детально описана в первой части этой серии. Однако текстовая аналитика, которая необходима в этом сценарии, не была рассмотрена с необходимой полнотой.

В этом примере вам необходимо понимать две основные точки, где используется обработка неструктурированных текстов:

- Данные о мобильности извлекаются из сетевых устройств, находящихся на базовых станциях сотовой связи. В четвертой статье этой серии мы опишем этот процесс более подробно. Общий принцип состоит в получении в реальном времени необработанных журнальных файлов от сетевых устройств и обработке их с помощью IBM InfoSphere® Streams для извлечения информации о долготе и широте. Используя эти данные, InfoSphere Streams преобразует координаты в геохэши, которые сохраняются в профилях мобильности клиентов. Преобразование текстов в InfoSphere Streams включает:
  a. Нормализацию данных
  b. Выделение признаков
- Информация социальных сетей обрабатывается и преобразуется в Платформе расширенной аналитики аналогично тому, как это происходило в примере с разметкой и профилированием.

### Пример: Формирование лучших предложений для предотвращения оттока клиентов

Прогнозирование оттока клиентов является одним из направлений исследований в области прогнозной аналитики для бизнеса в последние годы. Для прогнозирования склонности клиента отказаться от услуг провайдера необходимо множество параметров. В прошлом большая часть такой работы основывалась на профилировании исторических данных; анализ, предлагаемый в этом примере, работает аналогичным образом. Отличие от традиционных подходов к измерению вероятности ухода клиента состоит в использовании данных социальных сетей, анализе текстов диалогов со службами поддержки пользовательских сервисов и вычислении вероятности в реальном времени для большого объема объектов. Эта работа основывается на решениях, которые работают быстро, принимают множество параметров, сравнивают их с историческими значениями, отслеживают взаимодействия и оценивают количественно риски ухода клиента во время телефонного звонка.

## Программные продукты, используемые для анализа неструктурированных текстов

Этот краткий обзор описывает некоторые ключевые продукты для анализа неструктурированных текстов.

### IBM Social Media Analytics

IBM Social Media Analytics (SMA) – это коробочный продукт, использующий множество видов анализа для создания полного анализа обратной связи в социальных сетях для конкретной компании или предметной области. Это решение обеспечивает полный анализ, включающий в себя следующие внутренние задачи:

- Внутренние задачи, недоступные пользователю: нормализация, разметка, идентификация языка, разметка частей речи, удаление служебных слов.

- Внешние задачи: классификация текстов, анализ настроений, извлечение связей.

## SPSS Text Analytics

SPSS Text Analytics – это продукт, ориентированный на анализ текстовых результатов опросов. Основная функция этого продукта состоит в применении базовых функций анализа и разбора к неструктурированным текстам, получаемым в результате опросов, и формировании на их основе структурированной информации.

Решение включает следующие задачи: нормализация и распознавание именованных сущностей.

## BigInsights

IBM BigInsights – это платформа и система, использующая Hadoop и MapReduce и способная выполнять множество задач. В частности, BigInsights включает три группы программ для текстового анализа: Social Data Accelerator (SDA), Machine Data Accelerator (MDA) и Telecommunications Event Data Analytics (TEDA).

Внутренние задачи, недоступные пользователю, включают нормализацию, разметку, идентификацию языка, классификацию текстов для фильтрации спама, распознавание сущностей, интеграцию сущностей, анализ настроений, извлечение связей.

Эти задачи можно настраивать с помощью словарей для анализа настроений и списка сущностей для распознавания сущностей. Кроме того, разработаны специальные решения для извлечения отраслевых ключевых показателей для избранных отраслей.

BigInsights обрабатывает неструктурированный текст и может формировать структурированные данные. Эти результаты также можно использовать как исходные для других внешних обработчиков или для интеграции данных из нескольких внешних источников. Вы можете использовать эти результаты при решении таких задач, как классификация текстов, анализ настроений, извлечение связей и идентификация сущностей.

## IBM Content Analytics

IBM Content Analytics (ICA) - платформа, включающая множество инструментов для решения задач, обсуждавшхся в этой статье. В частности, этот продукт позволяет исполнять задачи анализа текстов низкого уровня индивидуально и предоставляет готовые сценарии для некоторых из них. В контексте текстовой аналитики ICA имеет решения для удаления служебных слов, разметки частей речи, классификации текстов и другие.

## SPSS Text Analytics for Surveys

SPSS Text Analytics for Surveys – это программная компонента, тесно интегрированная с другими продуктами SPSS. Она содержит базовые инструменты для выделения признаков, которые можно использовать для обработки и распознавания существенной информации, содержащейся в результатах опросов.

### Watson

Технология Watson представляет собой комбинацию множества технологий, предназначенных для реализации функциональности ответа на вопросы. На уровне ядра Watson представляет собой сложную систему ранжирования ответов на множество вопросов и систем извлечения информации, работающих параллельно и оптимизированных как по скорости, так и по точности.

Поскольку Watson использует множество технологий для формулирования ответов, большинство ее внутренних решений аналогичны ранее обсужденным, но снаружи доступна только подсистема, отвечающая на вопросы.

## Выводы

Шаблоны для анализа текстов, которые мы обсудили в этой статье, включают преобразования, маркировку, объединение и сжатие. Эти шаблоны могут быть детализированы таким образом, чтобы аналитик мог по–настоящему анализировать неструктурированные данные для получения результата, который невозможно получить с помощью структурированных данных. Несколько продуктов IBM поддерживают эти шаблоны, позволяя анализировать неструктурированные тексты в различных ситуациях, включая анализ оттока клиентов, анализ настроений клиентов, углубленное изучение клиентов.

Следующая статья посвящена другому типовому аналитическому шаблону, который называют шаблоном анализа местоположения. Сегодня большое количество геопространственной информации собирается провайдерами и доступно для других отраслей благодаря наличию доступа к GPS-данным. В части 4 мы расскажем, как платформа расширенной аналитики позволяет реализовать сложный анализ перемещений клиента, его нахождения с другими людьми, его вероятных местоположений. Такая информация может быть полезна для организации маркетинговых кампаний, включая краткосрочные, отправки локальных коммерческих предложений, предсказания ожидаемого потока клиентов и отслеживания лиц, причастных к противоправной деятельности.

# Об авторах

### Крис Лоза

Крис Лоза (Chris Loza) - ИТ-специалист с опытом разработки решений для текстовой аналитики и больших данных. Он работает аналитиком данных в Global Solution Center, заканчивая свою докторскую диссертацию в области автоматического семантического анализа для социальных сетей. Он имеет ряд статей и внутренних публикаций в IBM в области анализа социальных медиа, анализа настроений, создания интегрированных решений для анализа поведения клиентов.

---

### Арвинд Сатхи

Арвинд Сатхи (Arvind Sathi) - архитектор решений в области телекоммуникаций компании IBM для больших данных. Область его интересов состоит в разработке общих тенденций и дорожных карт для расширенной аналитики для основных клиентов IBM в области телекоммуникаций, СМИ, индустрии развлечений, энергетики и ЖКХ во всем мире. Он реализовал ряд стратегических контрактов со множеством клиентов в области телекоммуникаций.

---

### Мэтью Томас

Мэтью Томас (Mathews Thomas) работает ведущим архитектором в области телекоммуникаций в Глобальгном центре отраслевых решений IBM (IBM Global Industry Solution Center - GISC), который включает основной центр отраслевых решений IBM в Северной Америке в области телекоммуникаций, СМИ, индустрии развлечений, энергетики и ЖКХ. GISC также является центром аналитических решений IBM, где Мэтью является ведущим архитектором.