УДК 004.822

**ИНЖИНИРИНГ НОВЫХ ЗНАНИЙ ПРИ АНАЛИЗЕ ТЕКСТОВЫХ СООБЩЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ**

К.С. САМОЛЕТОВА

*Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский технологический университет»*

*119454, Москва, Россия*

Ценность знаний в связи с интенсивным ростом объема информации, генерируемой людьми в современном мире, повышается с течением времени. Очевидно, что подавляющая часть этих данных не имеет четкой структуры: аудио- и видеофайлы, изображения, тексты. При этом большинство средств, используемых аналитиками для анализа, оперируют только структурированной информацией, такой как время, даты, показатели в виде цифр и названия организаций. Так, количество неструктурированных сведений может достигать на предприятии 70%, а в социальных медиа это число равняется практически 100% [1]. Таким образом, неструктурированная информация – исключительный источник для инжиниринга знаний и составляет четыре пятых существующих данных в мире.

Особый интерес существует к онлайновым сервисам социальных сетей. В них люди в открытом доступе публикуют различные пользовательские сведения: факты биографии, дневники и заметки, переписки. Всё это открывает безграничные возможности для различных исследовательских задач, моделирования социальных, экономических, политических и других процессов от персонального до государственного уровня с целью разработки механизмов воздействия на эти процессы, а также создания инновационных аналитических и бизнес-приложений и сервисов [2]. Вместе с этим, работая с социальными данными необходимо учитывать, что качество пользовательского контента может быть крайне нестабильным, и принимать во внимание такие факторы как спам и фейковые аккаунты.

В общем виде обработка неструктурированной информации проводится в пять этапов, которые представлены на рисунке 1.

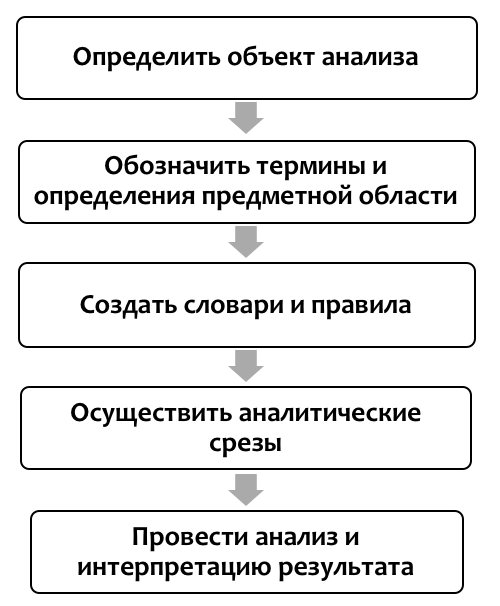


Рисунок 1 – Этапы обработки неструктурированной информации.

Первым этапом является выбор объекта анализа, в данной работе – это «Московский технологический университет». Для конкретизации рассматриваемой предметной области в качестве терминов и определений выступили названия структурных подразделений Университета, форм обучения, номера курсов, адреса организации, коды направлений подготовки и специальностей, набор на которые производиться в ВУЗе. Основываясь на этом третий шаг состоит в составлении словарей и правил.

В качестве инструмента для моделирования предмета исследования, осуществления аналитических срезов и перехода от них к анализу и интерпретации результатов был выбран IBM Watson Content Analytics. Данное решение позволяет представлять слова в нормализованной форме, создавать списки часто употребляемых словесных конструкций, а так же с высокой степенью корреляции [3]. Кроме того, среди функционала данного средства предусмотрена возможность добавления пользовательских коллекций словарей.

Проведенное исследование позволило выявить следующие характерные черты заданного Университета: популярные направления подготовки и специальности среди поступающих, актуальные тематики для учащихся различных годов обучения, наиболее важные вопросы абитуриентов и прочее.

Итак, результат работы позволяет сделать вывод, что одна из доминирующих тенденций развития социальных сетей состоит в понимании особенностей поведения людей на основе выявления новых знаний при анализе информации, которую они выкладывают в открытый доступ. Интерес представляет дальнейшая разработка метода мониторинга внутренних процессов Университета, основанного на анализе текстовых сообщений в сети Интернет.

Список использованных источников

1. Китова О. В., Дьяконова Л. П. Информационно-аналитические технологии и сервисы в экономике и образовании // Инжиниринг предприятий и управление знаниями. – 2016. – С. 30.

2. Коршунов А. и др. Анализ социальных сетей: методы и приложения // Труды Института системного программирования РАН. – 2014. – Т. 26. – №. 1.

3. Zhu W. D. J. et al. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content. – IBM Redbooks, 2014.