

Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра информатики

Зернов Алексей Викторович

Разработка системы автоматического анализа новостных публикаций на финансовом рынке

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Григорьев Д. А.

Рецензент:
позиция рецензента рецензент

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY
Faculty of Mathematics and Mechanics

Computer Science Department

Alexey Zernov

Development of the automatic analysis
system of a financial market's news
publications

Bachelor's Thesis

Scientific supervisor:
assistant professor Dmitry Grigoryev

Reviewer:
reviewer position reviewer

Saint-Petersburg
2017

Оглавление

Введение	4
1. Обзор существующих решений	5
1.1. IBM: Watson Developer Cloud	5
1.1.1. News Intelligence	5
1.1.2. Social Customer Care	5
1.1.3. News Explorer	6
1.1.4. Investment Advisor	6
1.2. Microsoft: Text Analytics	6
1.3. Медиалогия	6
2. Инструменты и методологии	8
2.1. Natural Language Toolkit	8
2.2. pymorphy2	8
2.3. Томита-парсер	9
2.4. Яндекс.Спеллер	9
2.5. OntosMiner	9
Список литературы	10

Введение

1. Обзор существующих решений

1.1. IBM: Watson Developer Cloud

Данный сервис представляет из себя набор различных API¹ для анализа текстовых, голосовых и визуальных данных. Среди их сравнительного большого количества я выделил несколько, о которых ниже будет рассказано подробнее.

1.1.1. News Intelligence

Одним из наиболее интересных сервисов для анализа новостных публикаций является приложение News Intelligence².

Приложение предлагает ввести название интересующей Вас компании, после чего предоставляются следующие результаты:

- Наиболее упоминаемые сущности (люди, темы, компании)
- Наиболее просматриваемые новостные публикации
- Анализ тональности новостных публикаций из десяти случайных источников
- Совместные упоминания и оценка их тональности

Важно отметить, что приложение News Intelligence является лишь примером использования данного инструмента, а не готовым продуктом.

1.1.2. Social Customer Care

Еще одним интересным примером использования Watson Developer Cloud является приложение Social Customer Care³. Оно осуществляет мониторинг социальных медиа, определяя потребности клиента или его запросы, а также автоматически отвечает в режиме реального времени.

¹Application Programming Interface

²<https://discovery-news-demo.mybluemix.net/>

³<https://social-customer-care.mybluemix.net/>

1.1.3. News Explorer

Следующий пример — News Explorer⁴. В данном приложении отображаются наиболее обсуждаемые запросы, наиболее часто встречающиеся совместные упоминания и список свежих новостей, разбитых по категориям.

В этом приложении также можно самостоятельно задать интересный запрос, после чего будет отображена визуализированная карта взаимосвязей разных сущностей и новостных публикаций в виде графа.

1.1.4. Investment Advisor

И последним рассматриваемым примером из данной группы является приложение Investment Advisor⁵. В данном приложении есть две группы людей: инвесторы и представители компаний. На основе анализа личностных особенностей людей по их постам, строятся определенные рекомендации по вложениям и наиболее подходящим для сотрудничества представителям компаний.

1.2. Microsoft: Text Analytics

Text Analytics⁶ позволяет провести анализ тональности текста, выделив ключевые слова. Microsoft предоставляет набор методов API для работы с данным сервисом. Подробнее о работе с ними будет написано ниже.

1.3. Медиалогия

Медиалогия⁷ — разработчик автоматической системы мониторинга и анализа СМИ в режиме реального времени.

Данная платформа предоставляет такие решения, как мониторинг СМИ компании (ее брендов, конкурентов и др.) и анализ СМИ и со-

⁴<http://news-explorer.mybluemix.net/>

⁵<http://investment-advisor.mybluemix.net/>

⁶<https://text-analytics-demo.azurewebsites.net/>

⁷<http://www.mlg.ru/>

общений с использованием уникальной технологии лингвистического анализа текстов.

Из предоставленных примеров отчетов⁸ на сайте можно увидеть, что сервис учитывает следующее:

- **Количество упоминаний.** Отслеживается динамика по кварталам и месяцам. Отслеживая динамику по дням, платформа определяет наиболее заметные информационные поводы, вызывающие более сильный всплеск упоминаний. Также учитывается цитируемость, совместные упоминания и распределение по тематическим рубрикам.
- **Качество упоминаний.** В рассмотренном отчете предоставлена информация о положительных и негативных сообщениях. Выделены пики и проанализирована связь со СМИ, которые способствуют больше благоприятному или отрицательному всплеску упоминаний.

Также в отчете были учтены распределения по уровням СМИ, по их географическому расположению, и прочее. Однако среди всего этого наиболее важным моментом является как раз анализ текста новости: является упоминание положительным или отрицательным, как это совместно упоминается с другими запросами и тому подобное.

⁸В качестве образца был взят аналитический отчет компании «Вымпелком» в СМИ за II квартал 2009 года

2. Инструменты и методологии

2.1. Natural Language Toolkit

NLTK⁹ является пакетом библиотек и программ для разработки программ на Python, работающих с естественным языком. Сопровождается обширной документацией, а также книгой¹⁰, объясняющей основные концепции проблем, для решения которых предназначен данный пакет. NLTK — свободное программное обеспечение, то есть доступное бесплатно.

Данный пакет подходит для таких областей как компьютерная лингвистика, эмпирическая лингвистика, когнитивистика, искусственный интеллект, информационный поиск и машинное обучение. NLTK используется преимущественно в качестве учебного пособия, индивидуального обучения или прототипирования и создания систем, ориентированных на научно-исследовательскую деятельность.

Изначально пакет предназначен для англоязычных текстов, но имеется возможность обучения классификаторов для остальных языков.

2.2. pymorphy2

Pymorphy2¹¹[1] написан на языке Python и имеет следующие возможности:

- Приведение слова к нормальной форме
- Ставить слово в нужную форму
- Возвращать грамматическую информацию о слове

Распространяется pymorphy2 под лицензией MIT¹², если используется в научной работе.

⁹<http://www.nltk.org>

¹⁰<http://www.nltk.org/book/>

¹¹<https://pymorphy2.readthedocs.io/en/latest/index.html>

¹²<https://opensource.org/licenses/MIT>

2.3. Томита-парсер

Томита-парсер¹³ способен извлекать структурированные данные из текстов на естественном языке. Как и почти во всех инструментах, рассматриваемых в данном разделе, Томита-парсер ориентирован преимущественно на русскоязычные тексты. В нем используются контекстно-свободные грамматики и словари ключевых слов. Код проекта¹⁴ находится в свободном доступе.

2.4. Яндекс.Спеллер

Яндекс.Спеллер¹⁵ выполняет задачу проверки орфографии в текстах на английском, русском и украинском языках. Для этого используется орфографический словарь. К тому же, предоставлен набор API методов для реализации данной проверки разработчиками сайтов или приложений.

2.5. OntosMiner

OntosMiner¹⁶ является решением компании Eventos¹⁷, занимающейся в большей степени разработкой продуктов в области лингвистического анализа текстовой информации, кластеризацией и классификацией информации. Конкретно OntosMiner является целой комплексной системой, дающей возможность распознавания связей между сущностями в текстах на естественном языке. Также, она позволяет определять общую тональность текста.

¹³<https://tech.yandex.ru/tomita/>

¹⁴<https://github.com/yandex/tomita-parser/>

¹⁵<https://tech.yandex.ru/speller/>

¹⁶<http://my-eventos.com/solution/ontosminer/>

¹⁷<http://my-eventos.com/solution/ontosminer/>

Список литературы

- [1] Korobov Mikhail. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts / Ed. by Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko et al.— Springer International Publishing, 2015.— Vol. 542 of Communications in Computer and Information Science.— P. 320–332.— URL: http://dx.doi.org/10.1007/978-3-319-26123-2_31.