

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:  
д. т. н., декан Мусаев А. А.

Санкт-Петербург  
2017

SAINT-PETERSBURG STATE UNIVERSITY  
Faculty of Mathematics and Mechanics

Computer Science Department

Zernov Alexey Viktorovich

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

Bachelor's Thesis

Scientific supervisor:  
Sc. C., associate professor Grigoryev D. A.

Reviewer:  
Sc. D., dean Musaev A. A.

Saint-Petersburg  
2017

# Оглавление

<b>Введение</b>	<b>5</b>
<b>1 Финансовый рынок</b>	<b>6</b>
1.1 Определение	6
1.2 Структура	7
1.3 Участники	8
<b>2 Интеллектуальный анализ текста</b>	<b>9</b>
2.1 Процесс интеллектуального анализа текста	10
2.1.1 Предварительная обработка текста	11
2.1.2 Преобразование текста	12
2.1.3 Поиск признаков	12
2.1.4 Методы анализа текста	12
2.1.5 Интерпретация и оценка	12
2.2 Области применения интеллектуального анализа текста	12
2.2.1 Извлечение информации	13
2.2.2 Информационный поиск	13
2.2.3 Обработка естественного языка	13
2.2.4 Интеллектуальный анализ данных	14
<b>3 Обзор существующих инструментов</b>	<b>15</b>
3.1 Natural Language Toolkit	15
3.2 Rymorphy2	15
3.3 Томиита-парсер	16
3.4 Яндекс.Спеллер	16
3.5 OntosMiner	16
<b>4 Программная часть</b>	<b>17</b>
4.1 Описание	17
4.2 Используемые инструменты	18
4.3 Структура программы	18
4.4 Результаты	19

<b>Заключение</b> . . . . .	<b>20</b>
<b>Список литературы</b> . . . . .	<b>21</b>

# Введение

# 1. Финансовый рынок

В данном разделе будет представлен краткий обзор основных терминов, связанных с самим финансовым рынком, его структурой и основными участниками. Более подробная информация может быть получена в [5].

## 1.1. Определение

В более общем виде **финансовый рынок** — совокупность экономических связей его участников, касающихся создания, поддержания и обращения капитала. Финансовый рынок является довольно абстрактным термином, и под ним часто подразумеваются более конкретные: рынок купонных и бескупонных облигаций, рынок акций (или фондовый рынок) или валютный рынок. Не смотря на выделение составляющих, каждая из них является частью единого механизма, в котором финансы перемещаются между каждым из конкретных рынков.

Каждый из финансовых рынков является рынком посредников между начальными владельцами финансов и их конечными пользователями. Если рынок основывается на финансах как на капитале, он называется фондовым рынком, и именно в этой роли выступает как составная часть всего финансового рынка.

В России финансовые рынки имеют следующие критерии, влияющие на их деятельность:

- Инвестиции в экономику страны
- Международные рынки, влияние тенденций глобализации
- Современные компьютерные технологии
- Уровень компьютерной и информационной развитости участников рынков

## 1.2. Структура

Финансовый рынок может быть:

- Первичным или вторичным
- Организованным или неорганизованным
- Биржевым или внебиржевым
- Традиционным или компьютеризированным
- Кассовым или срочным

**Первичный рынок** обеспечивает выход ценных бумаг в оборот, это своеобразное «производство» ценных бумаг. На **вторичном рынке** в обороте находятся уже выпущенные ранее ценные бумаги. Вторичный рынок представляет из себя совокупность всех операций с данными ценными бумагами, в результате которых они переходят от одних владельцев к другим.

**Организованный рынок** отличается от **неорганизованного рынка** тем, что в первом имеются единые для всех участников рынка правила, за соблюдением которых следят организаторы. Во неорганизованном рынке соблюдение единых правил для всех участников рынка не гарантируется.

**Биржевой рынок** — такой рынок, на котором в качестве инструмента торговли используется аукцион. Руководителем же является некоторый специалист, например, NYSE<sup>1</sup> или AMEX<sup>2</sup>. На **внебиржевых рынках** торги организуются при помощи электронных систем.

**Срочный рынок** чаще всего подразумевает отложенное исполнение сделки, в отличие от **кассового рынка**, когда сделки исполняются сразу. Обычно традиционные ценные бумаги (акции, облигации) идут в оборот на кассовом рынках, а контракты на производные инструменты рынка ценных бумаг — на срочных.

---

<sup>1</sup>New York Stock Exchange — Нью-Йоркская фондовая биржа

<sup>2</sup>American Stock Exchange - Американская фондовая биржа

### 1.3. Участники

**Участники** рынка ценных бумаг — это физические лица или компании, которые продают или приобретают ценные бумаги, обеспечивают их оборот или расчеты по ним.

Основными участниками рынка выступают **эмитенты**, выпускающие акции или облигации, с помощью которых привлекают финансирование, а также размещающие свободные на данный момент денежные средства. Эмитентами могут быть государство, субъекты государства или коммерческие предприятия. Целью эмитентов на первичном рынке является размещение запланированного транша по максимальной цене.

**Инвестор** — лицо, заинтересованное во вложении капитала в ценные бумаги. Их целью является как можно более выгодная покупка ценных бумаг максимально перспективных компаний.



## 2. Интеллектуальный анализ текста

В настоящее время можно заметить увеличение роли компьютеров в жизни каждого человека. Информация хранится преимущественно в цифровом виде, что значительно упрощает поиск или работу с ней. Но не смотря на это, многие данные все равно остаются довольно трудными для анализа, не смотря на оцифрованный вид, из-за чего можно подразделить их на следующие формы:

- Структурированные данные
- Частично структурированные данные
- Неструктурированные данные

Хорошим примером **структурированных данных** могут являться базы данных. **Частично структурированные данные** — это электронные письма, разнообразные файлы на языках разметок (HTML, XML и другие).

Если работа со структурированными или частично структурированными данными достаточно детерминированная, то **неструктурированные данные** представляют наибольший интерес в этом вопросе. Около 80% корпоративных данных находится именно в неструктурированном формате, в котором сложно проводить поиск или извлекать необходимую информацию. Для этого нужны специфические методы и алгоритмы обработки. И поскольку самая популярная форма хранения информации — это текст, интеллектуальный анализ текста (text mining) является более важным процессом, нежели интеллектуальный анализ данных (data mining).

Интеллектуальный анализ текста стоит на пересечении дисциплин, включая в себя: обработку web-данных, информационный поиск, компьютерную лингвистику и обработку естественного языка.

## 2.1. Процесс интеллектуального анализа текста

Концепция интеллектуального анализа текста представлена в [4]. В интеллектуальном анализе текста можно выделить два основных этапа (Рис. 1):

- Фильтрация текста
- Извлечение знаний

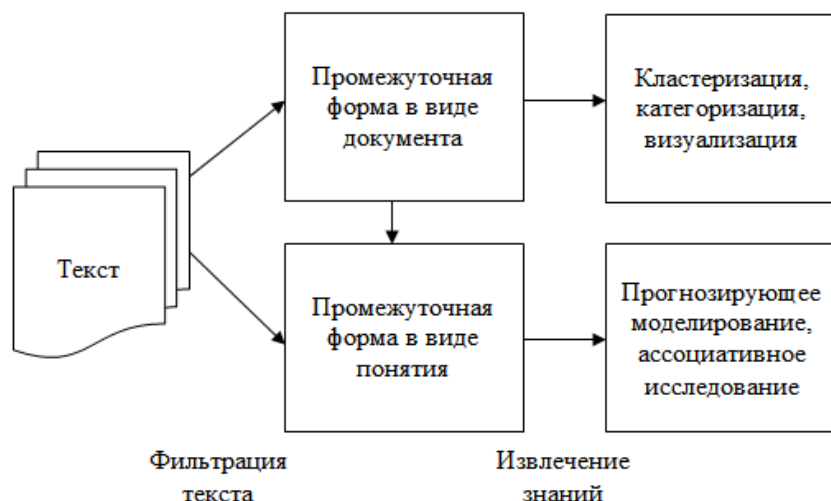


Рис. 1: Общий процесс интеллектуального анализа текста

**Фильтрация** (или очистка) преобразует исходный текстовый документ в некоторое промежуточное представление. **Извлечение знаний**, в свою очередь, получает знания или некоторые шаблоны из уже промежуточного представления. Промежуточное представление может быть как структурированным, так и частично структурированным. Промежуточное представление может быть как новым текстовым документом, так понятием, в котором составляющие являются данными или наборами данных из какой-либо предметной области.

Анализ промежуточного представления документов выдает образцы и связи между всеми документами. Примеры задач: *кластеризация, визуализация и категоризация документов*.

Анализ промежуточного представления понятий выдает образцы и связи между объектами или другими понятиями. Примеры задач: *прогнозирующее моделирование и ассоциативное исследование*.

Промежуточное представление в виде документа может быть преобразовано в промежуточное представление в виде понятия путем выделения релевантной информации, которая относится к необходимым объектам из какой-либо предметной области. Отсюда вытекает то, что промежуточное представление чаще не зависит от конкретной предметной области. К примеру, новостные потоки при фильтрации текста преобразуются в промежуточные представления в виде документов, соответствующим определенным статьям. Затем, в зависимости от поставленных задач визуализации или навигации, каждый документ (статья) проходит обработку знаний. Для извлечения же знаний в определенной предметной области промежуточное представление в виде документа может быть преобразовано в промежуточное представление в виде понятия в соответствии с необходимыми требованиями. К примеру, можно извлечь информацию, касающуюся определенного товара или услуги из промежуточного представления в виде документа и сформировать базу данных товаров или услуг для предоставления знаний о них.

Ниже будут рассмотрены шаги, выполняемые при интеллектуальном анализе текста.

#### **2.1.1. Предварительная обработка текста**

Предварительная обработка включает в себя:

1. Токенизацию
2. Удаление «стоп-слов»
3. Определение происхождения слов

**Токенизация.** Сначала текст разделяется на отдельные слова, освобождаясь от пробелов и знаков препинания.

**Удаление «стоп-слов».** На этом этапе происходит избавление от «ненужных» конструкций текста. Это могут быть HTML или XML теги, предлоги, артикли и прочее.

**Определение происхождения слов** представляет из себя выявление корней определенных слов. Преимущественно встречаются два типа происхождения: флективный и деривационный.

### **2.1.2. Преобразование текста**

Текстовый документ состоит из слов и информации об их происхождении. Два основных подхода представления документа: «мешок слов» («bag-of-words») и векторные пространства слов.

### **2.1.3. Поиск признаков**

Под признаками можно понимать переменные. То есть в результате этого шага отбирается подмножество наиболее значимых признаков для их дальнейшего применения при построении моделей. Убираются, например, признаки, которые избыточны или не несут никакой информации.

### **2.1.4. Методы анализа текста**

На данном шаге начинается построение модели с использованием разных методов, таких как кластеризация, классификация, информационный поиск и других. Данные методы распознавания данных также подходят и для интеллектуального анализа текста.

### **2.1.5. Интерпретация и оценка**

На последнем шаге (в зависимости от того, что требуется) проводится анализ результатов.

## **2.2. Области применения интеллектуального анализа текста**

Как уже упоминалось выше, интеллектуальный анализ текста стоит на пересечении разных дисциплин, включая в себя: извлечение информации, информационный поиск, обработку естественного языка и

интеллектуальный анализ данных. Ниже будет подробнее рассказано о каждой из областей.

### **2.2.1. Извлечение информации**

В процессе извлечения информации автоматически извлекается структурированная информация из неструктурированных данных. С помощью распознавания образов данная система определяет, например, где имена людей, где названия компаний, а где местоположение. То есть в документах происходит поиск predetermined последовательностей. Подобное решение позволяет получить элементы, подходящие для использования в базах данных для дальнейшего хранения, анализа или обработки.

### **2.2.2. Информационный поиск**

Наиболее известными системами информационного поиска являются поисковые системы Google. В данной задаче используются методы, используемые для хранения, представления и доступа к информации, которая преимущественно представлена в виде текстовых документов (а также новостных лент или книг), которые могут быть получены по запросу пользователя. Это своего рода расширение поиска по документам, позволяющее сужать набор документов, имеющих отношение к запросу пользователя. Эти системы значительно сокращают время, необходимое для поиска необходимой информации.

### **2.2.3. Обработка естественного языка**

Данная задача представляет из себя самую активную проблему в области искусственного интеллекта. Цель: исследовать естественный язык так, чтобы у компьютеров была возможность понимать языки, подобные тем, что используют для общения люди. Обработка естественного языка включает в себя распознавание и генерацию, которые отвечают за такие способности компьютера как «читать» и «говорить» на

естественном языке соответственно. Подобные системы включают в себя проверку грамматики, лексические, синтаксические и семантические анализаторы.

#### **2.2.4. Интеллектуальный анализ данных**

Данные задачи относятся к поиску знаний или релевантной информации в большом объеме данных. Система пытается обнаружить правила (статистически) и образцы (автоматически) от данных. Подобные системы имеют возможность предсказания, основываясь на «опыте», полученном в результате исследования.

## 3. Обзор существующих инструментов

В данном разделе будут рассмотрены основные инструменты, представленные в виде библиотек или отдельных сервисов. Внимание уделено в основном инструментам, работающим с русским языком.

### 3.1. Natural Language Toolkit

NLTK[3] является пакетом библиотек и программ для разработки программ на Python, работающих с естественным языком. Сопровождается обширной документацией, а также книгой<sup>3</sup>, 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.

### 3.2. Rymorphy2

Rymorphy2[2] написан на языке Python и имеет следующие возможности:

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

Распространяется rymorphy2 под лицензией MIT<sup>4</sup>, если используется в научной работе.

---

<sup>3</sup><http://www.nltk.org/book/>

<sup>4</sup><https://opensource.org/licenses/MIT>

### 3.3. Томита-парсер

Томита-парсер<sup>5</sup> 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 проекта<sup>6</sup> находится в свободном доступе.

### 3.4. Яндекс.Спеллер

Яндекс.Спеллер<sup>7</sup> 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.

### 3.5. OntosMiner

OntosMiner<sup>8</sup> является решением компании Eventos<sup>9</sup>, 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.

---

<sup>5</sup><https://tech.yandex.ru/tomita/>

<sup>6</sup><https://github.com/yandex/tomita-parser/>

<sup>7</sup><https://tech.yandex.ru/speller/>

<sup>8</sup><http://my-eventos.com/solution/ontosminer/>

<sup>9</sup><http://my-eventos.com/solution/ontosminer/>



## 4. Программная часть

В результате работы была написана программа, позволяющая автоматически анализировать новостные публикации сайта `mfd.ru`.

### 4.1. Описание

Программа способна выполнять следующие функции:

- Загружать заданное количество последних новостных публикаций определенной компании
- Загружать данные о котировках определенной компании за заданный промежуток времени
- Формировать и обучать рекуррентную нейронную сеть по заданным данным
- Предсказывать изменение цены по заданной новостной публикации

На вход программы подается название компании, выступающей в роли эмитента, количество новостей, начальная и конечные даты, в течение которых необходимо получить изменение изменения цен. В результате получают следующие файлы:

- `news/company.csv` — скаченные новости в формате csv с двумя колонками: дата и текст
- `stocks/company.csv` — скаченные котировки в формате csv с двумя колонками: дата, стоимость акций
- `stems/company.csv` — обработанные новости в формате, аналогичном новостям
- `connections/company.csv` — соединенные новости и котировки в формате csv с тремя колонками: дата, обработанный текст, изменение акции (положительное или отрицательное)

## 4.2. Используемые инструменты

Выбор инструментов основывался на тех задачах, которые нужно было решать в процессе написания программы. Исходя из поставленной задачи можно выделить следующие подзадачи:

- Загрузка данных с интернет-ресурсов, для чего необходима работа с web-запросами
- Преобразование содержимого web-страниц, для чего нужны инструменты преобразования содержимого html-файлов
- Преобразование текстовых документов в более пригодный для обучения вид
- Обучение рекуррентной нейронной сети, для чего необходимы соответствующие инструменты

В связи с подзадачами был выбран язык программирования Python версии 3.6.0 и библиотеки `urllib`<sup>10</sup> (работа с web-запросами) версии 1.21.1, `bs4`<sup>11</sup> (обработка html-файлов) версии 4.6.0, `nltk`<sup>12</sup>[3] (преобразование текстовых документов) версии 3.2.2 и `keras`<sup>13</sup>[1] (работа с рекуррентными нейронными сетями) версии 2.0.3. Возможность написания всех программных модулей на одном языке упрощает разработку и поддержку, что было еще одним преимуществом.

## 4.3. Структура программы

Всего в программе присутствует 6 основных файлов (модулей), каждый из которых отвечает за свою часть работы. Схема модулей приведена на Рис. 2. `news_getter.py` отвечает за скачивание новостей с сайта `mfd.ru`, за запись новостей в файл и за чтение новостей из файла. `stock_getter.py` отвечает за загрузку котировок с сайта `finam.ru`, за запись котировок в файл и за чтение котировок из файла. `connector.py`

---

<sup>10</sup><https://docs.python.org/3/library/urllib.html>

<sup>11</sup><https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

<sup>12</sup><http://www.nltk.org/>

<sup>13</sup><https://keras.io>

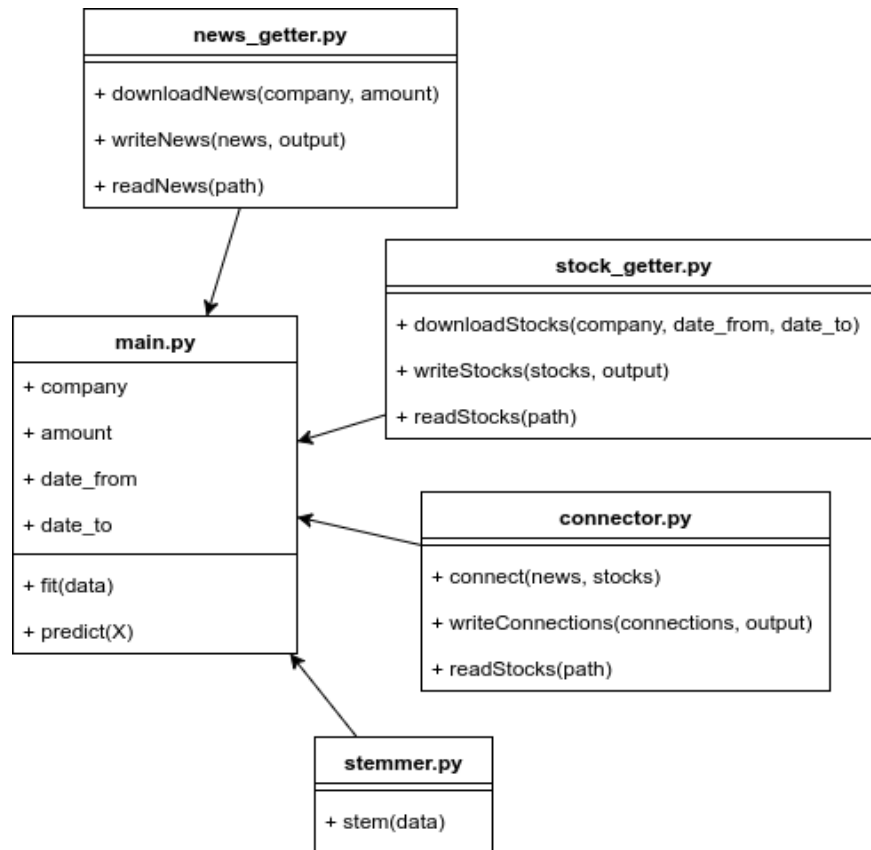


Рис. 2: Структура программы

## 4.4. Результаты

Че как работает

## Заключение

## Список литературы

- [1] Chollet François et al. Keras. — <https://github.com/fchollet/keras>. — 2015.
- [2] 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.
- [3] Loper Edward, Bird Steven. NLTK: The Natural Language Toolkit // Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. — ETMTNLP '02. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2002. — P. 63–70.
- [4] Sumathy K. L., Chidambaram M. Article: Text Mining: Concepts, Applications, Tools and Issues — An Overview // International Journal of Computer Applications. — 2013. — October. — Vol. 80, no. 4. — P. 29–32. — Full text available.
- [5] V.P. Romanov. Information technology modeling of financial markets - (Applied Information Technology) / Informatsionnye tekhnologii modelirovaniya finansovykh rynkov - ("Prikladnye informatsionnye tekhnologii"). — Finansy i statistika, 2010. — ISBN: 5279034444.