

**МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ  
УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА**  
**Факультет информатики и систем управления**  
**Кафедра теоретической информатики и компьютерных  
технологий**

Наброски дипломного проекта

«Автоматическое установление связей между сообщениями  
твиттера и новостными статьями»

Выполнил:

студент ИУ9-101

Выборнов А. И.

Руководитель:

Лукашевич Н.В.

Москва 2016

# Содержание

<b>Введение</b>	<b>4</b>
<b>1. Обзор литературы</b>	<b>6</b>
1.1. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media . . . . .	6
1.1.1. Перевод аннотации . . . . .	6
1.1.2. Идея статьи . . . . .	7
1.2. Linking Online News and Social Media . . . . .	8
1.2.1. Перевод аннотации . . . . .	8
1.3. Bridging Vocabularies to Link Tweets and News . . . . .	9
1.3.1. Основная идея . . . . .	9
1.4. TwitterStand: News in Tweets . . . . .	10
<b>2. Дополнительная литература</b>	<b>11</b>
2.1. Gibberish, Assistant, or Master? Using Tweets Linking to News for Extractive Single-Document Summarization . . . . .	11
2.2. Detecting Event-Related Links and Sentiments from Social Media Texts . . . . .	11
2.3. Определение тематической направленности текстового со- держимого микроблогов . . . . .	11
2.4. Разработка сервиса извлечения мнений . . . . .	11
<b>3. Полезные ссылки</b>	<b>11</b>
<b>4. Описание реализации</b>	<b>12</b>
4.1. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media . . . . .	12
4.1.1. Построение датасетов . . . . .	12
4.1.2. Evaluation . . . . .	12
4.1.3. WTMF . . . . .	13
4.1.4. Построение связей текст-текст . . . . .	14

4.1.5. WTMF-G . . . . .	15
<b>5. Формирование датасетов</b>	<b>16</b>
5.1. Примерно что делал . . . . .	16
<b>6. Используемое ПО</b>	<b>17</b>
6.1. tweepy . . . . .	17
<b>Список литературы</b>	<b>18</b>

# Введение

В современном мире всё больший вес приобретают социальные медиа (преимущественно социальные сети). Их главное отличие от традиционных медиа (газеты, тв) заключается в том, что контент порождается тысячами и миллионами людей. Социальные медиа не заменяют традиционные новостные источники, а дополняют их. Они могут служить полезным социальным датчиком того, насколько популярна история (тема) и как долго. Часто обсуждения в социальных медиа основаны на событиях из новостей и, наоборот, социальные медиа влияют на новостные события.

Одной из самых популярных социальных сетей является Twitter - социальная сеть для публичного обмена сообщениями. Главной особенностью Twitter является малый размер сообщений (140 символов), называемых твитами. Часто твиты являют собой описание, происходящего прямо сейчас события, отклик на него.

...

Выявление связи между сообщениями твиттера (твитов) и новостями позволит как расширить информативность твитов, так и обогатить новости.

...

Преимущества расширения новости с помощью твитов: определение отношения аудитории к новости, дополнительные признаки для тематической классификации новостей, дополнительная информация для аннотирования новостей.

Современные методы обработки естественного языка хорошо работают, используя большой массив текста в качестве входных данных, однако, они становятся неэффективными, когда применяются на коротких текстах, таких как твиты. Существенным преимуществом расширения твита с помощью новости является появляющаяся возможность использования большого количества методов обработки естественного языка

(Natural Language Processing).

...

Данная работа ставит целью исследование и разработку методов автоматического установления связей между сообщениями твиттера и новостными статьями.

Не существует стандартных решений, и есть считанное количество статей. На основе этих статей будет сделана попытка построить pipeline для получения подобной взаимосвязи.

# 1. Обзор литературы

В рамках предварительного исследования были разобраны несколько статей [1] [2] [3]. Ниже приводится краткое изложение основных идей, описанных в выбранных статьях.

## 1.1. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media

### 1.1.1. Перевод аннотации

Многие современные методы обработки естественного языка (NLP<sup>1</sup>) хорошо работают с большой массив текста в качестве входных данных. Однако они очень неэффективны при работе с короткими текстами (к примеру твиты). Преодоление этой проблемы мы видим в нахождении соответствующего твиту новостного документа. Решение этой задачи требует хорошего моделирования семантики коротких текстов.

Основной вклад статьи двойной:

1. представлено решение задачи нахождения взаимосвязи между твитами и новостями, из этого могут извлечь выгоду многие NLP задачи;
2. в отличие от предыдущих исследований, которые фокусируются на лексических особенностях коротких текстов (информация о связи текст-слово), мы предлагаем взаимосвязь, основанную на модели скрытой переменной, которая моделирует корреляцию между короткими текстами (информация о связи текст-текст). Необходимость этого обоснована наблюдением: твит обычно покрывает только один аспект события.

---

<sup>1</sup>Natural Language Processing

Мы покажем, что с помощью особенных признаков твита (хэштегов) и особых признаков новостей (именованных сущностей<sup>1</sup>) а также временных ограничений, мы можем получить взаимосвязь текст-текст, и, таким образом, дополнить семантическую картину короткого текста. Наши эксперименты показывают значительное преимущество нашей новой модели над baseline<sup>2</sup>.

### 1.1.2. Идея статьи

Современные методы обработки естественного языка плохо работают с короткими текстами. Для преодоления этого к твитам привязываются соответствующие новости.

Для формирования обучающей выборки, были выбраны твиты, которые имели ссылки на новости, опубликованные новостными агентствами (CNN или NYT) в тот же период.

Как показано в статье [5], добавление к твиту содержимого веб-страницы, ссылка на которую включена в этот твит, повышает **purity score** их кластеризации с 0.280 до 0.392.

Модели со скрытой переменной хорошо подходят для отображения коротких текстов в плотный малоразмерный вектор. В рамках решения задачи была применена модель со скрытой переменной, которая называется WTMF (Weighted Textual Matrix Factorization, подробное описание[6]), к твитам и к новостям. Модель была протестирована на двух схожих наборах данных из небольших сообщений. Как результат - используемая модель с большим запасом превзошла и LSA (Latent Semantic Analysis) и LDA (Latent Dirichlet Allocation). Эта модель позволила добавить информацию об отсутствующих словах в твит (модель WTMF добавляет более 1000 фичей к твиту, LDA лишь 14). Недостатком WTMF является то, что порождается только связь текст-слово, без

---

<sup>1</sup>Какой-то кривой перевод, найдо найти получше. In data mining, a named entity is a phrase that clearly identifies one item from a set of other items that have similar attributes.

<sup>2</sup>Как перевести?

учёта взаимосвязи между короткими текстами.

Ввиду разреженности исходных данных, возникает ещё одна проблема: твит обычно отражает, только один аспект события.

Полученный подход не учитывает следующих характеристик, которым обладает исходная выборка:

1. Хэштеги, которые являются прямым указанием на смысл твита.
2. **Named entities** новостей. Из новостей можно с высокой точностью извлекать **named entities**, используя инструменты для NER (Named Entity Recognition). Если несколько текстов содержат схожие **named entities** они наверняка описывают одно и то же событие.
3. Информация о времени публикации для твитов и новостей. Если несколько текстов опубликованы примерно в одно и то же время, то велик шанс, что они описывают одно и то же событие

В статье описывается решение проблемы поиска взаимосвязи между текстами, с использованием описанных выше характеристик. Два связанных текста, должны иметь схожий скрытый вектор (семантическая модель твита достраивается из схожих твитов).

Это дополнительная информация была добавлена в модель WTMF. Было также показано различное влияние на связь текст-текст жанра твита и жанра новости. Был получен на порядок более лучший результат чем при использовании исходной WTMF модель.

## 1.2. Linking Online News and Social Media

### 1.2.1. Перевод аннотации

Многое из того, что обсуждается в социальных медиа вдохновлено событиями, описанными в новостях и, наоборот, социальные медиа предоставляют механизм, позволяющий влиять на новостные события.



Мы обращаемся к следующей задаче: по новости, найти в социальных сетях высказывания, которые неявно на неё ссылаются. Используется трехступенчатый подход: сначала получаются несколько моделей запросов по исходной статье, затем модели используются для получения высказываний из индекса целевого социального медиа, результатом являются несколько ранжированных списков, которые объединяются с использованием особой техники слияния данных. Модель запроса создаётся как на основе структуры статьи, так и на основе явно связанных со статьей высказываний из социальных медиа. Для борьбы с дрейфом запроса<sup>1</sup> при большого объёме используемого текста (либо в новости, либо в явно связанных высказываниях из социальных медиа), предлагается основанный на графике метод для выбора отличительных условий.

В нашей экспериментальной оценки для порождения моделей запросов, использованы данные из Twitter, Digg, Delicious<sup>2</sup>, the New York Times Community, Wikipedia и блогосферы. Показано, что другие модели запросов, основанные на различных источниках данных, не только обеспечивают дополнительную информацию, но и влияют на получение различных высказываний из социальных медиа по нашему целевому индексу. Как следствие, методы слияния данных приводят к значительному повышению производительности в сравнении с индивидуальными подходами. Показано, что основанный на графике метод выделения условий помог улучшить как эффективность, так и продуктивность.

## 1.3. Bridging Vocabularies to Link Tweets and News

### 1.3.1. Основная идея

Значительную сложность при решении проблемы связывания твитов с новостями преимущественно вызывают малый размер твита и различия в словарях: в твитах используются аббревиатуры, неформальный

---

<sup>1</sup>Порождение менее подходящего запроса.

<sup>2</sup>Веб-сайт, бесплатно дающий зарегистрированным пользователям услугу хранения и публикации закладок на страницы Всемирной сети.

язык, сленг, в новостях, напротив, используется литературный язык. Также твиты очень зашумлены и не содержат полезного содержания.

Твиттер предлагает хештэги, как механизм для категоризации твитов. Но этот подход далеко не совершенен, так как не только далеко не все записи содержат хештеги, но и записи содержащие хештеги обладают рядом проблем. Таковыми как: хештег не содержит информацию о событии, хештег сформулирован в слишком общей форме, твит содержит несколько хештегов. Из этого делается вывод, что использование только хештегов приведёт к низкому качеству связывания твитов с новостями.

Предлагается следующий подход: Используется LDA для построения моделей тем поверх новостей. Затем среди твитов ищутся наиболее близкие к конкретному топику. Из полученных твитов извлекаются слова, которые служат “мостом” к другим твитами.

## 1.4. TwitterStand: News in Tweets

Формирование дайджеста новостей на основе твиттера, не очень коррелирует с темой. Но возможно в статье есть хорошие идеи, которые помогут с решением задачи.

## 2. Дополнительная литература

2.1. Gibberish, Assistant, or Master? Using Tweets Linking to News for Extractive Single-Document Summarization

2.2. Detecting Event-Related Links and Sentiments from Social Media Texts

2.3. Определение тематической направленности текстового содержимого микроблогов

2.4. Разработка сервиса извлечения мнений

## 3. Полезные ссылки

- Твиттер NYT: <https://twitter.com/nytimes> (20млн подписчиков, 200тыс твитов)
- Крайне отстойная статья на тему: <http://cyberleninka.ru/article/n/issledovanie-otklika-polzovateley-twitter-na-novosti-iz-smi>
- Идея для формирования train: <http://techcrunch.com/2013/08/19/twitter-related-headlines/>

## 4. Описание реализации

Из предложенных статей единственная полноценная - Linking tweets to news by guo.

### 4.1. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media

#### 4.1.1. Построение датасетов

За один и тот же промежуток выкачиваем твиты с помощью stream api, новости с помощью rss.

Твит задаётся кортежем: (time, author, text). Новость задаётся кортежем: (time, title, summary, url).

Train/test множества формируем из твитов, которые содержат единственную ссылку на новость, из выкаченных нами ранее, и не совпадают с заголовком новости.

#### 4.1.2. Evaluation

Используем метрику *ATOP* (метрика подробно описана в [7]). Рассмотрим что означает эта метрика в применении к нашей задаче (я немного модифицировал метрику, для более простого описания, полученная метрика полностью совпадает с описанной метрикой). Пусть  $T$  - это множество твитов,  $N \in \mathbb{N}$  - размер рассматриваемого топа новостей для твита (могут быть все новости вообще),  $k < N \in \mathbb{N}$ .  $TOPK_t(k) = 1$ , если твит  $t \in T$  соответствует хотя бы одной новости в top-k результатов, иначе  $TOPK_t(k) = 0$

$$TOPK(k) = \frac{\sum_{t \in T} TOPK_t(k)}{|T|},$$

$$ATOP = \frac{\sum_{k=1}^N TOPK_t(k)}{N} = \frac{1}{|T| * N} \sum_{k=\overline{1, N}, t \in T} TOPK_t(k).$$

Значения метрики  $ATOP$  лежат на отрезке  $[0, 1]$ . Чем ближе  $ATOP$  к 1 тем лучше.

#### 4.1.3. WTMF

WTMF - модель применяемая для анализа схожести между короткими текстами [6]. Модель рассматривает отсутствующие в тексте слова как признаки короткого текста. Отсутствующие слова это все слова корпуса рассматриваемых текстов за исключением слов из рассматриваемого короткого текста. Отсутствующие слова являются негативным сигналом для смысла коротких текстов.

WTMF похож на SVD, но использует не разложение, а непосредственный расчёт каждой ячейки. Модель раскладывает матрицу  $X \sim P^T Q$ .

Корпус рассматривается как матрица  $X$  размера  $M \times N$ : строки - это слова (всего  $M$ ), столбцы - короткие тексты (всего  $N$ ), ячейки - мера tf-idf. Как показано на рисунке 1 матрица  $X$  приближается произведением двух матриц  $P$  размера  $M \times K$  и  $Q$  размера  $K \times N$ .

$$X \approx P^T \times Q$$

Рисунок 1 — wtmf

Каждый текст  $s_j$  представлен в виде вектора  $Q_{\cdot, j}$  размерности  $K$ , каждое слово  $w_i$  представлено в виде вектор  $P_{i, \cdot}$ . Когда их скалярное произведение  $X_{ij}$  близко к нулю, то мы считаем, что это отсутствующее слово.

Задачей модели является минимизация целевой функции ( $\lambda$  - регуляризирующий член, матрица  $W$  определяет вес каждого элемента матрицы  $X$ ):

$$\sum_i \sum_j W_{ij} (P_{i,\cdot} \cdot Q_{\cdot,j} - X_{ij})^2 + \lambda \|P\|_2^2 + \lambda \|Q\|_2^2.$$

Для получения векторов  $P_{i,\cdot}$  и  $Q_{\cdot,j}$  используется алгоритм описанный в статье [8]. Сначала  $P$  и  $Q$  инициализируются случайными числами. Затем запускается итеративный пересчёт  $P$  и  $Q$  по следующим формулам (эффективный способ расчёта описан в [7]):

$$P_{i,\cdot} = (QW'_iQ^T + \lambda I)^{-1}QW'_iX_{i,\cdot}^T,$$

$$Q_{\cdot,j} = (PW'_jP^T + \lambda I)^{-1}PW'_jX_{\cdot,j}.$$

Здесь  $W'_i = \text{diag}(W_{i,\cdot})$  - диагональная матрица полученная из  $i$ -ой строчки матрицы  $W$ , аналогично  $W'_j = \text{diag}(W_{\cdot,j})$  - диагональная матрица полученная из  $j$ -ого столбца.

Определим матрицу  $W$  следующим образом:

$$W_{ij} = \begin{cases} 1, & \text{if } X_{ij} \neq 0, \\ w_m, & \text{otherwise.} \end{cases},$$

где  $w_m$  положительно и  $w_m \ll 1$ .

#### 4.1.4. Построение связей текст-текст

Твиты связываются с помощью хэштегов, named entities и времени.

Связь твитов с помощью хэштегов. Сначала извлекаем все хэштеги из твитов, затем превращаем в хэштеги все слова во всех твитах, которые совпали с ранее извлечёнными хэштегами. Для каждого твита и для каждого хэштега извлекаем  $k$  твитов, которые содержат этот этот хэштег, если хэштег появлялся в более чем  $k$  твитах берём  $k$  твитов наиболее

близких во времени к исходному.

Связь твитов с помощью named entities. Применяем методы NER к новостным summary и получаем множество named entities. Затем применяем тот же подход, что и к хэштегам, сначала превращаем в NE слова из твитов, которые совпали с полученными NE, а затем получаем  $k$  связей для каждого твита.

Связь твитов с помощью времени. Аналогично вышеописанному для каждого твита выбираем  $k$  связей с наиболее схожими твитами в окрестности 24 часов. Наиболее близкие находятся с помощью косинусной меры, рассчитываемой для векторов из таблицы X.

Новости связываются только по времени.

#### 4.1.5. WTMF-G

Добавление связей текст-текст в WTMF происходит с помощью влияния на regularization term. Для каждой пары связанных текстов  $j_1$  и  $j_2$ :

$$\lambda = \delta \cdot \left( \frac{Q_{\cdot, j_1} \cdot Q_{\cdot, j_2}}{|Q_{\cdot, j_1}| |Q_{\cdot, j_2}|} - 1 \right)^2,$$

коэффициент  $\delta$  задаёт степень влияния связей текст-текст.

Полученная модель и называется WTMF-G (WTMG on graphs).

Alternating Least Square используемый в [7] не применим из-за нового regularization term, который зависит от  $|Q_{\cdot, j}|$  (по-хорошему нужно понять почему). Для того, чтобы мы могли применить ALS мы вводим упрощение: длина вектора  $Q_{\cdot, j}$  не изменяется во время итерации. Получаем уравнения:

$$P_{i, \cdot} = (QW_i'Q^T + \lambda I)^{-1}QW_i'X_{i, \cdot}^T,$$

$$Q_{\cdot, j} = (PW_j'P^T + \lambda I + \delta L_j^2 Q_{\cdot, n(j)} \text{diag}(L_{n(j)}^2) Q_{\cdot, n(j)}^T)^{-1} (PW_j'X_{j, \cdot} + \delta L_j Q_{\cdot, n(j)} L_{n(j)}).$$

В этих формулах  $n(j)$  — список связанных текстов с текстом  $j$ .  $Q_{\cdot, n(j)}$  — матрица, состоящая из связанных векторов для  $Q_{\cdot, j}$ .  $L_j$  - длина вектора

$Q_j$  на начало итерации,  $L_n(j)$  — вектор длин векторов связанных с  $j$  i.e.  $Q_{\cdot, n(j)}$ , полученный на начало итерации.

## 5. Формирование датасетов

Для формирования датасетов нужно сделать:

1. Выбрать источники данных.
2. Выбрать формат хранения.
3. Выбрать БД, для записи результата.
4. Написать отказоустойчивый консьюмер.
5. В течение длительного времени собрать данные.

Предлагаю выбор формата хранения и БД отложить до тех пор, пока не будут получены все необходимые данные. А пока коллекционировать все возможные данные в текстовых файлах.

В качестве источников данных

### 5.1. Примерно что делал

Лог работы в raw формате :) будет много копиясты с используемых мануалов

1. Step 1: Getting Twitter API keys

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

Create a twitter account if you do not already have one. Go to <https://apps.twitter.com/> and log in with your twitter credentials. Click "Create New App" Fill out the form, agree to the terms, and



click "Create your Twitter application" In the next page, click on "API keys" tab, and copy your "API key" and "API secret". Scroll down and click "Create my access token and copy your "Access token" and "Access token secret".

2.

## 6. Используемое ПО

### 6.1. tweepy

<https://github.com/tweepy/tweepy>

# Список литературы

- [1] W. Guo, H. Li, H. Ji, and M. T. Diab. Linking tweets to news: A framework to enrich short text data in social media. - ACL, pages 239–249, 2013.
  - [2] Manos Tsagkias, Maarten de Rijke, Wouter Weerkamp. Linking Online News and Social Media. - ISLA, University of Amsterdam.
  - [3] T. Hoang-Vu, A. Bessa, L. Barbosa and J. Freire. Bridging Vocabularies to Link Tweets and News. - International Workshop on the Web and Databases (WebDB 2014), Snowbird, Utah, US, 2014.
  - [4] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, J. Sperling. TwitterStand: news in tweets. - 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009, Seattle, Washington.
  - [5] Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In Proceedings of the 20th ACM international conference on Information and knowledge management.
  - [6] Weiwei Guo and Mona Diab. 2012a. Modeling sentences in the latent space. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.
  - [7] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
  - [8] Nathan Srebro and Tommi Jaakkola. 2003. Weighted low-rank approximations. In Proceedings of the Twentieth International Conference on Machine Learning.
-