

Дипломная работа на тему:
«Автоматическое установление связей
между сообщениями твиттера и
новостными статьями»

Исполнитель: Выборнов А.И.
Руководитель: Лукашевич Н.В.

Актуальность

Установление связей твит-новость позволяет:

- обогатить текст новости, короткими характеристиками, с целью решения различных задач, таких как:
 - получение реакции аудитории на новость,
 - автоматическое аннотирование новостей,
 - классификация новостей;
- расширить текст твита, с целью применения методов обработки естественных языков:
 - тематическое моделирование,
 - классификация коротких текстов,
 - выявление тональности твита.

Постановка задачи

В рамках работы необходимо:

- Произвести исследование методов установления связей между твитами и новостными статьями.
- Собрать и разметить необходимые для решения данные
- Реализовать ПО, позволяющее различными методами для произвольного твита построить рекомендацию новостей и оценить качество методов.
- Определить наилучший метод для установления связей между твитами и новостями.

Получение данных

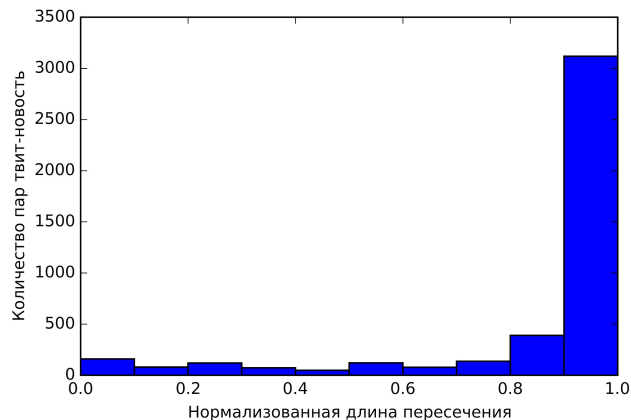
- **ТВИТЫ.**
 - Twitter Streaming API.
 - 1% от всех публикуемых ТВИТОВ.
- **НОВОСТИ.**
 - RSS-потoki: lifenews.ru, ria.ru, lenta.ru, russian.rt.com, gazeta.ru
 - $\approx 3\%$ от общего числа ссылок в твитах.
- **Данные собраны за период с 06.04.2016 по 17.04.2016.**

Рассматриваемое множество	Размер
ТВИТЫ	495552
Новости	13711
Твиты, содержащие URL	150510
Уникальные URL	101017
Твиты со ссылкой на новость	4324
Новости на которые есть ссылки из твита	2979

Наборы данных

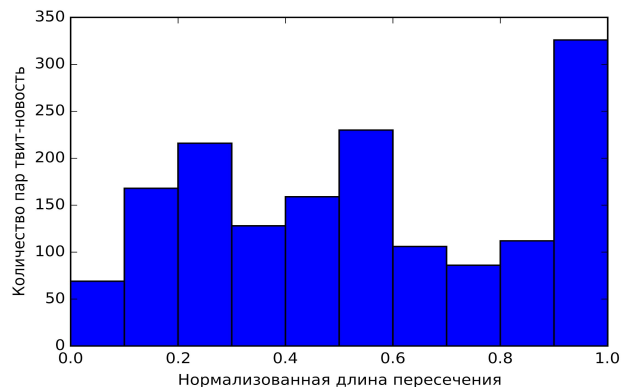
Автоматическая разметка

- Твитов / новостей - 4324 / 13711
- Всего связей - 4324
- Нетривиальных связей - 746



Ручная разметка

- Твитов / новостей - 1600 / 13711
- Всего связей - 1600
- Нетривиальных связей - 976



Установление связей

Реализовано три метода установления связей:

- TF-IDF
- WTMF
- WTMF-G

В качестве результата для каждого текста строится численный вектор, по которому производится сравнение текстов на схожесть.

На основе схожести численных векторов для каждого твита находятся наиболее похожие на него новости.

Метод TF-IDF

- Для каждого слова рассчитывается метрика TF-IDF.

$$TF-IDF = tf \cdot \log \frac{N}{df}$$

- На основе метрики заполняется TF-IDF матрица.

	t1	t2	t3	t4
w1	0	3.18	0	0.35
w2	0	6.1	1	0
w3	1.93	2.54	1.51	0
w4	2.37	1.51	0	0
w5	0	0	0.21	0.88
w6	0	0	4.15	1.96

Метод WTMF

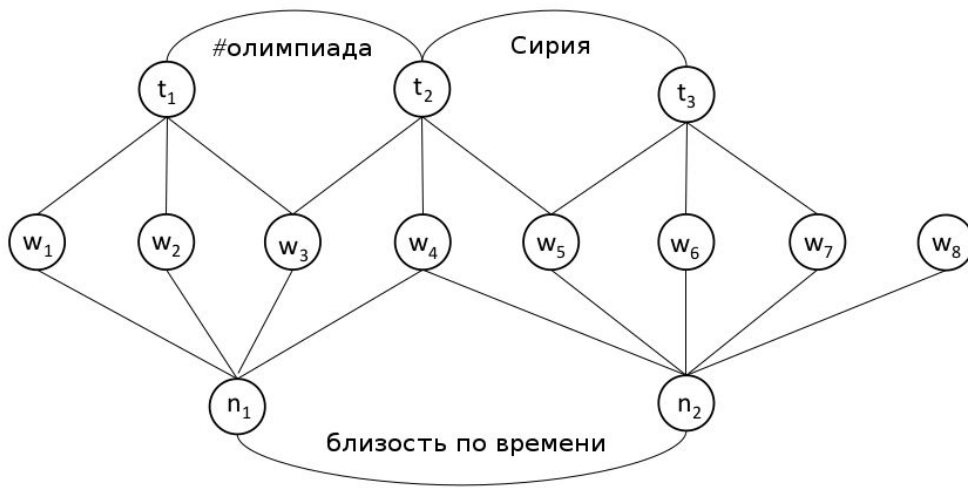
- Представляет собой особый способ численного разложения TF-IDF матрицы X на произведение двух матриц: P и Q .

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

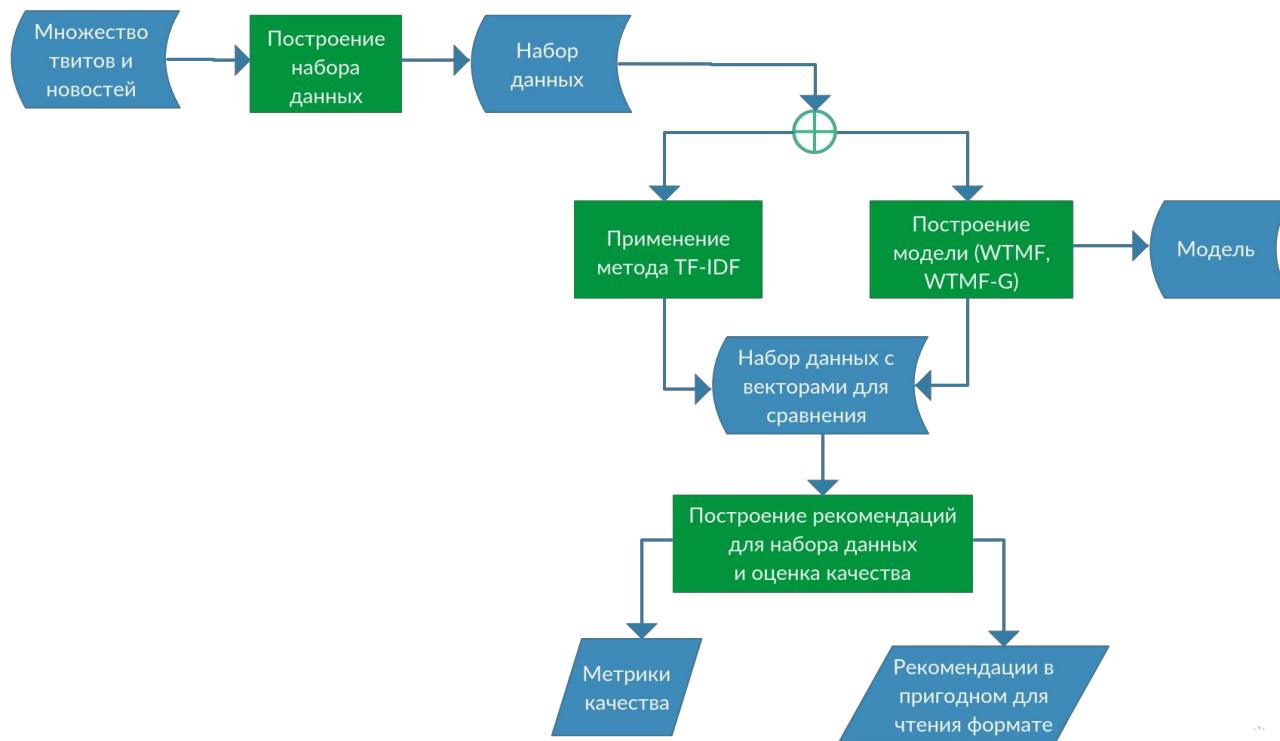
- Результат - матрица Q .
- Столбец матрицы Q - вектор для сравнения текстов.

Метод WTMF-G

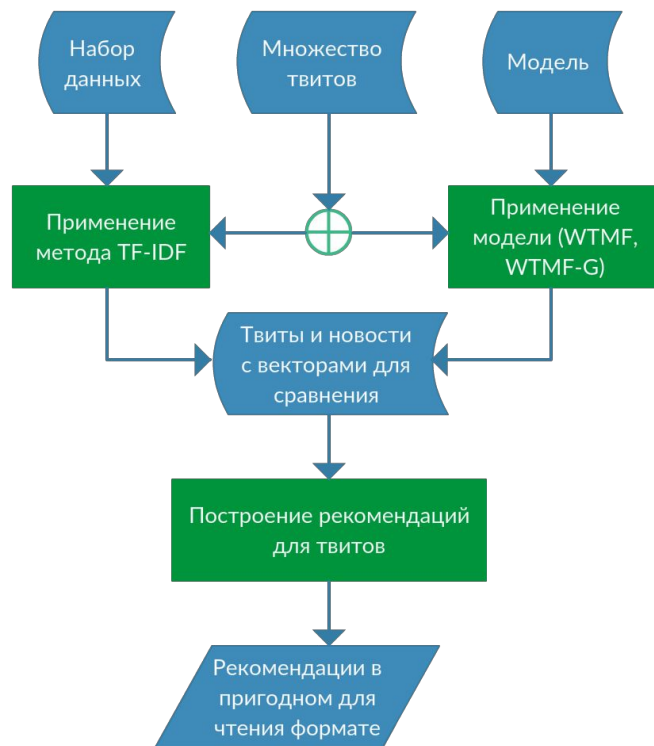
Метод WTMF-G основан на методе WTMF и позволяет учитывать дополнительные семантические связи между текстами:



Построение моделей и оценка качества



Построение рекомендаций



Используемые в работе метрики

- Средний обратный ранг

$$MRR = \frac{1}{n} \sum_{i=1}^n \frac{1}{rank_i}$$

- Процент попаданий в l первых результатов

$$TOP_l = \frac{1}{n} \sum_{i=1}^n Q_l(i)$$

ТВИТ

НОВОСТЬ

НОВОСТЬ

НОВОСТЬ

НОВОСТЬ

НОВОСТЬ

НОВОСТЬ

Результаты экспериментов

Метрика MRR

	WTMF	WTMF-G	TF-IDF
auto	0.8640	0.8685	0.8817
manual	0.7293	0.7854	0.8310
auto_nt	0.5297	0.5695	0.6035
manual_nt	0.6194	0.7000	0.7565

Метрика TOP3

	WTMF	WTMF-G	TF-IDF
auto	0.9148	0.9195	0.9299
manual	0.8193	0.8775	0.9137
auto_nt	0.5750	0.6099	0.6461
manual_nt	0.7223	0.8217	0.8688

Пример результата

- Твит: Землетрясение в Японии: много людей заблокированы под обломками зданий
Заголовки новостей:
 - В Японии после землетрясения госпитализированы более 760 человек
 - СМИ: жертвой землетрясения в Японии стал один человек
 - Лукашенко соболезнует в связи с землетрясениями в Японии и Эквадоре
 - Число жертв землетрясений в Японии возросло до 31
 - В Японии произошло землетрясение магнитудой 7,1
- Твит: Оруэлл в Омске
Заголовки новостей:
 - Владимир Путин пообещал починить дороги в Омске к юбилею
 - Мэр Омска: улицу Омска, которую показали Путину, отремонтируют
- Твит: Apple отказалась взломать iPhone по требованию суда в Бостоне
Заголовки новостей:
 - Суд обязал Apple помочь ФБР взломать iPhone возможного преступника
 - Apple помогла властям США взломать более 70 iPhone
 - Бостонский судья обязал Apple взломать iPhone предполагаемого бандита

Технико-экономическое обоснование

- Трудоёмкость - 740 дней.
- Общее время выполнения - 96 рабочих дней.
- Стоимость программного продукта 7967414 рублей.
- Итоговый чистый дисконтированный доход составил 2101032 рублей.

Заключение

- Решена задача автоматического установления связей между твитами и новостями.
- Решение реализовано тремя способами.
- На основе решения получено, что
 - некачественный набор данных завышает значения метрик;
 - дополнительные семантические связи улучшают результаты рекомендаций;
 - наилучший метод для решения задачи - TF-IDF.