

Автоматическое установление связей между сообщениями твиттера и новостными статьями

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

Цель и назначение разработки

- Цель: исследование и разработка методов автоматического установления связей между сообщениями твиттера и новостными статьями.
- Назначение: выявление связи между сообщениями твиттера (твитами) и новостями позволит как расширить информативность твитов, так и обогатить новости. Это позволит лучше решать такие задачи, как построение тематических моделей, классификация сообщений твиттера, определение тональности сообщения, автоматическое аннотирование новостей.

Обобщённая формулировка

- Установление семантических связей между короткими текстами.

Реализация

- Получение твитов и новостей.
- Разрешение ссылок.
- Предобработка данных.
- Построение датасета.
- Установление связей между новостями и твитами.
- Построение моделей WTMF, WTMF-G.
- Оценка качества построенных моделей.

Получение твитов и новостей

- Новости получаются с помощью парсинга rss-ленты. Для новостей сохраняем:
 - заголовок,
 - краткое описание,
 - время публикации,
 - ссылку на новость,
 - название новостного портала.
- Твиты выкачиваются с помощью Twitter Streaming API. Выкачивается 1% от общего количества твитов. Для твитов сохраняем:
 - текст,
 - время публикации,
 - является ли ретвитом.

Разрешение ссылок

- Очень медленный процесс.
- Оптимизации:
 - работа только с заголовками страниц;
 - использование быстрых и менее надёжных подходов, и только для не разрешённых использование медленных но более точных;
 - разрешение в несколько этапов.
- Идеи для дальнейшей оптимизации:
 - фильтрация ссылок основанная на метаданных твиттера;
 - кэширование dns.

Примеры разрешения ссылок

- <https://t.co/Ke8KMdNuX4>
 - <http://bit.ly/1QKpHMu>
 - <http://hamastery.com/interesnoe/divergent-glava-2-insurgent-skachat-torrent-1080.html>
- <https://t.co/oztuA8pfFy>
 - <https://twitter.com/leprasorium/status/707381506390425601>
- <https://t.co/98YuY5aoBu>
 - http://gigam.es/imtw_Tribez
 - <https://bitly.com/1mcWUoR>
 - https://apps.facebook.com/thetribez/?rid=viral_fb&pid=110
 - <https://apps.facebook.com/thetribez/>
- <https://t.co/VcJSsbBCUM>
 - <http://vk.cc/4T4xXI>
 - https://m.vk.com/wall264189899_1186

Немного статистики

- За неделю:
 - 341863 твитов
 - 134945 ссылок
 - 115940 уникальных ссылок
 - ~ 7000 новостей с сайтов lifenews и ria

домен	количество	% от общего числа
twitter.com	36807	31.75
apps.facebook.com	6234	5.38
www.youtube.com	3659	3.16
m.vk.com	2400	2.07
www.periscope.tv	2215	1.91
news.yandex.ru	2041	1.76
www.instagram.com	1798	1.55
su.epeak.in	1624	1.4
www.facebook.com	1406	1.21
lifenews.ru	888	0.77
ria.ru	863	0.74
l.ask.fm	803	0.69

Предобработка данных

- Токенизация.
 - Используется nltk.
- Удаление знаков препинания.
- Удаление стопслов.
 - Корпус стопслов из nltk.
- Лемматизация.
 - Английский текст - nltk WordNet.
 - Русский текст - pymorhy2.

Примеры предобработки данных

- Политолог: Россия стремится к миру в Сирии, сокращая группировку войск
 - политолог россия стремиться мир сирия сокращать группировка войско
- Пресс-конференция по итогам визита на Кубу Барака Обамы
 - пресс-конференция итог визит куб барак обама
- Президент Кипра соболезнает в связи с катастрофой в Ростове-на-Дону
 - президент кипр соболезнавать связь катастрофа ростов-на-дону

Построение датасета

- В датасет входят все новости и твиты, содержащие ссылки на эти новости.

Установление связей

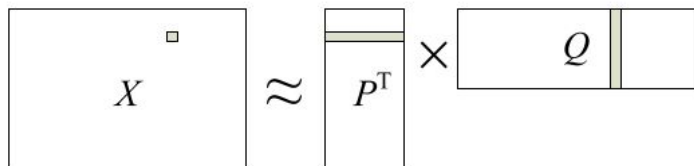
- Применяем методы NER к краткому описанию новостей и получаем множество named entities.
 - Polyglot
- Извлекаем все хэштеги из твитов, получаем множество хэштегов.
- Твиты связываются с помощью хэштегов, named entities и времени.
- Новости связываются только по времени.

Пример работы NER

- Вход:
 - Чешской армии необходимо три года, чтобы количественно дополнить свои ряды согласно требованиям НАТО, заявил в воскресенье начальник генштаба генерал-лейтенант Йозеф Бечварж, выступая по республиканскому ТВ.
- Выход:
 - I-ORG НАТО
 - I-PER Йозеф Бечварж
 - I-ORG ТВ

Построение модели WTMF

- Позволяет устанавливать семантические связи между короткими текстами основываясь на tf-idf матрице.
- Заключается в итеративном построении разложения tf-idf матрицы вида:

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


- В результате для каждого текста строится скрытый вектор, содержащий в матрице Q.
- Вектор представляет собой набор чисел с плавающей запятой.

Оптимизация построения модели WTMF

- Построение модели есть многократное перемножение матриц.
- Для оптимизации рассматривалась часть алгоритма построения модели, заключающаяся в перемножении нескольких больших матриц. Ниже приведено время за 100 итераций.

что сделали	время выполнения (с.)	во сколько раз стало быстрее
наивная реализация	205	1
перемножение с помощью OpenBlass	55	3.73
вынесение общих множителей	15.15	3.63
переход к работе с разреженными матрицами	0.75	20.2
удаление приведения матрицы к python list	0.63	1.21

Построение модели WTMF-G

- Расширение модели WTMF с использованием информации о связях между короткими текстами.

Оценка качества построенных моделей

- Используем метрику АТОР.
- $TOPK_t(k) = 1$, если твит t соответствует хотя бы одной новости в ТОП- k результатов, иначе $TOPK_t(k) = 0$

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

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

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

Результаты для построенной модели WTMF

- Рассматривался $k=10$.
- Для случайным образом выбранных 10 новостей метрика равна достаточно малому числу: 0.001.

размерность скрытого вектора\количество итераций алгоритма	1	2	3	4	5	6	7
10	0.27	0.37	0.38	0.35	0.32	0.39	0.26
20	0.45	0.58	0.56	0.45	0.37	0.32	0.28
30	0.57	0.69	0.59	0.52	0.39	0.34	-
50	0.65	0.75	-	-	-	-	-

Примеры

- Солдаты шиитской организации «Хезболла» нанесли ракетный удар по позициям террористов ИГИЛ* недалеко от ливано-сир...
 - СМИ: «Хезболла» нанесла ракетный удар по базе ИГИЛ на границе с Сирией
 - Глава Македонии: нелегальная миграция угрожает безопасности в регионе
 - СМИ: У Пальмиры уничтожены штаб-квартиры ИГИЛ и «Ан-Нусры»
 - Гройсман: парламентская коалиция Рады продолжает существовать
 - Власти Турции: Теракт в Стамбуле совершил боевик ИГИЛ
 - СМИ: Турция планирует расширить штат полиции на 15 тысяч человек
 - ООН направила гуманитарную помощь для жителей сирийского Блудана
 - Настоящее лицо УПА
 - Каир надеется на возобновление авиасообщения с Россией к лету
 - Переломная битва: наступление сирийской армии на Пальмиру

Примеры

- В Госдуме хотят ввести санкции против Хиллари Клинтон
 - © АР Ковалев: РФ вновь поднимет в ПА ОБСЕ вопрос санкциях против депутатов
 - Али Хаменеи возлагает на США вину за проблемы Ирана в банковской сфере
 - На место крушения Boeing выдвинулась мобильная группировка СКРЦ МЧС
 - Сирийский политолог: будущее Сирии во многом зависит от решений России
 - СМИ: В Грузии арестованы обвиняемые по делу о скандальных видеозаписях
 - Календарь событий 19марта – 20 апреля (Часть 2)
 - Календарь событий 19марта – 20 апреля (Часть 1)
 - Лидера рязанского отделения ПАРНАСа приговорили к обязательным работам
 - Андрей Воробьев попал в топ-15 самых цитируемых блогеров РФ
 - Сирийская операция. Итоги

Что дальше?

- Подготовка более хорошего датасета с большим количеством метаинформации (уже сделано).
- Адаптация алгоритмов к новому формату данных.
- Реализация модели WTMF-G.
- Проба word2vec к искомой задаче.

Автоматическое установление связей между сообщениями твиттера и новостными статьями

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