1. Описание реализации

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

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

1.1.1. 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.

1.1.2. 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.

1.1.3. WTMF

WTMF - модель применяемая для анализа схожести между короткими текстами [6]. Модель рассматривает отсутствующие в тексте слова как признаки короткого текста. Отсутсвующие 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 $K \times N$.

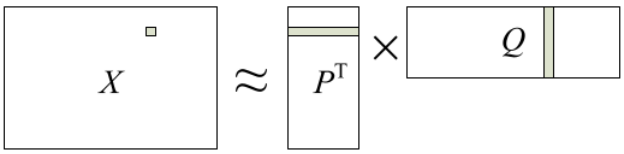

Рисунок 1 — wtmf

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 (λ - 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]):

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

$$Q_{\cdot,j} = (PW_j'P^T + \lambda I)^{-1}PW_j'X_{j,\cdot}.$$

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

Определим матрицу W следующим образом:

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

где w_m положительно и $w_m << 1$.

1.1.4. Построение связей текст-текст

Твиты связываются с помощью хэштегов, named entities и времени.

Связь твитов с помощью хэштэгов. Сначала извлекаем все хэштеги из твитов, затем превращаем в хэштеги все слова во всех твитах, которые совпали с ранее извлечёнными хэштэгами. Для каждого твита и для каждого хэштэга 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, расчитываемой для векторов из таблицы X.

Новости связываются только по времени.

1.1.5. 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:

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

коэффициент δ 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:

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

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

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.

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.