Opinion Integration Through Semi-supervised Topic Modeling

А.В. Фадеева

НИУ ВШЭ ФКН

Москва, 2017

План

- Opinion Mining
- Задача Opinion Integration
- Модель PLSA
- Semi-supervised PLSA
- Эксперименты и применение

- **Жантры текста:** отзывы, твиты [1], новости и тд

Альянс ХДС/ХСС канцлера Германии Ангелы Меркель вполне ожидаемо победил на выборах в бундестаг. Однако эксперты запугивают, что итоги выборов угрожают изменить сложившуюся после 1949 года политическую систему страны. С чего бы это?

Банкомат @sberbank при попытке внести кэш сожрал и деньги, и карту. Обещали разобраться "в приоритетном порядке за 1-2 дня". Шестой день уж)

[1] Лукашевич Н. В., Рубцова Ю. В. Объектно-ориентированный анализ твитов по тональности: результаты и проблемы. 2015. С. 499-500

Opinion Mining

- **Эксплицитные оценки**
  - словари оценочной лексики [2]
- **Имплицитные оценки**
  - выделение аспектов, fact extraction, topic models, ...

[2] http://linis-crowd.org

Opinion Mining

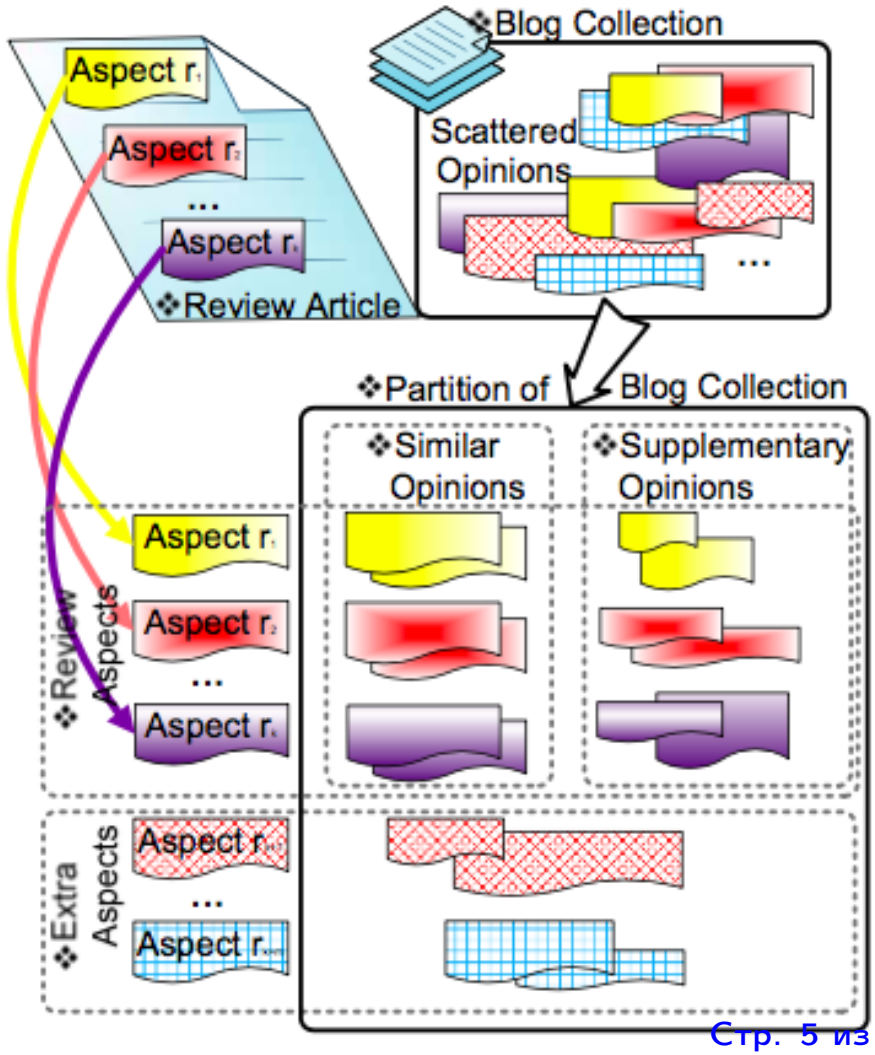
- **Эксплицитные оценки**
  - словари оценочной лексики [2]
- **Имплицитные оценки**
  - выделение аспектов, fact extraction, topic models, ...

“В супе плавала муха.”
“Памяти могло бы быть больше.”

[2] http://linis-crowd.org

Opinion Integration

Задача: выделить мнения из текстов и сопоставить их с экспертным обзором.



Opinion Integration

Дано: *экспертный обзор* по теме T $R = \{r_1, \dots r_k\}$, где r_i – это сегмент
текстовая коллекция $D = \{d_1, \dots d_{|D|}\}$, $d_i = \{s_{i1}, \dots s_{i|d_i|}\}$ где s_{ij} – это сегмент
Найти: *Integrated Opinion Summary* $(R, S^{sim}, S^{supp}, S^{extra})$
$S^{sim} = \{S_1^{sim}, \dots S_k^{sim}\}$ – сегменты документов, наложенные на обзор и схожие с экспертными;
$S^{supp} = \{S_1^{supp}, \dots S_k^{supp}\}$ – дополняющие сегменты;
$S^{extra} = \{S_1, \dots S_m\}$ – дополнительные группы сегментов из документов, которые не представлены в обзоре;
$S_i \subseteq S$, где S –это набор всех сегментов.

Opinion Integration подзадачи

1. Поиск сегментов S_0 релевантных теме T
   - разведывательный поиск по запросу темы T [3]
2. Наложить сегменты документов на сегменты экспертного обзора
3. Разделить сегменты внутри групп на схожие и дополняющие сегмент экспертного обзора

[3] C. Zhai and J. Laferty. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of CIKM 2001, pages 403–410.

Тема – вероятностное распределение $p(w|t)$ термина w в теме t.

Дано: коллекция текстов:n_{dw} сколько раз термин w встречается в документе d

Найти: параметры ϕ и θ $p(w|d) = \sum_t \phi_{wt}\theta_{td}$

$\phi_{wt} = p(w|t)$ – вероятность термина w в теме t

$\theta_{td} = p(t|d)$ – вероятность темы t в документе d

ML:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

ограничения:

$$\phi_{wt} \geqslant 0 \quad \sum_w \phi_{wt} = 1 \quad \theta_{td} \geqslant 0 \quad \sum_d \theta_{td} = 1$$

PLSA:

$$\ln p(S_0|\Phi,\Theta) = \sum_{s\in S_0}\sum_{w\in s} n_{sw} \ln \sum_{j=1}^{k+m} \phi_{wt_j}\theta_{t_j s} \to \max_{\Phi,\Theta}$$

Semi-supervised PLSA: Введем prior распределение для параметров ϕ_j на основе распределения слов в сегменте экспертного обзора r_j: $Dir(\{\sigma_j p(w|r_j)\}w \in V)$. Для дополнительных тем $j > k$ $\sigma_j = 0$.
MAP: $p(S_0|\Phi,\Theta)p(\Phi) \to \max_{\Phi,\Theta}$

$$\ln p(S_0|\Phi,\Theta) + \ln p(\Phi) \to \max_{\Phi,\Theta}$$

$\ln p(\Phi) \propto \sum_{j=1}^{k+m}\sum_{w\in W} \sigma_j p(w|r_j) \ln \phi_{wt_j}$

Semi-supervised PLSA

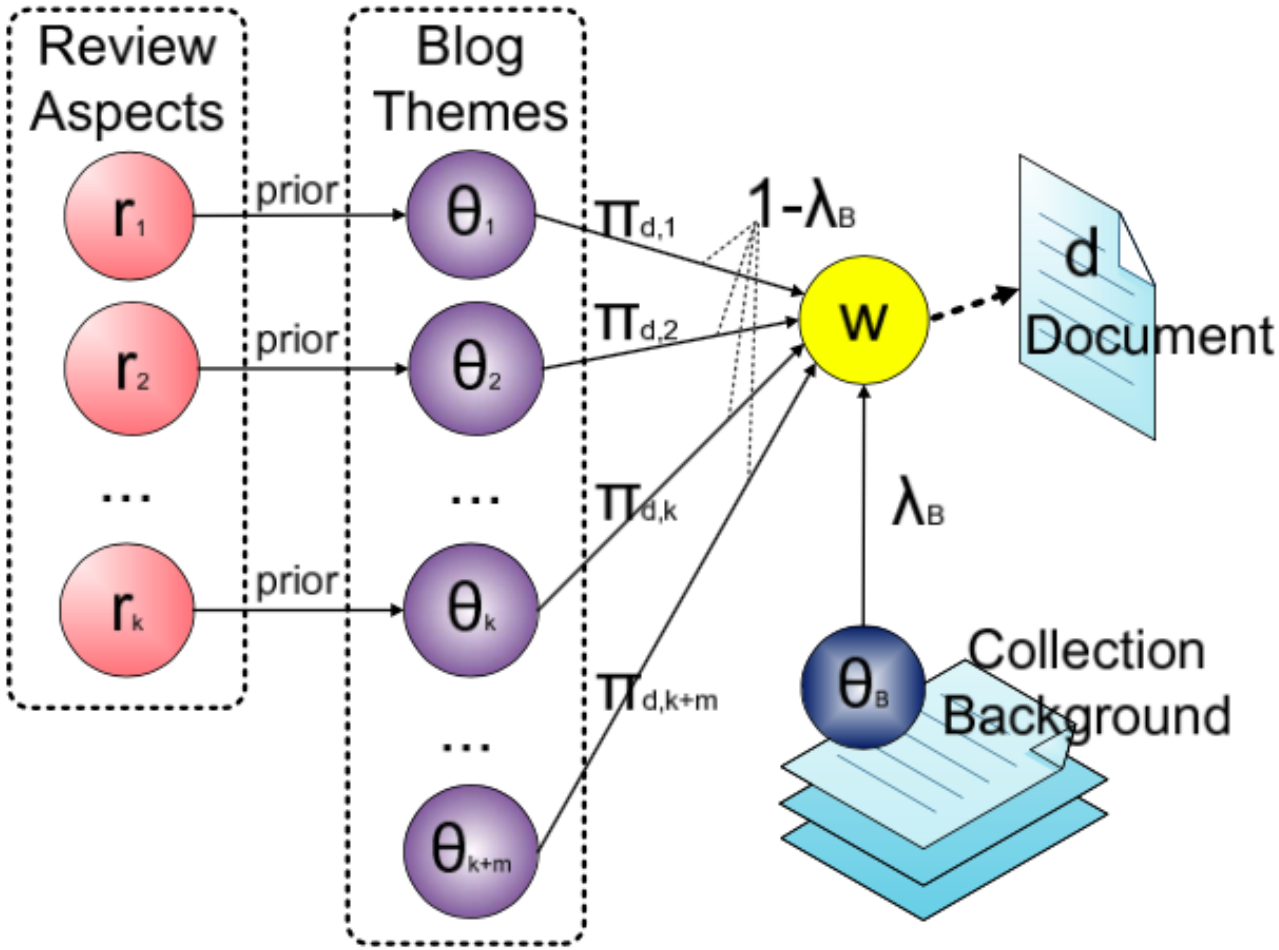


Figure 2: Generation Process of a Word

Opinion Integration подзадачи

1. Поиск сегментов S_0 релевантных теме T
   - разведывательный поиск по запросу темы T
2. Наложить сегменты документов на сегменты экспертного обзора
   - Semi-supervised PLSA на $m + k$ тем с prior на существительных
3. Разделить сегменты внутри групп на схожие и дополняющие сегмент экспертного обзора
   - Semi-supervised PLSA на две темы с prior на всех словах

Эксперименты

- Качественные эксперименты
  - обзор нового IPhone
  - статья из википедии о Бараке Обаме
- Количественные эксперименты: 34 предложения и 3 ассесора

Эксперименты

Aspect	Review	Similar Opinions	Supplementary Opinions
Background	Even with the new $399 price for the 8GB model (down from an original price of $599), it's still a lot to ask for a phone that lacks so many features and locks you into an iPhone-specific two-year contract with AT&T.		[**support=19**]The iPhone will come in two versions, a 4GB 499 model, and an 8GB 599 model with a two year contract.
			[**support=16**]The Price: 499 (4GB) or 599(8GB) with a two year contract , by the time the contract is over your iPhone will probably be scratched all over like the Nano or be made obsolete by better phone on the market.
			[**support=12**]Recently, Apple decided to cut down price of iPhone from 399 to 200 , giving rise to much rage from consumers bought the phone before.
Activation	You can make emergency calls, but you can't use any other functions, including the iPod music player.		[**support=10**]Several other methods for unlocking the iPhone have emerged on the Internet in the past few weeks, although they involve tinkering with the iPhone hardware or more complicated ways of bypassing the protections for AT T's exclusivity.
Battery	Battery life The Apple iPhone has a rated battery life of 8 hours talk time, 24 hours of music playback, 7 hours of video playback, and 6 hours on Internet use.	[**support=19**] iPhone will Feature Up to 8 Hours of Talk Time, 6 Hours of Internet Use, 7 Hours of Video Playback or 24 Hours of Audio Playback	[**support=7**]Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly 9 freaking hours of continuous playback with cell and WiFi on (but Bluetooth off).

Table 3: iPhone Example: Opinion Integration with Review Aspects

Применение

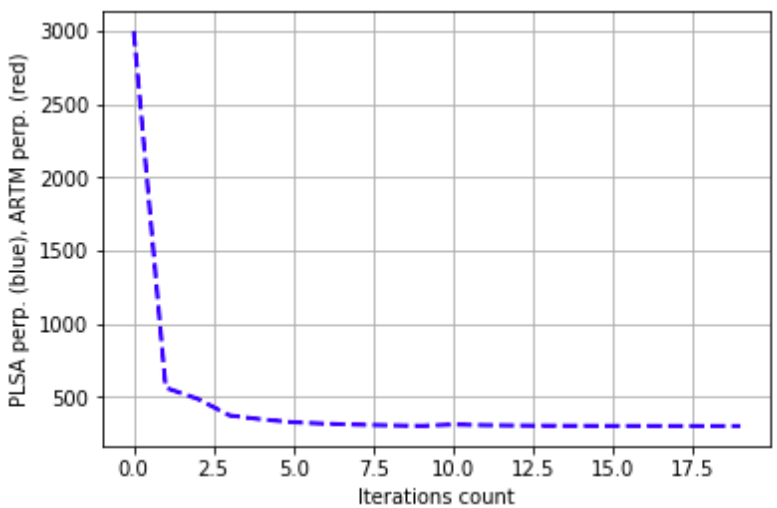
Задача: сформировать тестовую выборку для алгоритма opinion mining
Дано: по одному тексту для каждого мнения

- Достоинства
  - Сильно уменьшает трудозатраты на разметку
  - При изменении данных модель не нужно изменять
  - Выделение мнений из дополнительных тем
- Недостатки
  - Подбор параметров регуляризации [4]

[4] http://bigartm.org

Перплексия: мера несоотвествия модели $p(w|d)$ наблюдаемым терминам w в текстовой коллекции

$$P(D,p) = e^{-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)}$$



Список литературы

- Bing Liu. Sentiment Analysis and Opinion Mining, Morgan Claypool Publishers, May 2012.
- Yue Lu, Chengxiang Zhai Opinion Integration Through Semi-supervised Topic Modeling. 2008. In Proceedings of the 17th international conference on World Wide Web
- Воронцов К. В., Потапенко А. А. Регуляризация робастность и разреженность веротностных тематических моделей. Компьютерные исследования и моделирование. – 2012. – Т4 – С. 693-706
- R. Balasubramanyan, W. Cohen, D. Pierce, and D. Redlawsk. Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News? ICWSM, The AAAI Press, (2012)
- Лукашевич Н. В., Рубцова Ю. В. Объектно-ориентированный анализ твитов по тональности: результаты и проблемы. 2015. С. 499-500