Тематический анализ сообществ в социальных сетях

Бусыгина Мария, 461 группа

Научный руководитель:
к. ф.-м.н. Д.Ю. Бугайченко

Рецензент: Дзюба А.А.

2014

Тематическое моделирование

- Неструктурированные текстовые данные
- Представление в пространстве тем
- Тема — нечеткий кластер семантически связанных терминов

Подходы:

- Векторное представление
- Латентно-семантический анализ
- Вероятностное моделирование
  - Вероятностный латентно-семантический анализ
  - Скрытое размещение Дирихле

Тематическое моделирование

- Неструктурированные текстовые данные
- Представление в пространстве тем
- Тема — нечеткий кластер семантически связанных терминов

Подходы:

- Векторное представление
- Латентно-семантический анализ
- Вероятностное моделирование
  - Вероятностный латентно-семантический анализ
  - Скрытое размещение Дирихле

Постановка задачи

Цель работы: определение тем сообществ социальных сетей с помощью методов вероятностного тематического моделирования на примере постов из сообществ социальной сети «Одноклассники»

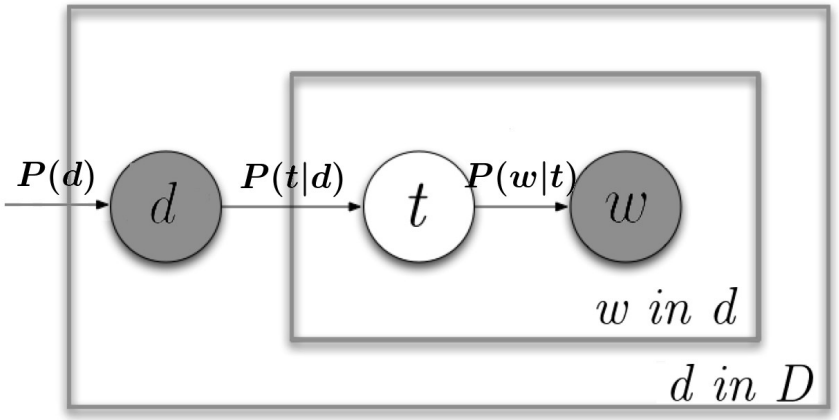
Для этого необходимо решить следующие подзадачи:

1. Исследовать существующие методы тематического моделирования
2. Реализовать алгоритм построения тематической модели
3. Оценить качество полученной тематической модели на тестовых данных

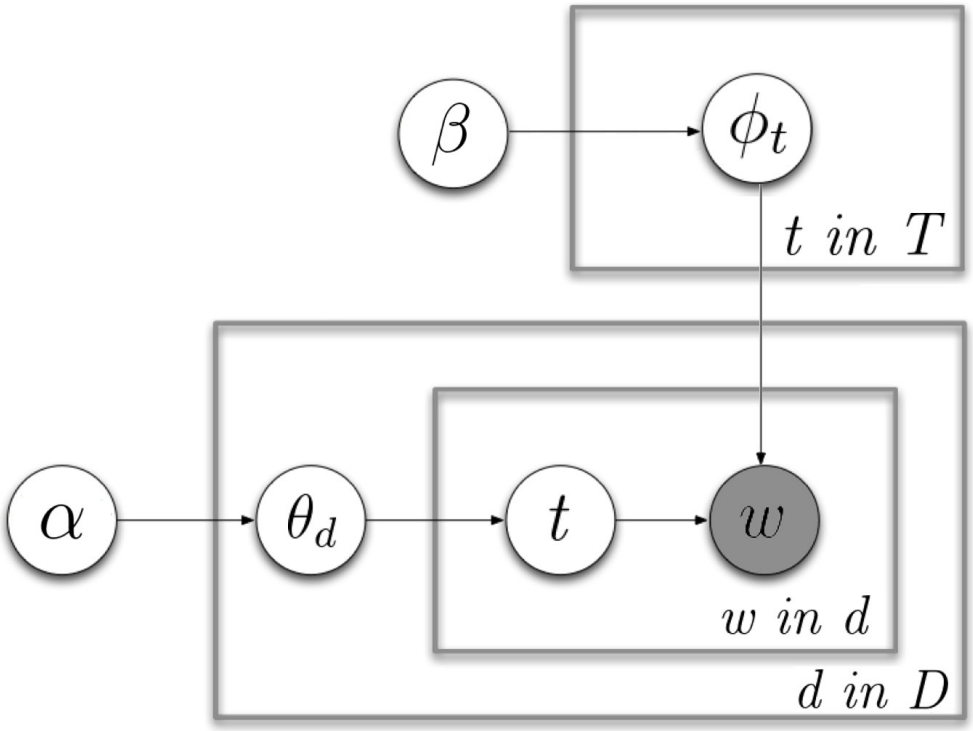
Вероятностное моделирование

Вероятностный латентно-семантический анализ (Probabilistic latent semantic analysis, PLSA)

$$P(d,w) = \sum_{t \in T} P(d)P(t|d)P(w|t)$$



Латентное размещение Дирихле (Latent Dirichlet Allocation, LDA)



EM-алгоритм

- E-шаг: $$P(t_k|d_i,w_j) = \frac{P(w_j|t_k)P(t_k|d_i)}{\sum_{l=1}^{K} P(w_j|t_l)P(t_l|d_i)}$$

- M-шаг: $$P(w_j|t_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(t_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(t_k|d_i,w_m)}$$

$$P(t_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(t_k|d_i,w_j)}{n(d_i)}$$

Решение задачи

- определение языка
- удаление форматирования, идентификаторов, временных отметок, стоп-слов...
- стемминг

- вероятностный латентно-сематический анализ
- шум и фон
- латентное размещение Дирихле

- ручная
- перплексия
- использование в работе приложений - прогноз количества «Классов!»

Темы сообществ

ДТП на федеральной трассе около Марьяновки: 4 человека в больнице

---

22 августа на трассе Омск-Исилькуль, на 64-м километре, близ деревни Татьяновка Марьяновского района столкнулись «Хюндай» и «Тойота». На место ДТП выехала скорая помощь, аварийная служба, два экипажа ДПС. - Водитель «Хюндая» двигался в сторону Омска...

Водитель скрылся с места ДТП, в котором погиб житель Марьяновского района. Вчера на обочине автодороги Челябинск — Новосибирск у села Лузино был обнаружен труп мужчины 35–40 лет, погибшего, предположительно, вследствие наезда транспортного средства...

Агрессивный водитель арестован на 15 суток. Мужчина за рулём «копейки» устроил в Марьяновке гонку с преследованием. Сотрудники экипажа ДПС, находясь на автодороге Марьяновка-Любино, обратили внимание на автомобиль «ВАЗ-2101», водитель которого не был пристегнут ремнем безопасности. ..

Пожар в церкви д. Орловка, Марьяновский район 15 сентября, в 12 часов 59 минут местного времени, на пульт диспетчера "01" поступило сообщение о пожаре в здании церкви. На момент прибытия первого подразделения происходило пламенное горение внутри закрытой церкви, обрушение кровли по всей площади. В результате пожара повреждены сгораемые конструкции...

В селе Новая Шараповка Марьяновского района ночью сгорел сарай у фермера и погибло в общей сложности 160 животных.

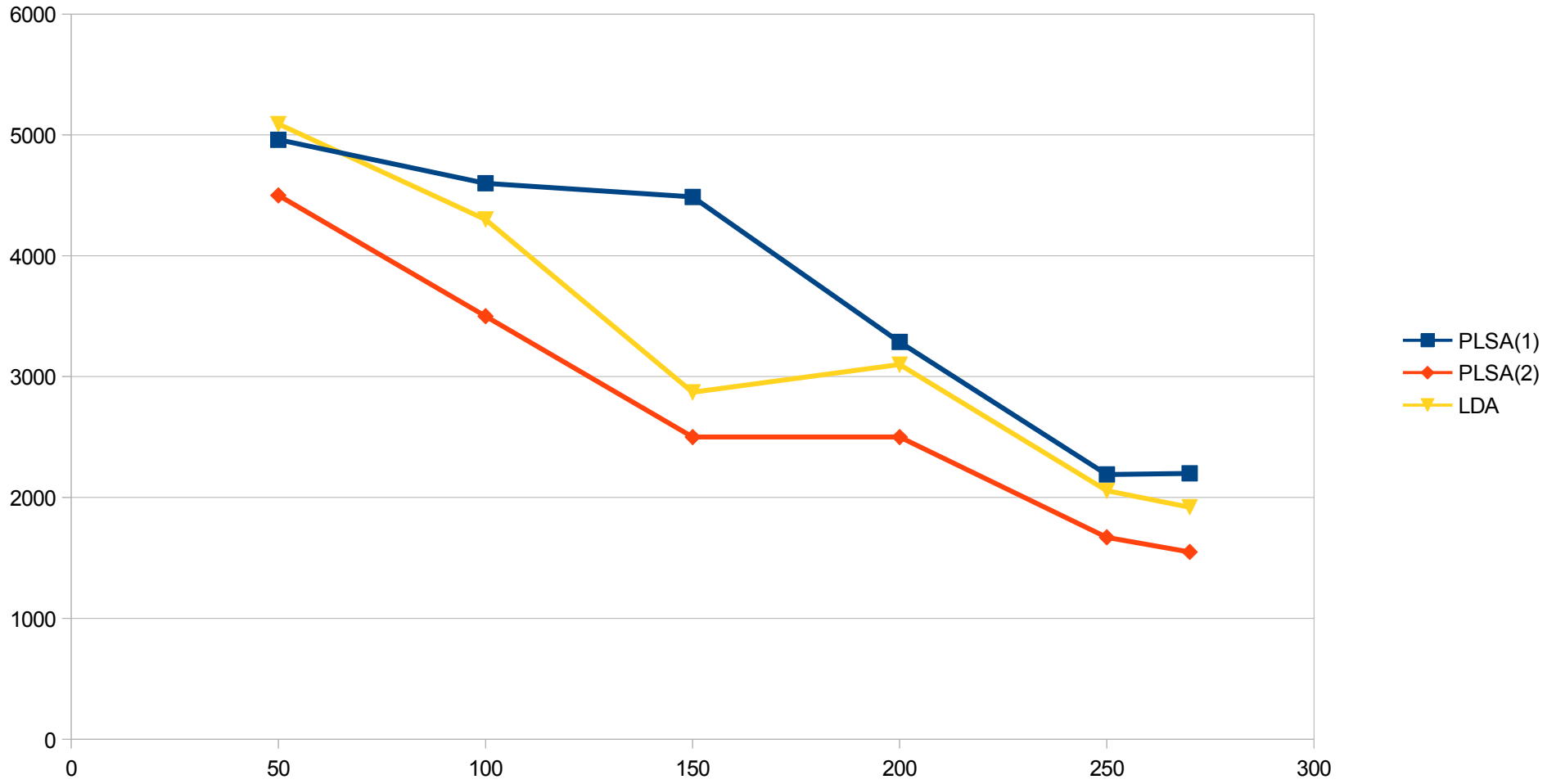
В селе Заря Свободы Марьяновского района горел неэксплуатируемый склад. В результате огонь уничтожил 1 тысячу 470 квадратных метров кровли здания. Причина пожара и материальный ущерб еще устанавливаются.

автомобиль
водитель
ДТП
трасса
километр
автодорога
наезд
транспортное
руль
столкнулись
ДПС

пожар
горел
сгорел
уничтожил
огонь
горение
ущерб
сгорание
пламенное

Перплексия

$$Perplexity = exp(-\frac{1}{n}\sum_{d \in D}\sum_{w \in d} n_{dw} ln(p(w|d)))$$



Прогноз количества «Классов»

R^2: $score(f,p) = (1 - \frac{Var(f|p)}{Var(f)}) * 1000$

features	R^2
PLSA(1)	233,649
PLSA(2)	259,630
LDA	220,348
количество слов, количество символов, количество знаков, препинания, среднее количество символов в слове	153,334

Результаты

- Проанализированы основные методы тематического моделирования
- Реализованы алгоритмы построения тематической модели для больших объемов данных
- Проведен сравнительный анализ качества тематических моделей