



Автоматическая обработка текстов на естественном языке. Тематическое моделирование.

Евгений Борисов

NLP Тематическое моделирование

схема системы обработки текстов

подбор текстов для обучения

извлечение признаков из текста

обучение модели ML

тестирование результата

NLP Тематическое моделирование

извлечение признаков из текста

предварительная очистка

токенизация

дополнительная очистка

составление словаря

частотный анализ текстов по словарю

(bag of words, BoW)

NLP Тематическое моделирование

извлечение признаков из текста

очистка

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)

«смайлики» - отдельное слово

преобразование чисел, интернет ссылок и т.п.

лемматизация - приведение слов к нормальному виду

или

стеминг - выделение основ слов

Законодательная дума Хабаровского края (duma.khv.ru) Состоялось очередное заседание Думы На последнем перед каникулами очередном заседании Законодательной Думы Хабаровского края, состоявшемся 28

```
['законодательн',  
'дум',  
'хабаровск',  
'кра',  
'url',  
'состоя',  
'очередн',  
'заседан',  
'дум',  
'последн',  
'перед',  
'каникул',  
'очередн',  
'заседан',  
'законодательн',  
'дум',  
'хабаровск',  
'кра',  
'состоя',  
'digit',
```

NLP Тематическое моделирование

извлечение признаков из текста
составление словаря

из очищенного текста извлекаем словарь

```
[  
  'digit',  
  'url',  
  'администрац',  
  'большинств',  
  'бурн',  
  'бюджетн',  
  'верхнебуреинск',  
  'власт',  
  'возьмет',  
  'войдет',  
  'вопрос',  
  'врем',  
  'втор',  
  'вызва',  
  'год',  
  ...  
]
```

NLP Тематическое моделирование

извлечение признаков из текста

частотный анализ текстов по словарю

простой частотный анализ
считаем в тексте t количество повторов x_i
каждого слова v_i из словаря V

текст должен содержать слова в достаточном количестве

NLP Тематическое моделирование

извлечение признаков из текста

частотный анализ текстов по словарю

простой частотный анализ
считаем в тексте t количество повторов x_i
каждого слова v_i из словаря V

значения x зависят от размера текста t ,
чем больше текст тем больше повторов

нормализованный частотный анализ (TF, term frequency)
значения частоты x делятся на общее число слов в тексте t .

$$TF(t, V) = x(t, V) / \text{size}(t)$$

NLP Тематическое моделирование

извлечение признаков из текста частотный анализ текстов по словарю

Удалять часто употребляемые слова или нет?

TF-IDF - компромиссный вариант формирования вектор-признаков.

не выбрасывает часто употребляемые слова из словаря
но уменьшает их вес в вектор-признаке

коэффициент обратной частоты (IDF, inverse document frequency)
чем чаще встречается слово тем меньше значение его IDF

$$IDF(v) = \log \text{size}(T) / \text{size}(T(v))$$

количество текстов T
разделить на
количество текстов T содержащих слово v

$$TF-IDF(t, T, v) = TF(t, v) * IDF(v, T)$$

NLP Тематическое моделирование

Тематическое моделирование

автоматическое извлечение тем из набора текстов

наборы ключевых слов

NLP Тематическое моделирование

Тематическое моделирование

W - конечное множество слов

D - конечное множество документов

T - конечное множество тем

NLP Тематическое моделирование

Тематическое моделирование

W - конечное множество слов

D - конечное множество документов

T - конечное множество тем

слово w в документе d связано с темой t

$D \times W \times T$ - дискретное вероятностное пространство

порядок слов в документе не важен

d, w - наблюдаемые, t - скрытая

NLP Тематическое моделирование

Тематическое моделирование

W - конечное множество слов

D - конечное множество документов

T - конечное множество тем

слово w в документе d связано с темой t

$D \times W \times T$ - дискретное вероятностное пространство

порядок слов в документе не важен

d, w - наблюдаемые, t - скрытая

гипотеза независимости $p(w|d,t)=p(w|t)$

NLP Тематическое моделирование

Тематическое моделирование

W - конечное множество слов

D - конечное множество документов

T - конечное множество тем

слово w в документе d связано с темой t

$D \times W \times T$ - дискретное вероятностное пространство

порядок слов в документе не важен

d, w - наблюдаемые, t - скрытая

гипотеза независимости $p(w|d,t)=p(w|t)$

гипотеза разреженности - документ d и термин w связаны с небольшим числом тем t , значительная часть вероятностей $p(t|d)$ и $p(w|t)$ должна обращаться в нуль.

NLP Тематическое моделирование

Тематическое моделирование

слово w в документе d связано с темой t

d, w - наблюдаемые, t - скрытая

гипотеза независимости $p(w|d,t)=p(w|t)$

гипотеза разреженности - документ d и термин w связаны с небольшим числом тем t , значительная часть вероятностей $p(t|d)$ и $p(w|t)$ должна обращаться в нуль.

тематическая модель:

$$p(w|d) = \sum_t p(w|t)p(t|d)$$

NLP Тематическое моделирование

Тематическое моделирование

частотный анализ

матрица частот употребления слова w в документе d

[слова \times документы]

вероятность $p(w|d)$ "слово w принадлежит документу d ".
можно оценивать как частоту

NLP Тематическое моделирование

Тематическое моделирование

разложение частотной матрицы

$$[\text{слова} \times \text{документы}] = [\text{слова} \times \text{темы}] * [\text{темы} \times \text{документы}]$$

$$p(w|d) = p(w|t) \cdot p(t|d)$$

$p(w|t)$ - матрица с описанием тем [слова x темы]
или оценки вероятностей "слово w принадлежит теме t ",

$p(t|d)$ - матрица [темы x документы],
или оценки вероятностей "тема t описывает документ d ".

NLP Тематическое моделирование

Тематическое моделирование

разложение частотной матрицы

$$[\text{слова} \times \text{документы}] = [\text{слова} \times \text{темы}] * [\text{темы} \times \text{документы}]$$

$$p(w|d) = p(w|t) \cdot p(t|d)$$

задача стохастического матричного разложения

стохастическая матрица — неотрицательная, сумма любого столбца/строки = 1

методы решения

PLSA - probabilistic latent semantic analysis

LDA - latent Dirichlet allocation / латентное размещение Дирихле

NMF - non-negative matrix factorization / неотрицательная матричная факторизация

NLP Тематическое моделирование

примеры текстов

Около 18 тысяч человек покинули подконтрольные боевикам районы Алеппо. За минувшие сутки из подконтрольных боевикам районов сирийского города Алеппо было выведено около 17,971 тысячи жителей, в их числе 7,542 тысячи детей. Об этом в субботу, 10 декабря, сообщает ТАСС со ссылкой на российский Центр примирения враждующих сторон в Арабской Республике.

Лидер Радикальной партии Украины Олег Ляшко назвал Надежду Савченко госизменницей. Политик призвал лишить наводчицу мандата народного депутата "То, что сейчас чудит Савченко, – это государственная измена. За подобные действия ей надо немедленно запретить доступ к государственной тайне, отозвать из ПАСЕ и лишить мандата народного депутата Украины", – написал Ляшко на странице в Facebook.

Финальная распродажа! Chery Tiggo от 19990 руб (199,9 млн) «Китайские автомобили» объявляют финальную распродажу популярных кроссоверов Chery Tiggo FL! На автомобили в максимальной комплектации установлена специальная цена 19 990 рублей (199,9 млн). Количество автомобилей ограничено!

NLP Тематическое моделирование

Темы и ключевые слова

Тема 0: рублей млн компания компании млрд модели долларов

Тема 1: трамп сша трампа дональд президент избранный президента

Тема 2: by tut декабря фото беларуси ноября беларусь

Тема 3: дтп водитель результате мвд области происшествия аварии

Тема 4: савченко украины надежда заявила партии лидер действия

Тема 5: народов севера коренных малочисленных края фестиваль июля

Тема 6: ученые университета специалисты исследователи жизни часов человека

Тема 7: flash adobe player javascript браузер проигрывателя html5

Тема 8: россия путин рф президент заявил глава президента

Тема 9: динамо матче чемпионата очков матча лиги шахтера

NLP Тематическое моделирование

Литература

`git clone https://github.com/mechanoid5/ml_lectorium.git`

К.В. Воронцов Вероятностные тематические модели коллекций текстовых документов.

Евгений Борисов Автоматизированная обработка текстов на естественном языке, с использованием инструментов языка Python
<http://mechanoid.su/ml-text-proc.html>

Евгений Борисов О задаче определения темы текста на естественном языке
<http://mechanoid.su/ml-topic-modeling.html>

Sebastian Raschka Python Machine Learning - Packt Publishing Ltd, 2015