Лекция 16:
Автоматическая обработка текстов на естественном языке.
Метод частотного анализа.

Евгений Борисов

NLP частотный анализ

обработка текстов на естественом языке,
natural language processing (NLP)

<u>частотный анализ</u>

сортировка по заданным темам

определение авторства

определение тона текста

поиск похожих текстов

текст должен содержать слова в достаточном количестве

NLP частотный анализ

схема системы обработки текстов

подбор текстов для обучения

извлечение признаков из текста

обучение модели ML

тестирование результата

NLP частотный анализ

извлечение признаков из текста

токенизация

очистка

составление словаря

частотный анализ текстов по словарю

(bag of words, BoW)

NLP частотный анализ

извлечение признаков из текста

<u>токенизация</u>

разбиения текста на отдельные слова
и/или словосочетания

n-gram - последовательность из n слов

```
Законодательная дума Хабаровского края (duma.khv.ru)

[ 'Законодательная', 'дума', 'Хабаровского', 'края', '(duma.khv.ru)' ]
```

NLP частотный анализ

извлечение признаков из текста

<u>очистка</u>

способ очистки зависит от задачи

NLP частотный анализ

извлечение признаков из текста

<u>очистка</u>

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

NLP частотный анализ

извлечение признаков из текста

<u>очистка</u>

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)
«смайлики» - отдельное слово

NLP частотный анализ

извлечение признаков из текста

<u>очистка</u>

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.

NLP частотный анализ

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

NLP частотный анализ

извлечение признаков из текста

<u>очистка</u>

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
<u>или</u>
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 частотный анализ

извлечение признаков из текста
<u>составление словаря</u>

из очищенного текста извлекаем словарь

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

NLP частотный анализ

извлечение признаков из текста

<u>частотный анализ текстов по словарю</u>

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 частотный анализ

извлечение признаков из текста

<u>частотный анализ текстов по словарю</u>

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

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

$$TF(t,V) = x(t,V) / size(t)$$

NLP частотный анализ

извлечение признаков из текста
частотный анализ текстов по словарю

Удалять часто употребляемые слова или нет?

NLP частотный анализ

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

NLP частотный анализ

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

$$IDF(v) = \log size(T) / size(T(v))$$

количество текстов T
разделить на
количество текстов T содержащих слово v

$$TF\text{-}IDF(t,T,v) = TF(t,v) * IDF(v,T)$$

NLP частотный анализ

извлечение признаков из текста
частотный анализ текстов по словарю

хэш-векторизация

заменяем слова на их хэш ограниченной длины

сокращаем размер словаря
и число признаков

экономия ресурсов для больших датасетов

NLP частотный анализ

практическое применение

сортировка по заданным темам - классификация
собираем и размечаем тексты
чистим текст
применяем частотный анализ
обучаем классификатор
тестируем

NLP частотный анализ

практическое применение

сортировка по заданным темам - классификация
собираем и размечаем тексты
чистим текст
применяем частотный анализ
обучаем классификатор
тестируем

определение авторства - классификация
собираем и размечаем тексты
чистим текст (частота употребления предлогов - важный признак)
применяем частотный анализ
обучаем классификатор
тестируем

NLP частотный анализ

практическое применение

сортировка по заданным темам - классификация
собираем и размечаем тексты
чистим текст
применяем частотный анализ
обучаем классификатор
тестируем

определение авторства - классификация
собираем и размечаем тексты
чистим текст (частота употребления предлогов - важный признак)
применяем частотный анализ
обучаем классификатор
тестируем

поиск похожих текстов - кластеризация
собираем тексты
чистим текст
применяем частотный анализ
выполняем кластеризацию (размечаем тексты)

NLP частотный анализ

Тематическое моделирование

автоматическое извлечение тем из набора текстов

наборы ключевых слов

NLP частотный анализ

Тематическое моделирование

W - конечное множество слов
D - конечное множество документов
T - конечное множество тем

слово *w* в документе *d* связано с темой *t*

D x W x T - дискретное вероятностное пространство

порядок слов в документе не важен

d, w - наблюдаемые, *t* - скрытая

<u>гипотеза независимости</u> $p(w|d,t)=p(w|t)$

<u>гипотеза разреженности</u> - 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.

<u>тематическая модель:</u> $$p(w|d)=\sum_t p(w|t)p(t|d)$$

NLP частотный анализ

Тематическое моделирование

частотный анализ

матрица частот употребления слова w в документе d

[слова x документы]

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

NLP частотный анализ

Тематическое моделирование

разложение частотной матрицы

[слова x документы] = [слова x темы] * [темы x документы]

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

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

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

NLP частотный анализ

Тематическое моделирование

разложение частотной матрицы

*[слова x документы] = [слова x темы] * [темы x документы]*

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

задача стохастического матричного разложения

методы решения

PLSA - probabilistic latent semantic analysis

LDA - latent Dirihlet allocation / латентное размещение Дирихле

NMF - non-negative matrix factorization / неотрицательная матричная факторизация

NLP частотный анализ

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 частотный анализ

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: россии путин рф президент заявил глава президента

Тема 9: динамо матче чемпионата очков матча лиги шахтера

NLP частотный анализ

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.kiev.ua/ml-text-proc.html

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

Sebastian Raschka Python Machine Learning - Packt Publishing Ltd, 2015

NLP частотный анализ

Вопросы ?

NLP частотный анализ : практика

источники данных для экспериментов

sklearn.datasets
UCI Repository
kaggle
www.ruscorpora.ru

задание

- применить TF-IDF для сортировки текстов
- применить методы тематического моделирования на наборах текстов