Введение в Data Science
Занятие 4. Naive Bayes и классификация текстов

Николай Анохин Михаил Фирулик

23 марта 2014 г.

План занятия

Обработка текстов

Naive Bayes

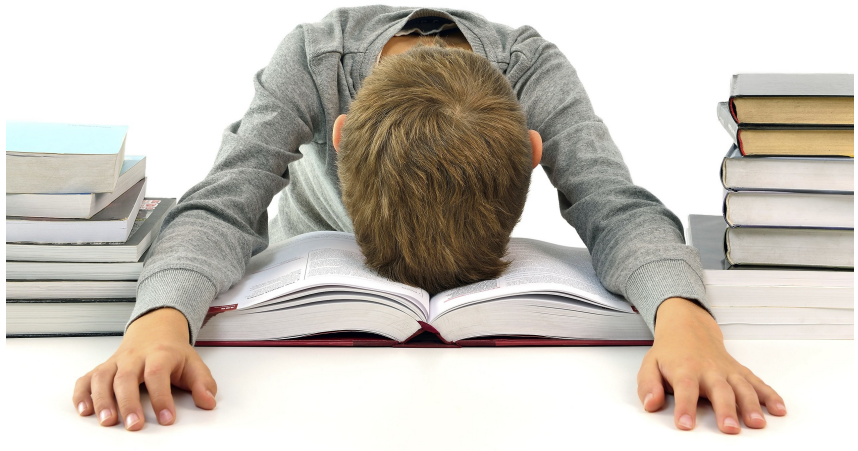
Data Mining vs Text Mining

Data Mining:
извлечение *неочевидной* информации

Text Mining:
извлечение *очевидной* информации

Трудности

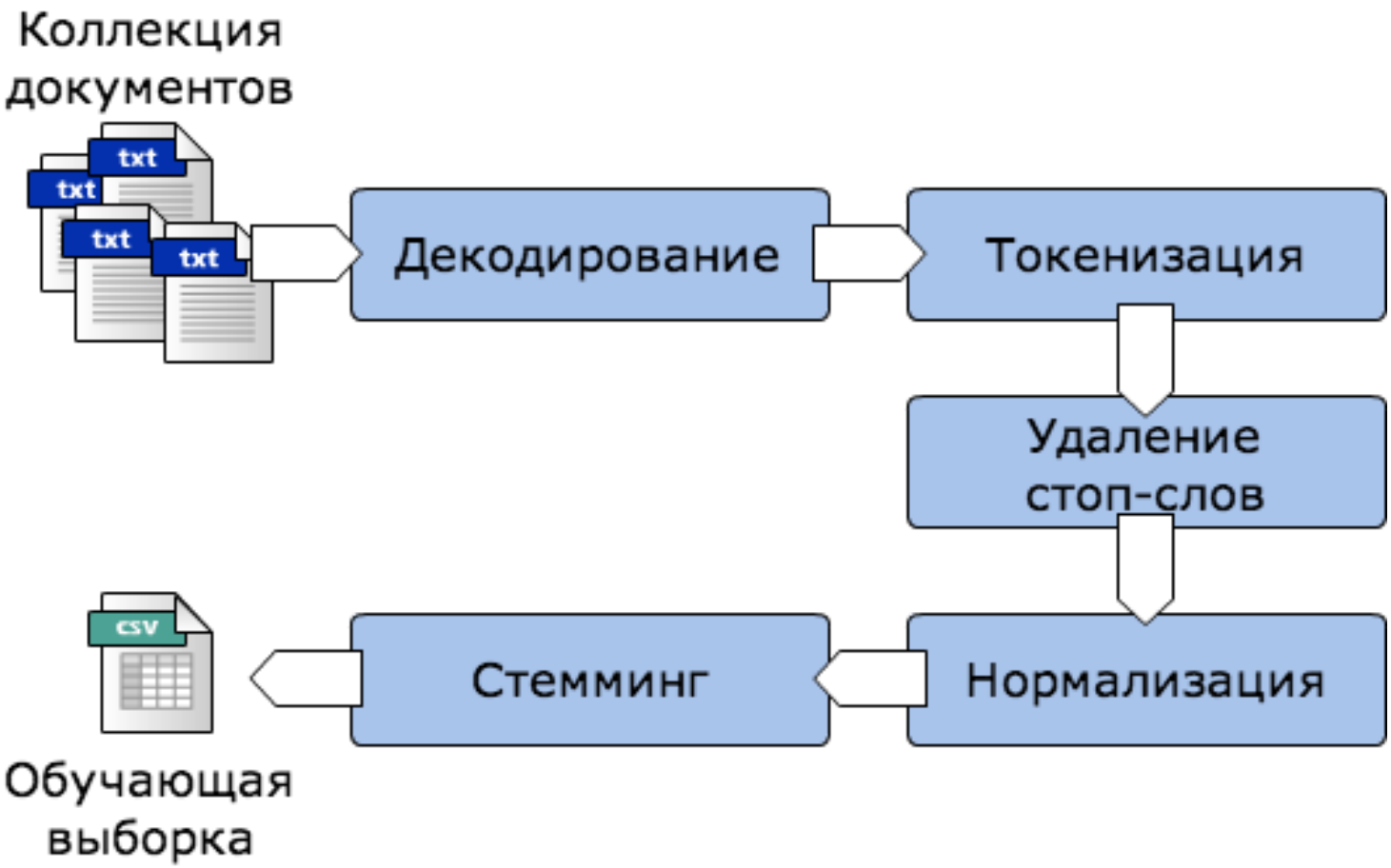
- Огромные объемы
- Отстутсвие структуры



Задачи Text Mining

- Суммаризация текста
  аггрегация новостей
- Классификация и кластеризация документов
  категоризация, фильтрация спама, эмоции
- Извлечение метаданных
  определение языка, автора, тегирование
- Выделение сущностей
  места, люди, компании, почтовые адреса

Этапы обработки текста



Декодирование

Def.

перевод последовательности байт в последовательность символов

- Распаковка
  plain/.zip/.gz/...
- Кодировка
  ASCII/utf-8/Windows-1251/...
- Формат
  csv/xml/json/doc...

Кроме того: что такое документ?

Разбиение на токены

Def.

разбиение последовательности символов на части (токены), возможно, исключая из рассмотрения некоторые символы

Наивный подход: разделить строку пробелами и выкинуть знаки препинания

> *Трисия любила* **Нью-Йорк**, *поскольку любовь к Нью-Йорку могла положительно повлиять на ее карьеру.*

Проблемы:

- n.anokhin@corp.mail.ru, 127.0.0.1
- C++, C#
- *York University* vs *New York University*
- Зависимость от языка (“Lebensversicherungsgesellschaftsangestellter”, “l’amour”)

Альтернатива: n-граммы

Разбиение на токены

```
>>> from nltk.tokenize import RegexpTokenizer
>>> tokenizer = RegexpTokenizer('\w+|[^\w\s]+')
>>> s = u'Трисия любила Нью-Йорк, поскольку любовь \
... к Нью-Йорку могла положительно повлиять на ее карьеру.'
>>> for t in tokenizer.tokenize(s)[:7]: print t + " ::",
...
Трисия :: любила :: Нью :: - :: Йорк :: , :: поскольку ::
```

Стоп-слова

Def.

Наиболее частые слова в языке, не содержащие никакой информации о содержании текста

```
>>> from nltk.corpus import stopwords
>>> for sw in stopwords.words('russian')[1:20]: print sw,
...
в во не что он на я с со как а то все она так его но да ты
```

Проблема: “To be or not to be"

Нормализация

Def.

Приведение токенов к единому виду для того, чтобы избавиться от поверхностной разницы в написании

Подходы

- сформулировать набор правил, по которым преобразуется токен

$$\textit{Нью-Йорк} \to \textit{нью-йорк} \to \textit{ньюйорк} \to \textit{ньюиорк}$$

- явно хранить связи между токенами

$$\textit{машина} \to \textit{автомобиль},\ \textit{Windows} \not\to \textit{window}$$

Нормализация

```
>>> s = u'Нью-Йорк'
>>> s1 = s.lower()
>>> print s1
нью-йорк
>>> s2 = re.sub(ur"\W", "", s1, flags=re.U)
>>> print s2
ньюйорк
>>> s3 = re.sub(ur"й", u"и", s2, flags=re.U)
>>> print s3
ньюиорк
```

Стемминг и Лемматизация

Def.

Приведение грамматических форм слова и однокоренных слов к единой основе (lemma):

- Stemming – с помощью простых эвристических правил
  - Porter (1980)
    5 этапов, на каждом применяется набор правил, таких как

$$sses \rightarrow ss \quad (\text{caresses} \rightarrow \text{caress})$$

$$ies \rightarrow i \quad (\text{ponies} \rightarrow \text{poni})$$

  - Lovins (1968)
  - Paice (1990)
  - еще 100500
- Lemmatization – с использованием словарей и морфологического анализа

Стемминг

```
>>> from nltk.stem.snowball import PorterStemmer
>>> s = PorterStemmer()
>>> print s.stem('tokenization'); print s.stem('stemming')
token
stem
>>> from nltk.stem.snowball import RussianStemmer
>>> r = RussianStemmer()
>>> print r.stem(u'Авиация'); print r.stem(u'национальный')
авиац
национальн
```

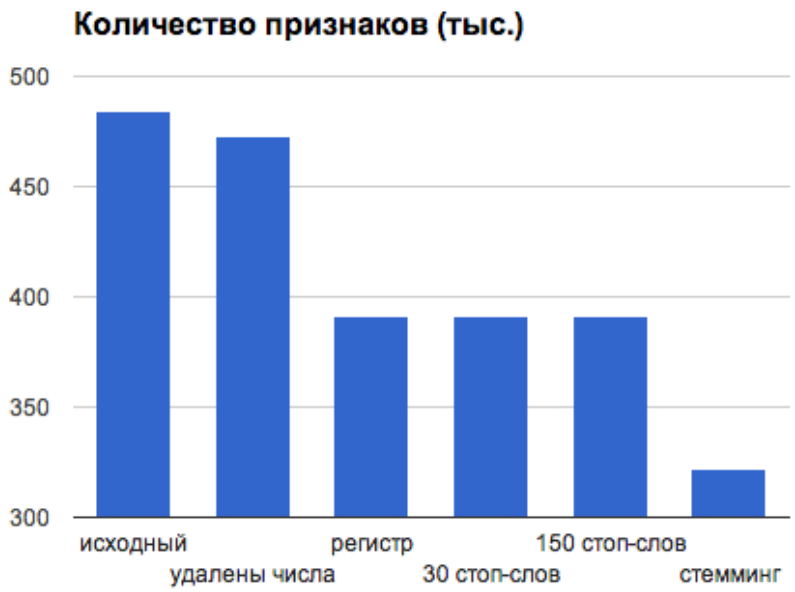
Наблюдение

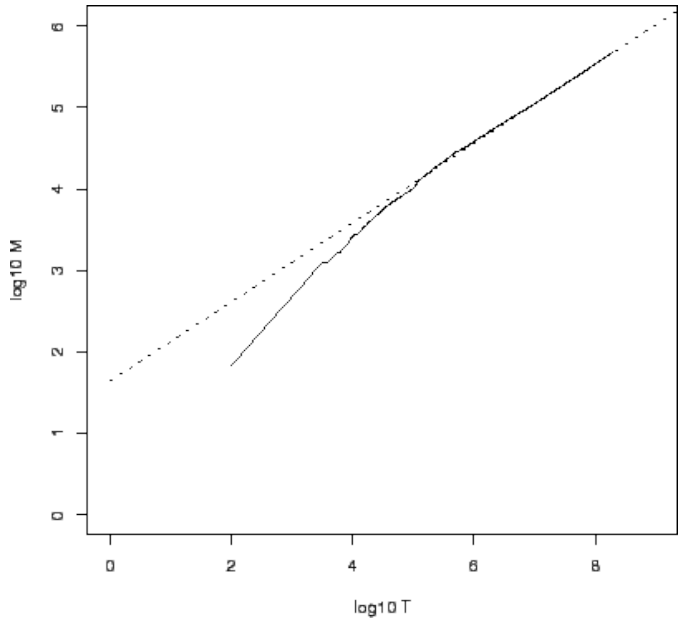
для сложных языков лучше подходит лемматизация

Heap's law

$M = kT^{\beta}$, M – размер словаря, T – количество слов в корпусе

$$30 \leq k \leq 100, \; b \approx 0.5$$





Представление документов

Boolean Model. Присутствие или отсутствие слова в документе

Bag of Words. Порядок токенов не важен

> *Погода была ужасная, принцесса была прекрасная.*
> *Или все было наоборот?*

Координаты

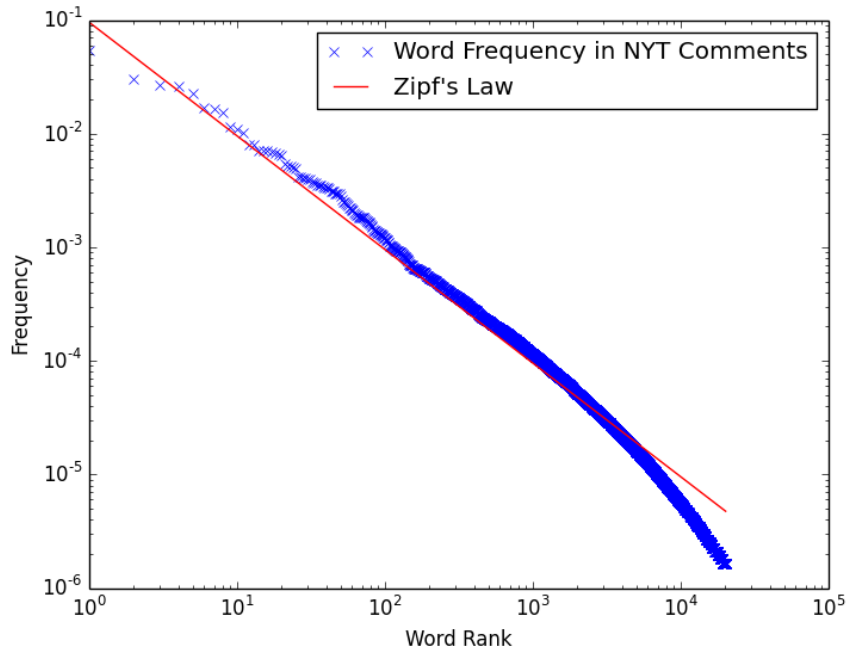
- Мультиномиальные: количество токенов в документе
- Числовые: взвешенное количество токенов в документе

Zipf's law

$t_1, \ldots, t_N$ – токены,
отранжированные по убыванию частоты
$f_1, \ldots, f_N$ – соответствующие частоты

Закон Ципфа

$$f_i = \frac{c}{i^k}$$



Что еще? Посещаемость сайтов, количество друзей, население городов...

Задача

Дана коллекция, содержащая 10^6 (не уникальных) токенов.
Предполагая, что частоты слов распределены по закону

$$f_i = \frac{c}{(i+10)^2},$$

оцените

- количество вхождений наиболее часто встречающегося слова
- количество слов, котоые встречаются минимум дважды

Подсказка: $\sum_{i=11}^{\infty} \frac{1}{i^2} \approx 0.095$

BoW & TF-IDF

Количество вхождений слова t в документе d

$$TF_{t,d} = term\text{–}frequency(t, d)$$

Количество документов из N возможных, где встречается t

$$DF_t = document\text{–}fequency(t)$$

$$IDF_t = inverse\text{–}document\text{–}frequency(t) = \log \frac{N}{DF_t}$$

TF-IDF

$$TF\text{–}IDF_{t,d} = TF_{t,d} \times IDF_t$$

Пример

Коллекция документов: Cersei Lannister, Tyrion Lannister
$d_1 = \{\text{cersei:1}, \text{tyrion:0}, \text{lannister:0}\}$
$d_2 = \{\text{cersei:0}, \text{tyrion:1}, \text{lannister:0}\}$

Байесовский классификатор

Дано
$\mathbf{x} \in \mathbf{X}$ – описание документа d из коллекции D
$C_k \in C,\ k = 1, \ldots, K$ – целевая переменная

Теорема Байеса

$$P(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} \propto p(\mathbf{x}|C_k)p(C_k)$$

Принцип Maximum A-Posteriori

$$C_{MAP} = \arg\max_k p(C_k|\mathbf{x})$$

Naive Bayes

X_j – токен на j-м месте в документе $\mathbf{x}$, $x_i \in V$ – слово из словаря V

Предположения

1. conditional independence

$$p(X_i = x_i, X_j = x_j | C_k) = p(X_i = x_i | C_k) p(X_i = x_i | C_k)$$

2. positional independence

$$P(X_i = x_i | C_k) = P(X_j = x_i | C_k) = P(X = x_i | C_k)$$

Получаем

$$p(\mathbf{x} | C_k) = p(X_1 = x_1, \ldots, X_{|\mathbf{x}|} = x_{|\mathbf{x}|} | C_k) = \prod_{i=1}^{|\mathbf{x}|} p(X = x_i | C_k)$$

Почему NB хорошо работает?

Корректная оценка дает правильное предсказание, но правильное предсказание *не требует* корректной оценки

Варианты NB

MAP

$$C_{MAP} = \arg\max_k \prod_{i=1}^{|\mathbf{x}|} p(X = x_i|C_k)P(C_k) =$$

$$= \arg\max_k \left[\log P(C_k) + \sum_{i=1}^{|\mathbf{x}|} \log P(X = x_i|C_k) \right]$$

Априорные вероятности

$$P(C_k) = N_{C_k}/N$$

Likelihood $p(X = x_i|C_k)$

- BernoulliNB $P(X = x_i|C_k) = D_{x_i,C_k}/D_{C_k}$, D – кол-во документов
- MultinomialNB $P(X = x_i|C_k) = T_{x_i,C_k}/T_{C_k}$, T – кол-во токенов
- GaussianNB $P(X = x_i|C_k) = \mathcal{N}(\mu_k, \sigma_k^2)$, параметры из MLE

Обучение NB

```
train_nb(D, C):
  V = словарь токенов из D
  N = количество документов в D
  for C_k ∈ C:
    N_{C_k} = количество документов класса C_k
    p(C_k) = N_{C_k} / N
    D_{C_k} = документы класса C_k
    for x_i ∈ V:
      p(X = x_i | C_k) = считаем согласно выбранному варианту
  возвращаем V, p(C_k), p(X = x_i | C_k)
```

Алгоритмическая сложность: $O(|D|\langle|\mathbf{x}|\rangle + |C||V|)$

Применение MultinomialNB

apply_nb(d, V, $p(C_k)$, $p(x_i|C_k)$, C):
  $\mathbf{x}$ = разбиваем d на токены, используя V
  for $C_k \in C$:
    $score(C_k|\mathbf{x}) \mathrel{+}= \log p(C_k)$
    for $x_i \in \mathbf{x}$:
      $score(C_k|\mathbf{x}) \mathrel{+}= \log p(x_i|C_k)$ считаем согласно выбранному варианту
  возвращаем $\arg\max score(C_k|\mathbf{x})$

Алгоритмическая сложность: $O(|C||\mathbf{x}|)$

Задача

d	Текст	Класс
1	котики такие котики	мимими
2	котики котики няшки	мимими
3	пушистые котики	мимими
4	морские котики мокрые	не мимими
5	котики котики мокрые морские котики	???

С помощью алгоритма MultinomialNB вычислить $p(\text{мимими}|d_5)$

Сглаживание

Проблема: $p(\text{пушистые}|\text{не мимими}) = 0$

Решение:

$$p(X = x_i|C_k) = \frac{T_{x_i,C_k} + \alpha}{T_{C_k} + \alpha|V|}$$

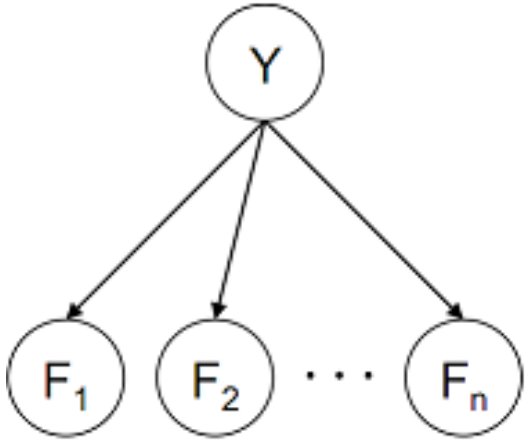
если $\alpha \geq 1$ – сглаживание Лапласа, если $0 \leq \alpha \leq 1$ – Лидстоуна

Упражнение

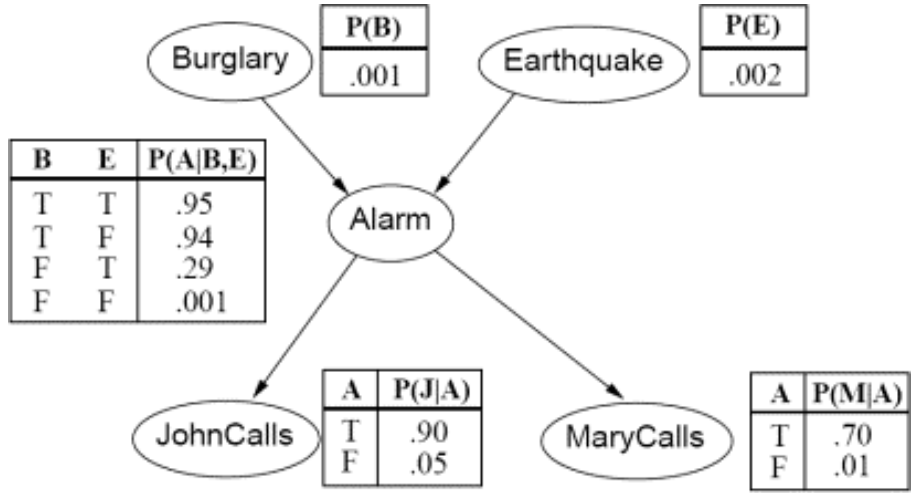
С учетом сглаживания вычислить

$$p(\text{пушистые}|\text{не мимими}),\ p(\text{пушистые}|\text{мимими}).$$

Байесовские сети



Naive Bayes



Bayes Network

Итоги

+ Генеративная модель
+ (Удивительно) неплохо работает
+ Стабилен при смещении выборки (aka concept drift)
+ Оптимальный по производительности

– Наивные предположения
– Требует отбора признаков

Определение языка текста

Определение языка на основании n-грамм

- Нормализация
  Нижний регистр, заменяем акценты на обычные буквы
- Токенизация
  Разбиваем документы на n-граммы
- Выбор признаков
  Берем топ-k признаков из каждого языка
- Инициализация модели
  Используем один из вариантов NB из sklearn
- Анализ
  Как зависит точность предсказания от n и k?

Домашнее задание 3

Байесовский классификатор

Реализовать

- алгоритм Naive Bayes для задачи классификации
- алгоритм Naive Bayes для задачи регрессии

Варианты: multinomial, bernoulli, gaussian

Ключевые даты

- До 2014/03/29 00.00 выбрать задачу и ответственного в группе
- До 2014/04/05 00.00 предоставить решение задания

Спасибо!

Обратная связь