



Блок

# FEATURE ENGINEERING



ЕГОР  
САЧКО

Lead Data Scientist

Сбербанк



egsachko@gmail.com



sachkoe



**МАТЕРИАЛЫ ПО БЛОКУ**

## МАТЕРИАЛЫ ПО БЛОКУ

1

**“Learning scikit-learn:  
Machine Learning in Python”**

Raul Garreta,  
Guillermo Moncecchi,  
2013,  
Packt

2

**“Hands-On Machine  
Learning with Scikit-Learn  
and Tensorflow:  
Concepts, Tools and  
Techniques to Build  
Intelligent Systems”**

Geron, A., 2017, O'Reilly Media

3

<https://www.analyticsvidhya.com/blog> - много интересных статей и tutorиалов

4

[blog.kaggle.com/](https://blog.kaggle.com/) -  
[No Free Hunch](#)



# ЦЕЛИ ЗАНЯТИЯ

## ЦЕЛИ ЗАНЯТИЯ

# В КОНЦЕ ЗАНЯТИЯ ВЫ СМОЖЕТЕ

# 1

Использовать  
**методы  
декомпозиции  
матриц**

# 2

Применять на  
практике **LDA**  
**модели**

# 3

Работать с  
**разряженными  
матрицами**



**ЧТО БУДЕМ ОБСУЖДАТЬ**

## ПЛАН ЗАНЯТИЯ

---

1

Работа со sparse матрицами

2

Методы разложения матриц

3

Singular Value decomposition

4

Latent Dirichlet Allocation

5

Обсуждение домашнего задания



---

# Часть 1

# **Sparse матрицы**

## Плотные матрицы

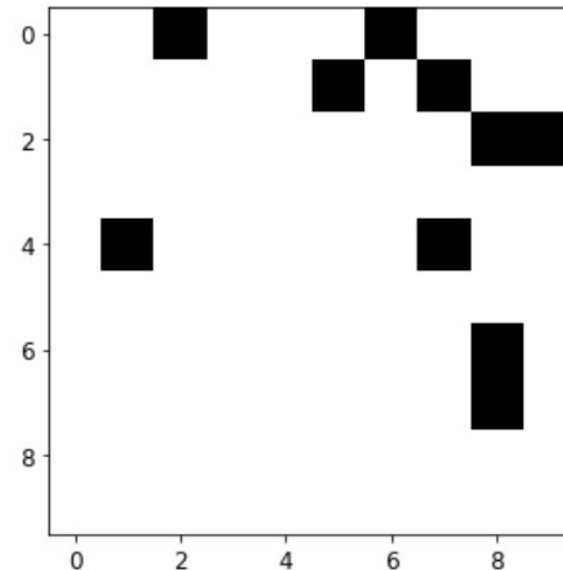
- $m$  - количество строк,  $n$  - количество столбцов
- Для хранения плотной матрицы требуется объем памяти, пропорциональный  $m \times n$
- Хранить больше плотные матрицы дорого

## Плотные матрицы

- $m$  - количество строк,  $n$  - количество столбцов
- Для хранения плотной матрицы требуется объем памяти, пропорциональный  $m \times n$
- Хранить больше плотные матрицы дорого
- ... а если 99% данных нули? невыгодно

## Главное

- Разреженная матрица - матрица, в которой большая часть элементов нули
- Черное - 1, белое - 0
- Density (= 1 - Sparsity)
- Форматы хранения:
  - DOK, LIL, COO, CSR, CSC, ...
- Требуют адаптации алгоритмов





## Экономия

- Матрица 10 x 10, density = 0.1
  - Плотная матрица: 800B
  - CSR матрица: 164B
- Матрица 1000 x 1000, density = 0.001
  - Плотная матрица: 7.8MB
  - CSR матрица: 15.6kB

## Три буквы

- DOK, LIL, COO:
  - Эффективны: для быстрого создания и изменения матриц
  - Не эффективны: для арифметических операций
- CSR, CSC:
  - Эффективны: для арифметических операций
  - Не эффективны: для быстрого создания и изменения матриц
- ... бывают исключения из правил

- Бинарная матрица 10 x 10
- Плотность: 0.09
- Количество ненулевых ячеек (NNZ): 9

[illegible]

[illegible]



```
{(0, 1): 1.0,
 (1, 1): 1.0,
 (3, 3): 1.0,
 (3, 4): 1.0,
 (4, 3): 1.0,
 (5, 0): 1.0,
 (5, 2): 1.0,
 (5, 5): 1.0,
 (6, 9): 1.0}
```

[illegible]

row_id	данные
0	[(1, 1.0)]
1	[(1, 1.0)]
2	[]
3	[(3, 1.0), (4, 1.0)]
4	[(3, 1.0)]
5	[(0, 1.0), (2, 1.0), (5, 1.0)]
6	[(9, 1.0)]
7	[]
8	[]
9	[]

[illegible]



- Хранится в виде трех 1D-массивов
- A массив ненулевых ячеек матрицы
- IA массив указателей строк
  - $IA[0] = 0$
  - $IA[i] = IA[i - 1] + (\text{NNZ на } i - 1 \text{ строке})$
- JA массив id столбцов ненулевых ячеек

[illegible]

- A: [1, 1, 1, 1, 1, 1, 1, 1, 1]
- IA: [0, 1, 2, 2, 4, 5, 8, 9, 9, 9]
- JA: [1, 1, 3, 4, 3, 0, 2, 5, 9]

[illegible]

- Аналогично CSR
- IA массив указателей столбцов
- JA массив id строк ненулевых ячеек

[illegible]

- A: [1, 1, 1, 1, 1, 1, 1, 1, 1]
- IA: [0, 1, 3, 4, 6, 7, 8, 8, 8, 8, 9]
- JA: [5, 0, 1, 5, 3, 4, 3, 5, 6]

[illegible]

---

Практика

**Разряженные матрицы в  
NumPy**



---

Часть 2

# Методы декомпозиции матриц

## Разложение матрицы

- Представление матрицы в виде произведения матриц
- Новые матрицы обладают некоторыми определёнными свойствами (ортогональность, симметричность, диагональность)
- Много способов разложения: каждый способ используется в определенном классе задач

## Singular Value Decomposition

- Матрица **M** ( $m \times n$ ) раскладывается на три новые матрицы **U**,  **$\Sigma$**  и  **$V^*$** , так что:  **$M = U\Sigma V^*$**
- **U** - матрица  $m \times m$ , левые сингулярные векторы матрицы **M**
- **V** - матрица  $n \times n$ , правые сингулярные векторы матрицы **M**
- **$V^*$**  - сопряженно транспонированная матрица к **V**
- **$\Sigma$**  - диагональная матрица  $m \times n$ , числа на диагонали - сингулярные числа

## Сингулярные числа и векторы

- **u** - вектор единичной длины размерности  $m$
- **v** - вектор единичной длины размерности  $n$
- $\sigma$  - сингулярное число  $\Leftrightarrow \mathbf{M}\mathbf{v} = \sigma\mathbf{u}$  &  $\mathbf{M}^*\mathbf{u} = \sigma\mathbf{v}$
- **u** - левый сингулярный вектор
- **v** - правый сингулярный вектор

## Low-Rank Approximation

- $k$  - количество компонент

$$M \approx M_k = U_k \Sigma_k V_k^T$$

- $\mathbf{U}$  - матрица  $m \times k$
- $\mathbf{V}^T$  - матрица  $k \times n$
- $\Sigma_k$  - диагональная матрица  $k \times k$ , числа на диагонали - первые  $k$  наибольших сингулярных чисел

## Non-Negative Matrix Factorization

- NMF - альтернативный способ матричного разложения
- Предполагает, что данные и компоненты не негативны
- Исходная матрица **M** раскладывается на две матрицы **W** и **H**, оптимизируя евклидову норму:

$$\arg \min_{W, H} \frac{1}{2} \|M - WH\|_{Fro}^2 = \frac{1}{2} \sum_{i,j} (M_{ij} - W_i H_j)$$

## Non-Negative Matrix Factorization

- $M$  - матрица  $m \times n$
- $W$  - матрица  $m \times p$
- $H$  - матрица  $p \times n$
- $p \ll \min(m, n)$
- Регуляризация

$$\arg \min_{W, H} \frac{1}{2} \|M - WH\|_{Fro}^2 + \alpha \rho \|W\|_1 + \alpha \rho \|H\|_1 + \frac{\alpha(1 - \rho)}{2} \|W\|_{Fro}^2 + \frac{\alpha(1 - \rho)}{2} \|H\|_{Fro}^2$$

---

Практика

**Декомпозиция матриц в  
sklearn**



---

Часть 3

# Latent Dirichlet Allocation

## Latent Dirichlet Allocation

- Вероятностная модель для коллекции дискретных датасетов (например, корпус текстов)
- Используется для моделирования абстрактных топиков (latent topics, latent features)
- Конечная цель: описать принадлежность документа к множеству топиков

## Latent Dirichlet Allocation

- Документ - распределение топиков (распределение Dirichlet)
- Топик - распределение слов (распределение Dirichlet)
- Документ - множество слов (наблюдаемая информация)

## Topics and terms

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

## **Классификация групп ВК**

1. На какие темы пишут в группе-паблике?
2. Что интересует пользователя, подписанного на эти паблики?
3. Какие слова (terms) наиболее характерны для этих пабликов?

# Характерные слова топиков по группам ВК

## Интересные страницы 97



Лентач  
Пропаганда здорового  
смысла



hype  
Фан-клуб одиночных  
игр



Медуза  
СМИ



Геймеры



Сериал Ведьмак |  
Netflix`s The Witcher  
Только новости и ничего  
лишнего 🔥

## Интересные страницы 47



MDK  
мемес деливери  
корпорейшн



Celebrity  
Новости. Фото.  
Премьеры. Факты.



Смейся до слёз :D  
~\(\ツ)\\_/~



E-squire  
Умный журнал для  
успешных людей!



Шедевры рекламы  
Этот день настал: мы  
открыли комменты!

## Интересные страницы 503



30 Дней Стройности с  
Ильёй Павловым.  
Когда вы думаете, что  
уже слишком поздно  
что-то начинать,



Шедевры кулинарии |  
Простые рецепты  
Собрание лучших  
рецептов 🔥



Зайка Развивайка  
Для родителей и детей



Калуга Даром  
Всё лучшее в жизни -  
Бесплатно! :D



Нужные люди Калуги  
Найди нужного  
человека! Расскажи о  
себе! БЕСПЛАТНО!

## Интересные страницы 149



AdMe.ru  
Вдохновение.  
Творчество. Позитив.



Бумажный самолётик  
Делаем ваш день!



Почему?  
Ответы на все ваши  
"почему"!



Комментатор от Бога  
Твой шанс стать мемом!



Бумажный кораблик  
Мир комиксов

## Характерные слова топиков по пабликам ВК

1. 2300 популярных пабликов
2. Из каждого паблика скачиваются последние 100 постов
3. Паблик = документ из слов, которые встретились в 100 последних постах
4. Итого получаем 2300 документов, можно обучить LDA модель
5. LDA модель из gensim:  
<https://radimrehurek.com/gensim/models/ldamodel.html>

# Характерные слова топиков по пабликам ВК

## LdaModel performance

- 0 чтоб только можно этом потом было есть всем менить очень
- 1 сознание духовный истина воля осознать истинный страдание человеческий разум общество
- 2 пацан ~~говно~~ жопа бабка короче ~~блять~~ тупой водка орать ладный
- 3 рубль телефон группа цена писать репост запись ноябрь фото звонить
- 4 джон фильм американский герой смерть режиссёр война история жанр роль
- 5 группа участие друг состояться ждать ноябрь участник конкурс победитель музыка
- 6 комплект задний цена диск колесо авто передний двигатель продать автомобиль
- 7 вещество нагрузка поверхность организм температура способствовать свойство витамин применение мышца
- 8 доставка фото який можный буде супер репост куртка тільки комплект
- 9 российский россия сотрудник государственный житель область владимир страна центр александр
- 10 человек любить жизнь знать друг твой думать жить видеть любовь
- 11 матч лига чемпион игрок чемпионат сборный соперник забить победа тренер
- 12 остров озеро северный река парк путешествие гора турист берег расположить
- 13 сайт ссылка информация возможность интересный компания количество команда уровень результат
- 14 дизайн покрытие отработка ноготок ноготь гель-лак укрепление педикюр ногтевой гель
- 15 человек парень любить девушка жизнь забирать 2017 когда атмосферноаж сериал
- 16 ингредиент соус приготовление сливочный рецепт яйцо духовка вкусно блюдо вкусный
- 17 love world music live time black life night like come
- 18 нокаут прогнозы👍 емельяненко рефери спарринг весовой нокаутировать fight конора полутяжёлый
- 19 цвета красивый цвет натуральный платье фото размер оттенок образ форма



---

Практика

**Классификация групп VK**

---

Часть 4

# Обсуждение домашнего задания

# Сравнение интересов аудитории телеканалов НТВ и Дождь с помощью тематического моделирования LDA





**ЧТО МЫ СЕГОДНЯ УЗНАЛИ**

## ИТОГИ

1

Какие существуют **методы**  
**декомпозиции матриц**

2

Как устроено **SVD**

3

Принцип работы **LDA**

4

Как работать со **sparse**  
**матрицами**

---

**СПАСИБО ЗА ВНИМАНИЕ**