



ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



# Онлайн-занятие по подготовке участников к основному этапу Всероссийской олимпиады по искусственному интеллекту 2023

5 октября 13:00





# Программа онлайн-занятия

1. **Регламент участия в основном этапе Олимпиады**
2. **Подготовка к решению заданий основного этапа**
  - Задания по математике
  - Задания по машинному обучению
  - Задания по алгоритмам



ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



# Основной этап



# Основной этап

Основной этап состоит **из 6 заданий**, которые можно решать в произвольном порядке. Итоговый рейтинг строится по сумме баллов за каждое задание:

- 2 задания по математике
- 2 задания по алгоритмам
- 2 задания по анализу данных и ML

В заданиях «Алгоритмы» и «Математика» количество баллов определяется количеством пройденных тестов.

В заданиях категории «Машинное обучение» критерием присвоения баллов является точность предсказания обученной модели.

# Основной этап



## Задания по математике

Два задания на проверку знаний по темам, непосредственно связанным с анализом данных:

- Комбинаторика
- Основы теории вероятностей
- Основы теории графов
- Базовые математические знания

## Задания по алгоритмам

Два задания по алгоритмам, требующие умения писать код. В задачах есть математическая подоплека, то есть математическую формулировку необходимо перевести в код. Задачи проходят через набор автоматических тестов.



# Основной этап

## Задания по машинному обучению

Более продвинутые задания, чем в отборочном этапе.

Для подготовки к этапу участникам рекомендуется попрактиковаться в решении задач:

- Регрессии
- Классификации
- Кластеризации
- Построения рекомендательных систем
- Компьютерного зрения
- Области автоматической обработки языка
- Поиска аномалий в данных



# Основной этап

## Языки программирования:

- Для задач по алгоритмам рекомендуются языки **C++** и **Python**
- Для задач по машинному обучению ожидается знание **Python**

## Рекомендуемые требования к рабочим местам для комфортного решения заданий отборочного этапа:

- Процессор с тактовой частотой ядра не менее 2,1ГГц и количеством ядер не менее 6,
- Оперативная память не менее 16 гб,
- Жесткий диск SSD на менее 128 гб,
- Видеокарта внешняя Nvidia с объемом видеопамати от 8 гб,
- ОС Windows,
- Монитор не менее FHD (1920\*1080), 60Hz,
- Клавиатура, мышь.

*\*Оборудование участника может отличаться в меньшую сторону от заявленного в требованиях*





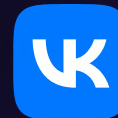
ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



# Подготовка к решению заданий по машинному обучению





# Задания по машинному обучению

В основном этапе будет **два задания**. Оба посвящены различным областям применения машинного обучения. В частности, могут встретиться задания по следующим темам:

- Анализ временных рядов,
- Построение рекомендательных систем,
- Задачи работы с изображениями,
- Задачи работы с текстами,
- Другие задачи.



# Пример темы 1: рекомендательные системы

Рекомендательная система автоматически предсказывает товары/фильмы/музыку, которые могут заинтересовать пользователя на основе:

- прошлого поведения,
- похожести на других пользователей,
- похожести товаров/фильмов/музыки,
- контекста (например: пользователь находится в поисковой выдаче по запросу "ipad").



# Основные подходы

- **Collaborative Filtering:** рекомендуем товары, основываясь на прошлом поведении пользователя и всех остальных пользователей.
- **Content-based:** рекомендации, основанные на схожести свойств товаров.
- **Matrix Factorization:** рекомендации, основанные на разложении матрицы оценок "пользователь-товар" в произведение матриц меньшей размерности.
- **Neural Networks:** рекомендации, полученные с помощью нейросетевых подходов.



ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ






# Рекомендации: Collaborative filtering

# Collaborative filtering



Рассмотрим матрицу взаимодействий «пользователь-товар»:

						
	1	1	0		1	
	0	1	1			1
				1	1	0
		1	1		0	
		1				1



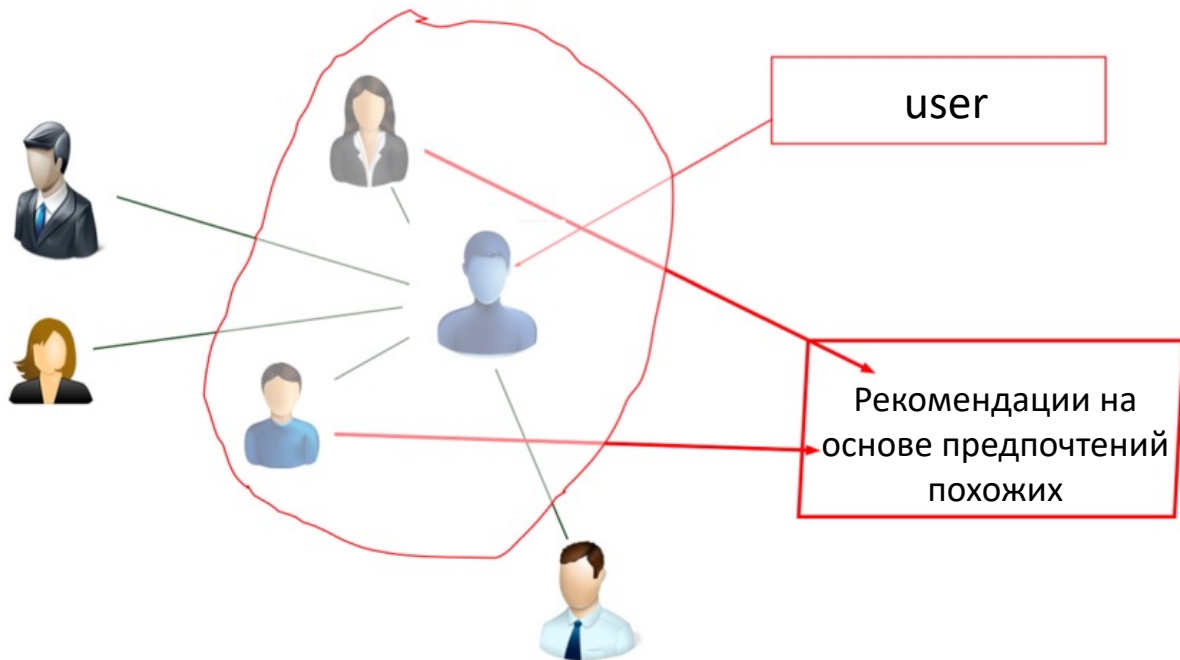
# User-based CF

Как сделать рекомендацию для пользователя user?

**Идея: найдем похожих на user пользователей и порекомендуем ему понравившиеся им товары.**

Такой подход называется user-based collaborative filtering.










# User-based CF





# Collaborative filtering

Какие фильмы рекомендовать выделенному пользователю?

						
	1	1	0		1	
	0	1	1			1
				1	1	0
		1	1		0	
		1				1

?

# Collaborative filtering

Найдем пользователей, смотревших те же фильмы:

						
	1	1	0		1	
	0	1	1			1
				1	1	0
		1	1		0	
		1				1

# Collaborative filtering




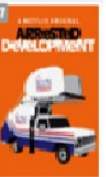


Найдем пользователей, смотревших те же фильмы:

	4 	5 	6 	7 	8 	9 
1 	1	1	0		1	
2 	0	1	1			1
3 				1	1	0
4 		1	1		0	
5 		1				1

Похожие пользователи

# Collaborative filtering

Предложим нашему пользователю фильм, который он не смотрел, но смотрели похожие на него пользователи:

	1	2	3	4	5	6
4						
1	1	1	0		1	
2	0	1	1			1
3				1	1	0
4		1	1		0	
5		1				1

Похожие пользователи



ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



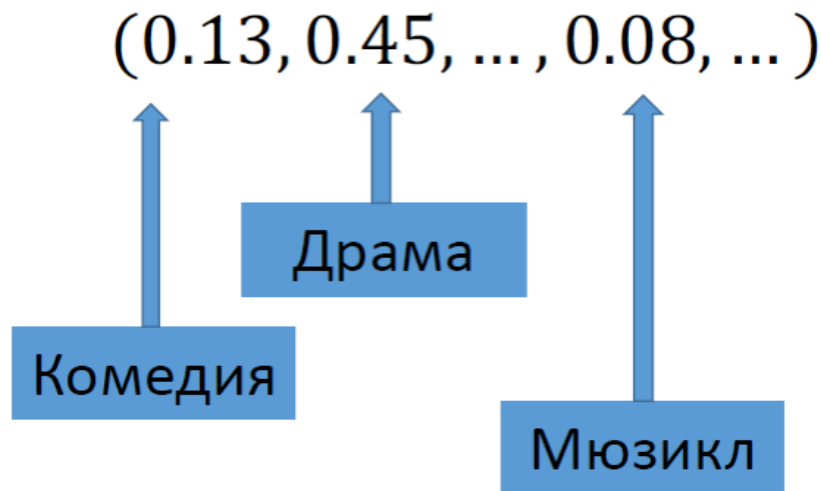
# Рекомендации: Матричные разложения

# Векторы интересов

Решаем задачу рекомендации пользователям различных фильмов.

Можно описать пользователя и фильм векторами интересов:

- для пользователя – насколько он интересуется каждым жанром,
- для фильма – насколько он относится к каждому жанру.



# Рейтинг

Будем определять *заинтересованность* как *скалярное произведение* вектора пользователя и вектора фильма:

$$(0.1, 0.5, 0.01, 0.92) \times (0, 0, 0.1, 0.95) = 0.875$$

$$(0.1, 0.5, 0.01, 0.92) \times (0.9, 0, 0, 0.1) = 0.182$$

Пользователь

Фильм



# Модели со скрытыми переменными

У нас есть матрица рейтингов для задачи «пользователь-фильм»:

2	5	
5		4
	1	
	2	5

**Цель:** найти такие векторы пользователей и векторы фильмов, скалярное произведение которых максимально близко к рейтингам из таблицы.

# Модели со скрытыми переменными

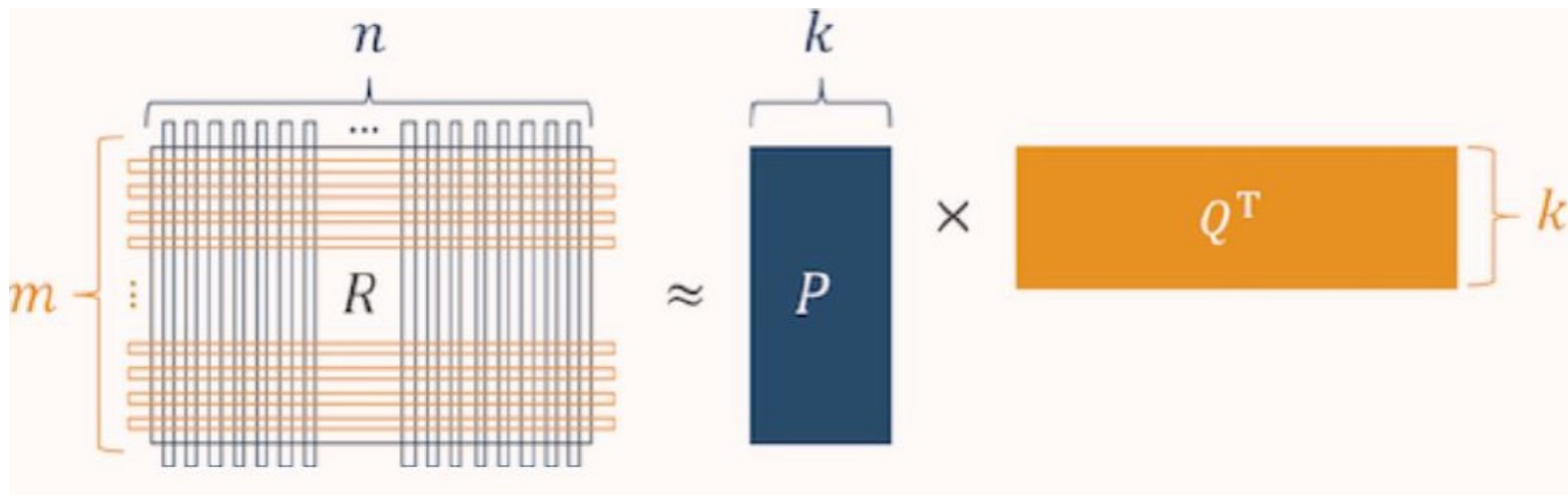
У нас есть матрица рейтингов для задачи «пользователь-фильм»:

	(0.9, 0.05)	(0.02, 1.1)	(1.05, 0.01)
(2.1, 5)	2	5	
(4.6, 0)	5		4
(0, 1)		1	
(4.9, 0.9)		1	5

**Цель:** найти такие векторы пользователей и векторы фильмов, скалярное произведение которых максимально близко к рейтингам из таблицы.

# Матричные разложения

Эту задачу можно решить с помощью матричной факторизации, а именно, **представить матрицу рейтингов как произведение двух матриц:**



- в матрице  $P$  находятся векторы интересов пользователей
- в матрице  $Q$  находятся векторы фильмов

# SVD для построения рекомендаций

Матрица товарных предпочтений (матрица, где строки - это пользователи, а столбцы - это продукты, с которыми пользователи взаимодействовали) представляется произведением трех матриц:

The diagram illustrates the SVD decomposition of a matrix  $A$ . On the left, matrix  $A$  is shown as a grid of blue vertical lines (columns) and orange horizontal lines (rows). It is labeled with  $n$  for the number of columns and  $m$  for the number of rows. An approximation symbol  $\approx$  follows. To the right, the decomposition is shown as the product of three matrices: a dark blue matrix  $U$  with  $k$  columns, a green square matrix  $\Sigma$ , and an orange matrix  $V^T$  with  $k$  rows. The matrices are separated by multiplication symbols  $\times$ . Brackets above  $U$  and below  $V^T$  indicate their dimensions are  $k$ .

После восстановления исходной матрицы, клетки, где у пользователя были нули, а появились «большие» числа, показывают степень латентного интереса к товару. Упорядочим эти цифры и получим список товаров, релевантных для пользователя.

# Практика по рекомендательным системам

Ноутбук с примером реализации коллаборативной фильтрации и матричных разложений для построения рекомендательной системы пользователям, которые читают статьи на онлайн-ресурсе:





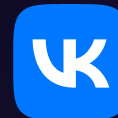
ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



# Методы кодирования текстовых данных

## Пример темы 2. Работа с текстами

Чтобы работать с текстом, необходимо разбить его на токены. В простейшем случае токены – это слова (а также наборы букв, знаки препинания и т.д.).





# Bag of words (мешок слов)

По корпусу создадим словарь из всех встречающихся в нем слов (можно убрать общеупотребительные часто встречающиеся слова и очень редкие слова).

Каждое слово закодируем вектором, в котором стоит единица на месте, соответствующем месту этого слова в словаре, все остальные компоненты вектора – 0.

Для кодирования документа сложим коды всех его слов.

Raw Text	Bag-of-words vector
it is a puppy and it is extremely cute	it 2
	they 0
	puppy 1
	and 1
	cat 0
	aardvark 0
	cute 1
	extremely 1
	...
	...

## Bag of words (пример)

Пусть корпус состоит из следующих документов:

- D1 - “I am feeling very happy today”
- D2 - “I am not well today”
- D3 - “I wish I could go to play”

Кодировка этих документов будет такой:

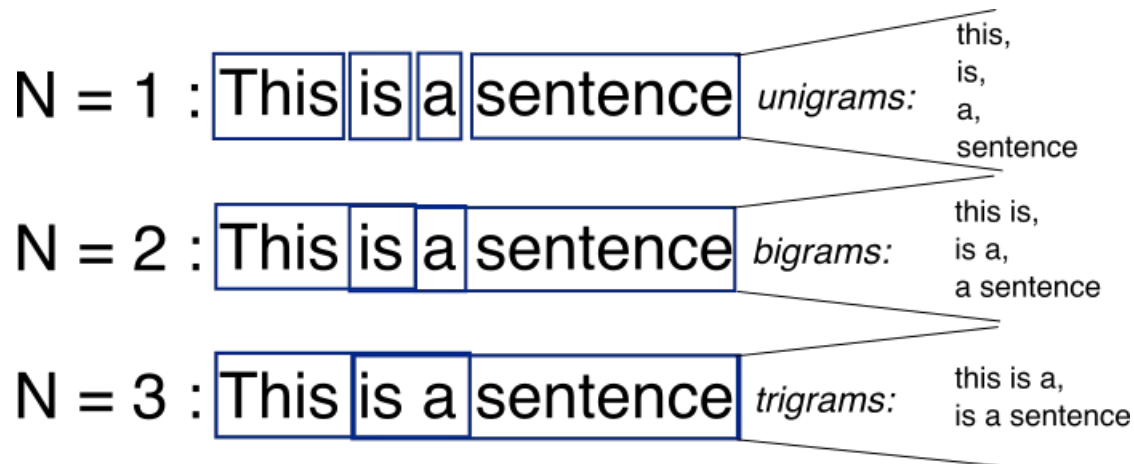
	I	am	feeling	very	happy	today	not	well	wish	could	go	to	play
D1	1	1	1	1	1	1	0	0	0	0	0	0	0
D2	1	1	0	0	0	1	1	1	0	0	0	0	0
D3	2	0	0	0	0	0	0	0	1	1	1	1	1

# N-gram bag of words

В качестве слов в словаре можно использовать:

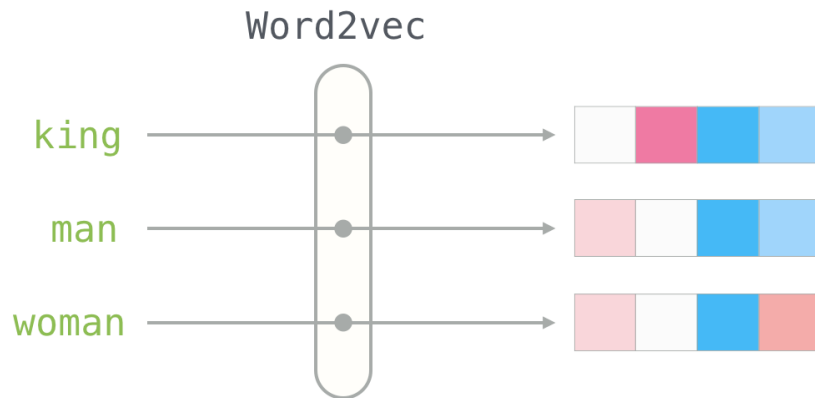
- N-граммы из букв (наборы букв длины N в слове)
- N-граммы из слов (наборы фраз длины N в документе)

*Такой подход поможет учесть сходственные слова и опечатки.*



# Другие способы векторизации текстов

- Из простых: tf-idf
- Из deep learning подходов:
  - Word2vec, fasttext и другие
  - Bert и другие



# Практика по работе с текстами

Ноутбук с примером работы с текстовыми данными

- [Ссылка](#)

Бесплатные курсы по работе с текстовыми данными на платформе Stepik:

- [First Step in NLP](#)
- [Second Step in NLP](#)

# Другие полезные ссылки и ноутбуки

Анализ временных рядов:

- [Ноутбук](#)
- [Курс на Stepik](#)

Обработка изображений базовыми методами машинного обучения:

- [Курс на Stepik](#)



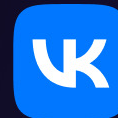
ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



# Подготовка к решению заданий по математике





# Задания по математике

В основном этапе будет **два задания**.

Стоит обратить внимание на прикладные темы из математики, которые необходимо знать для работы с данными:

- Дискретная математика (комбинаторика, графы),
- Основы теории вероятностей.



# Дискретная математика (комбинаторика и основы теории графов)

Из конкретных тем, на которые стоит обратить внимание:

- Бином Ньютона,
- Числа Каталана,
- Рекуррентные соотношения в комбинаторике.

Потренируйтесь решать комбинаторные задачи путем написания кода!  
Часто формулы комбинаторики можно представить в виде рекуррентного соотношения, которое удобно представить в виде кода программы и быстро посчитать на компьютере.



# Рекуррентные соотношения

Существует известная рекуррентная формула, связывающая между собой биномиальные коэффициенты:

$$C_n^k = C_{n-1}^k + C_{n-1}^{k-1}$$

Благодаря этой формуле мы можем, используя программу, посчитать очень большие значения  $C_n^k$ , которые напрямую посчитать не можем.

Лучше использовать динамическое программирование, нежели рекурсивный подход.

[Пример вычисления биномиальных коэффициентов в Python.](#)



# Теория вероятностей

Из конкретных тем, на которые стоит обратить внимание:

- Формулы произведения и суммы вероятностей,
- Формула Байеса,
- Формула полной вероятности.

Можно порешать задачи с портала [problems.ru](https://problems.ru).



ВСЕРОССИЙСКАЯ  
ОЛИМПИАДА ПО  
ИСКУССТВЕННОМУ  
ИНТЕЛЛЕКТУ



МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРО  
СВЕТ

ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ  
ПРОСВЕЩЕНИЯ



Если у вас появятся вопросы по  
Олимпиаде, вы всегда можете написать  
нам на почту [ai@guppros.ru](mailto:ai@guppros.ru) или в  
официальный чат участников ВКонтакте:

