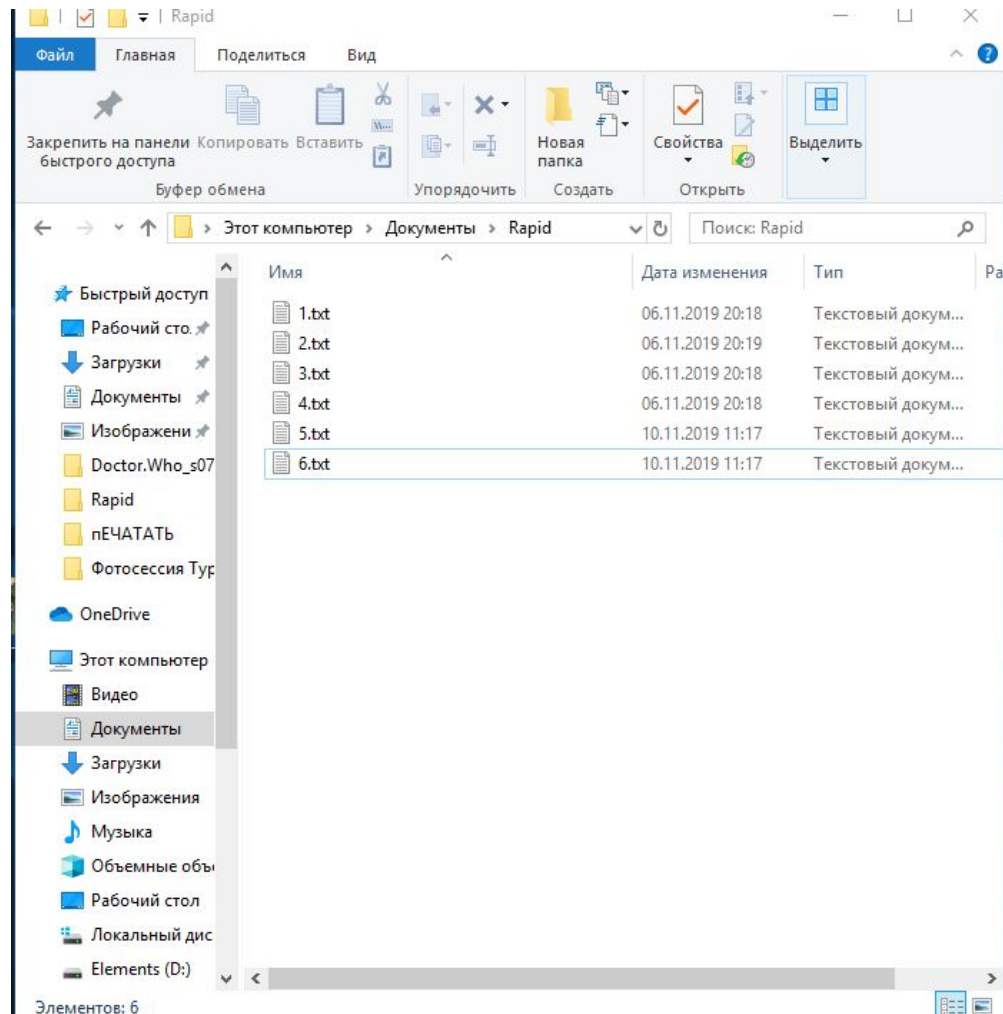


# Практическое задание



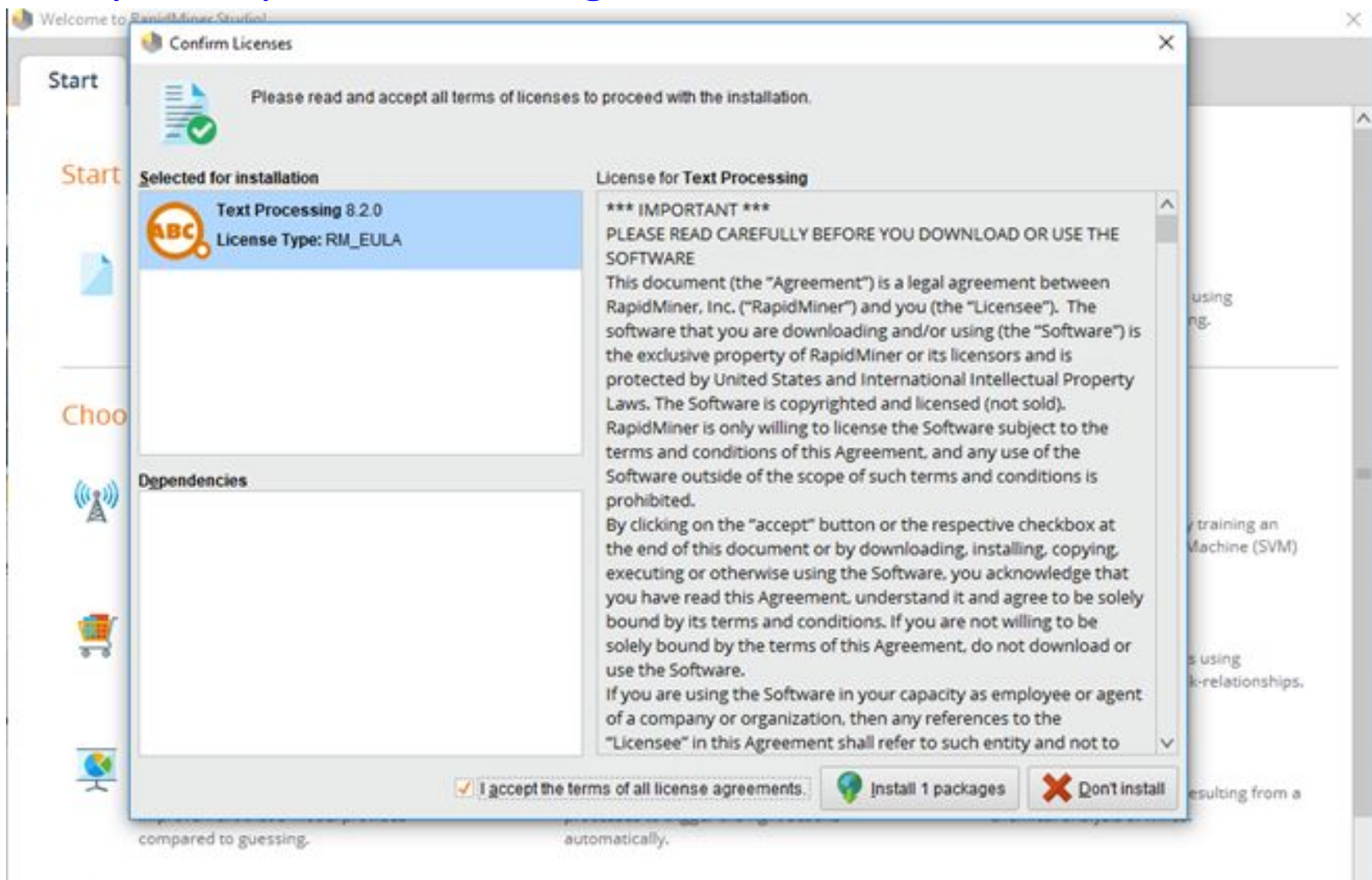
# 1) Подготовка данных



2)

# Установка RapidMiner. Установка КОМПОНЕНТОВ textMining

<https://rapidminer.com/get-started/>



# КОМПОНЕНТЫ- Process Dociment from file s и различных фильтров (МИНИМУМ-3).

The screenshot displays the RapidMiner Studio interface. The main workspace shows a process diagram with a single operator named "Process Documents from Files". The operator has an input port labeled "inp" and an output port labeled "res". The operator icon features a document symbol and a warning triangle, indicating a configuration issue.

The "Parameters" panel on the right is open, showing the configuration for the "Process Documents from Files" operator:

- text directories: Edit List (0...)
- file pattern: \*
- use file extension as type:
- vector creation: TF-IDF
- add meta information:
- keep text:

The "Operators" panel on the left shows a search for "text" with a list of operators under "Text Processing (11)". The operator "Process Documents from Files" is highlighted in blue.

A "Help" panel at the bottom right provides additional information about the operator, including tags: Text Processing.

At the bottom of the interface, a message reads: "Leverage the Wisdom of Crowds to get operator recommendations based on your process design!" with a green checkmark and the text "Activate Wisdom of Crowds".

At the very bottom, a footer note states: "Double-click to enter subprocess, drag to move."

Views: Design Results Turbo Prep Auto Model

### Repository

+ Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- DB (Legacy)
- Local Repository (Оксана)

### Operators

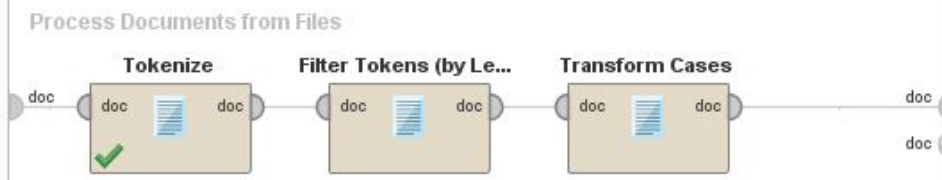
transfo

- Text Processing (11)
  - Transformation (11)
    - Transform Cases**
    - Replace Tokens
    - Remove Document Parts
    - Keep Document Parts
    - Generate n-Grams (Characters)
    - Generate n-Grams (Terms)
    - Cut Document
    - Window Document
    - Combine Documents

We found "Semweb", "RapidMiner Finance and Econom..." and one more result in the Marketplace. [Show me!](#)

### Process

Process Documents from Files 100%



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

# 4) Проведение кластеризации документов

The screenshot displays a machine learning workflow in a software interface. The main workspace shows a process flow starting with 'Process Documents...' and followed by 'Clustering'. The 'Clustering' operator is highlighted in orange. The interface is divided into several panels:

- Repository:** Shows data sources like 'Training Resources', 'Samples', 'Community Samples', 'DB (Legacy)', and 'Local Repository (Оксана)'. It includes an 'Import Data' button.
- Operators:** A search bar contains 'cl'. A list of operators is shown under 'Segmentation (14)', with 'k-Means' selected. Other operators include 'k-Means (Kernel)', 'k-Means (fast)', 'X-Means', 'k-Medoids', 'DBSCAN', 'Expectation Maximization Clustering', and 'Support Vector Clustering'.
- Process:** Shows the workflow diagram with 'Process Documents...' and 'Clustering' operators. The 'Clustering' operator has 'k' set to 3 and 'max runs' set to 10. A 'Wisdom of Crowds' notification is visible at the bottom of the process area.
- Parameters:** A detailed view of the 'Clustering (k-Means)' operator. It includes checkboxes for 'add cluster attribute', 'add as label', and 'remove unlabeled'. The 'k' parameter is set to 3, and 'max runs' is set to 10. Other options include 'determine good start values', 'Hide advanced parameters', and 'Change compatibility (9.4.001)'.
- Help:** A section titled 'k-Means' with the sub-heading 'Concurrency'. It lists tags: 'Unsupervised', 'Clustering', 'Segmentation', 'Grouping', 'Similarity', 'Similarities', 'Euclidean', 'Distances', 'Centroids', 'K Means', 'K means', 'Kmeans'. The synopsis states: 'This Operator performs clustering using the *k-means* algorithm.' A link to 'Jump to Tutorial Process' is provided.

Views: Design Results Turbo Prep Auto Model Deployments More ▾

Result History Cluster Model (Clustering) X

- ^
- Description
- Folder View
- Graph
- Centroid Table
- Plot
- Annotations

- root
  - cluster\_0
    - 3.0
    - 4.0
    - 5.0
  - cluster\_1
    - 1.0
    - 2.0
  - cluster\_2
    - 6.0

## 5) Численная оценка качества алгоритма

(точность, полнота, F-мера)

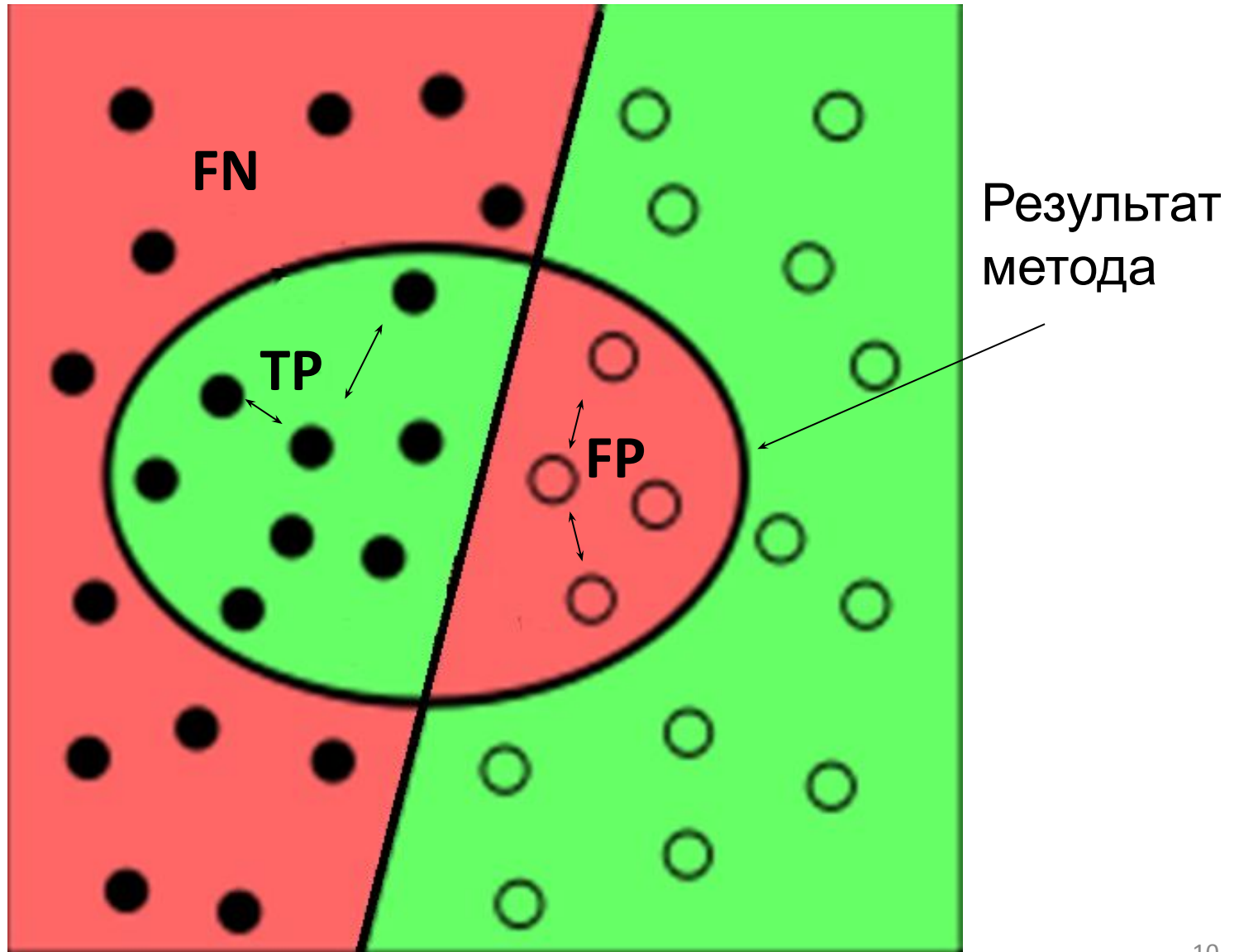


# Точность и полнота

Категория $i$		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

- TP — истинно-положительное решение;
- TN — истинно-отрицательное решение;
- FP — ложно-положительное решение;
- FN — ложно-отрицательное решение.

# Пример (наглядность)



# Точность и полнота

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

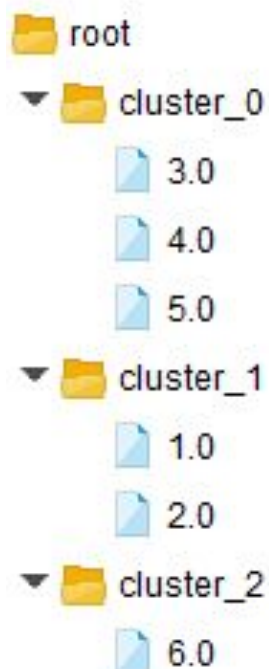
# F-мера

$$F = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

$$F = (\beta^2 + 1) \frac{Precision * Recall}{\beta^2 Precision + Recall} \quad \beta^2 \in [0, \infty]$$

Где  $\beta$  принимает значения в диапазоне  $0 < \beta < 1$ , если Вы хотите отдать приоритет точности, а при  $\beta > 1$  приоритет отдается полноте.

При  $\beta=1$  формула сводится к предыдущей и вы получаете сбалансированную F-меру (также ее называют  $F_1$ )



Кластер	Название файлов, которые должны были попасть	Название файлов, которые попали	Кол-во правильно попавших текстов	Кол-во не попавших текстов	Кол-во не правильно попавших текстов
0	3,4	3,4,5	2	0	1
1	1,2	1,2	2	0	0
2	5,6	6	1	1	0

- TP — 5
- FP — 1
- FN — 1

$$\text{Precision} = 5/6 = 0.83$$

$$\text{Recall} = 5/6 = 0.83$$

$$F = 2 * \frac{0.83 * 0.83}{0.83 + 0.83} = 2 * \frac{0.6869}{1.66} = 0.82$$

## 6) Анализ полученных результатов (полноценный анализ работы)

- Вывод по работе алгоритмов
- Обоснование полученных результатов