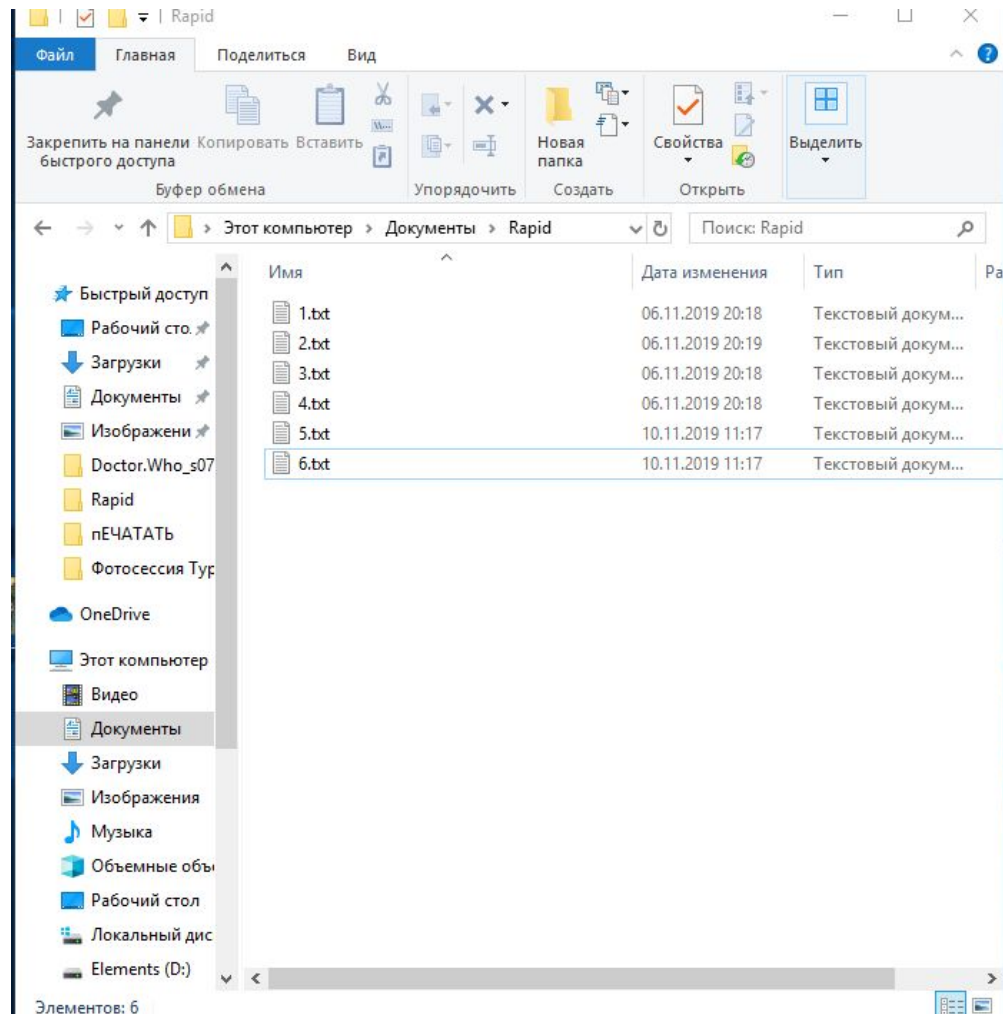


Практическое задание



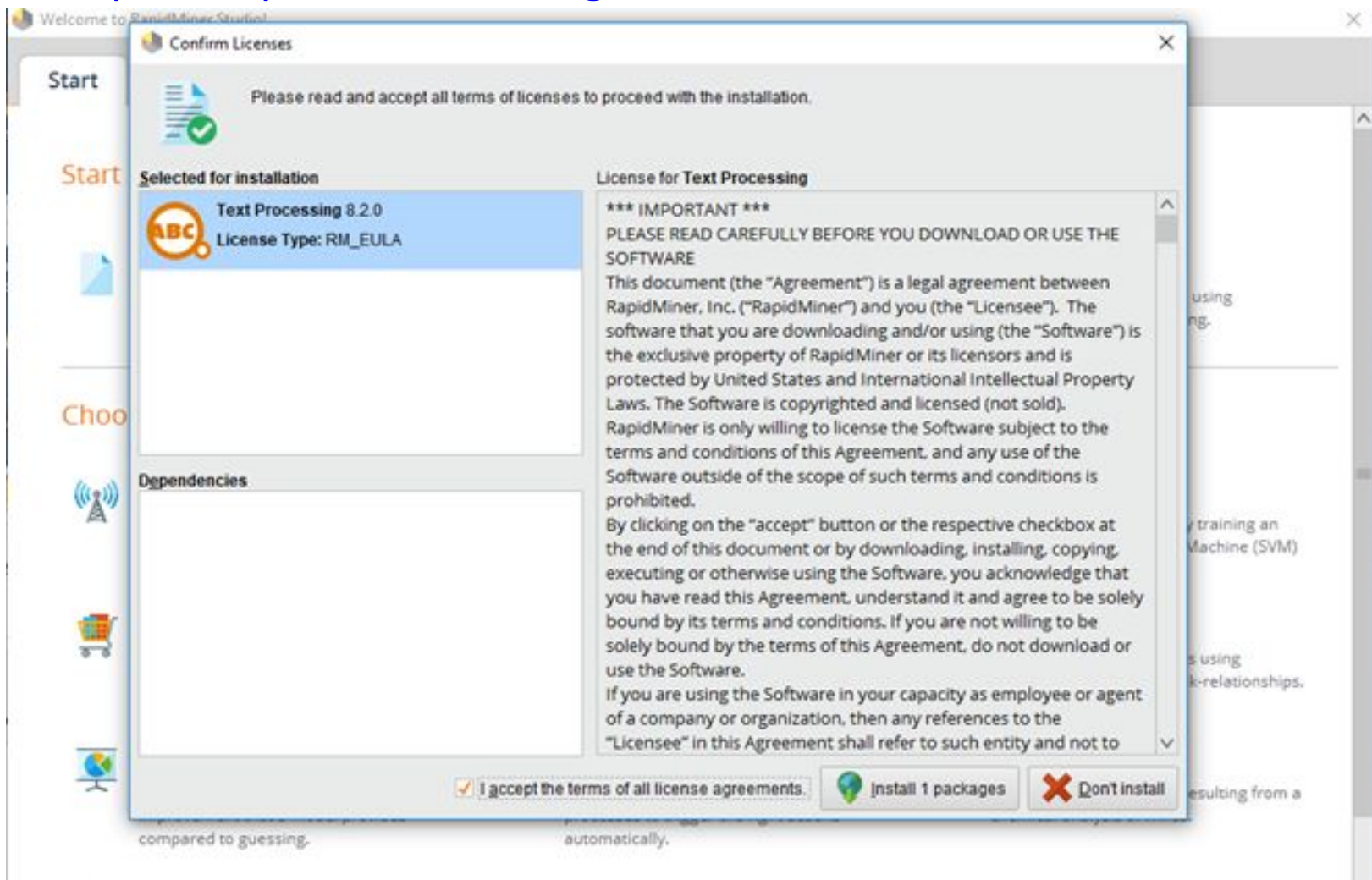
1) Подготовка данных



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, "Process Documents from Files", which is highlighted with a yellow warning icon. The interface includes a menu bar (File, Edit, Process, View, Connections, Settings, Extensions, Help), a toolbar with icons for file operations and execution, and a top navigation bar with tabs for Design, Results, Turbo Prep, Auto Model, and More. On the left, there are panels for "Repository" (containing folders like Training Resources, Samples, Community Samples, and DB) and "Operators" (with a search for "text" and a list of text processing operators). The "Process Documents from Files" operator is selected in the Operators list. On the right, the "Parameters" panel is open, showing configuration options for the operator: "text directories" (with an "Edit List" button), "file pattern" (set to "*"), "use file extension as type" (checked), "vector creation" (set to "TF-IDF"), "add meta information" (checked), and "keep text" (unchecked). A "Help" panel at the bottom right provides additional information about the operator. At the bottom of the interface, a message states: "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".

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%

Process Documents from Files

```
graph LR; In((doc)) --> Tokenize[Tokenize]; Tokenize --> Filter[Filter Tokens (by Le...)]; Filter --> Transform[Transform Cases]; Transform --> Out1((doc)); Transform --> Out2((doc));
```

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 data science software interface with the following components:

- Repository:** Shows data sources including Training Resources, Samples, Community Samples, DB (Legacy), and Local Repository (Оксана).
- Operators:** A search bar contains 'cl'. The 'Segmentation (14)' category is expanded, listing operators such as k-Means, k-Means (Kernel), k-Means (fast), X-Means, k-Medoids, DBSCAN, Expectation Maximization Clustering, and Support Vector Clustering.
- Process:** A workflow diagram showing a 'Process Documents...' operator connected to a 'Clustering' operator. The process is running at 100%.
- Parameters (Clustering (k-Means)):**
 - add cluster attribute
 - add as label
 - remove unlabeled
 - k: 3
 - max runs: 10
 - determine good start values
 - [Hide advanced parameters](#)
 - [Change compatibility \(9.4.001\)](#)
- Help (k-Means):** Provides a synopsis: "This Operator performs clustering using the *k-means* algorithm." and includes tags like Unsupervised, Clustering, Segmentation, etc.

Views: Design Results Turbo Prep Auto Model Deployments More ▾

Result History Cluster Model (Clustering) ×

^

- 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

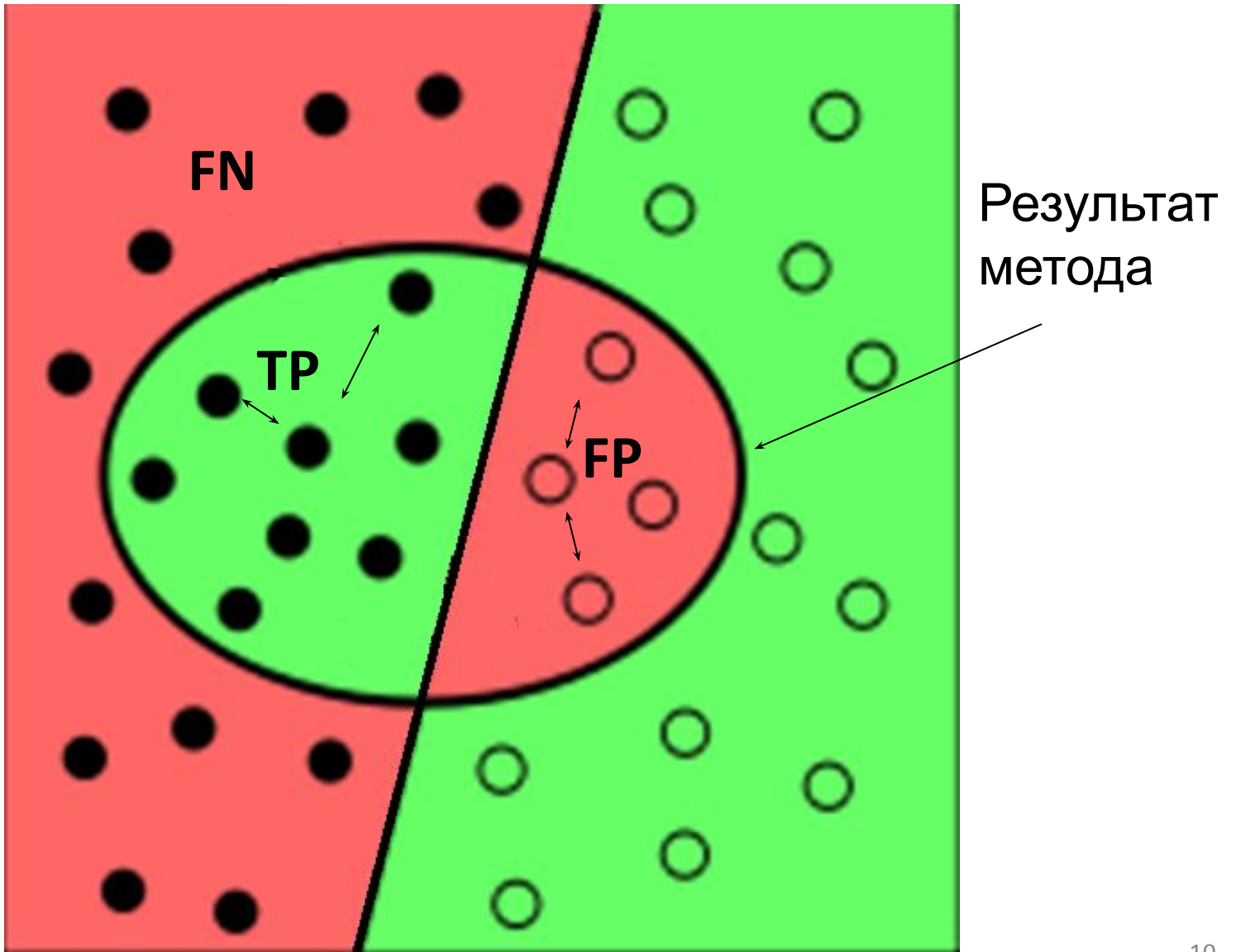
(точность, полнота, F-мера)

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

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

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

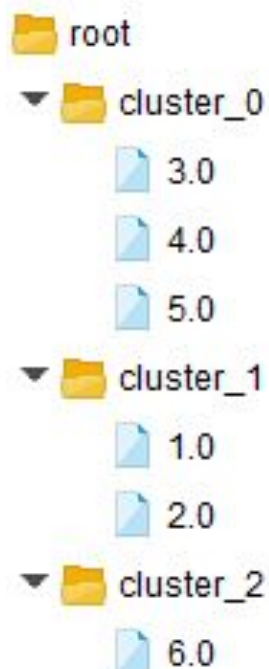
F-мера

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

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

Где β 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