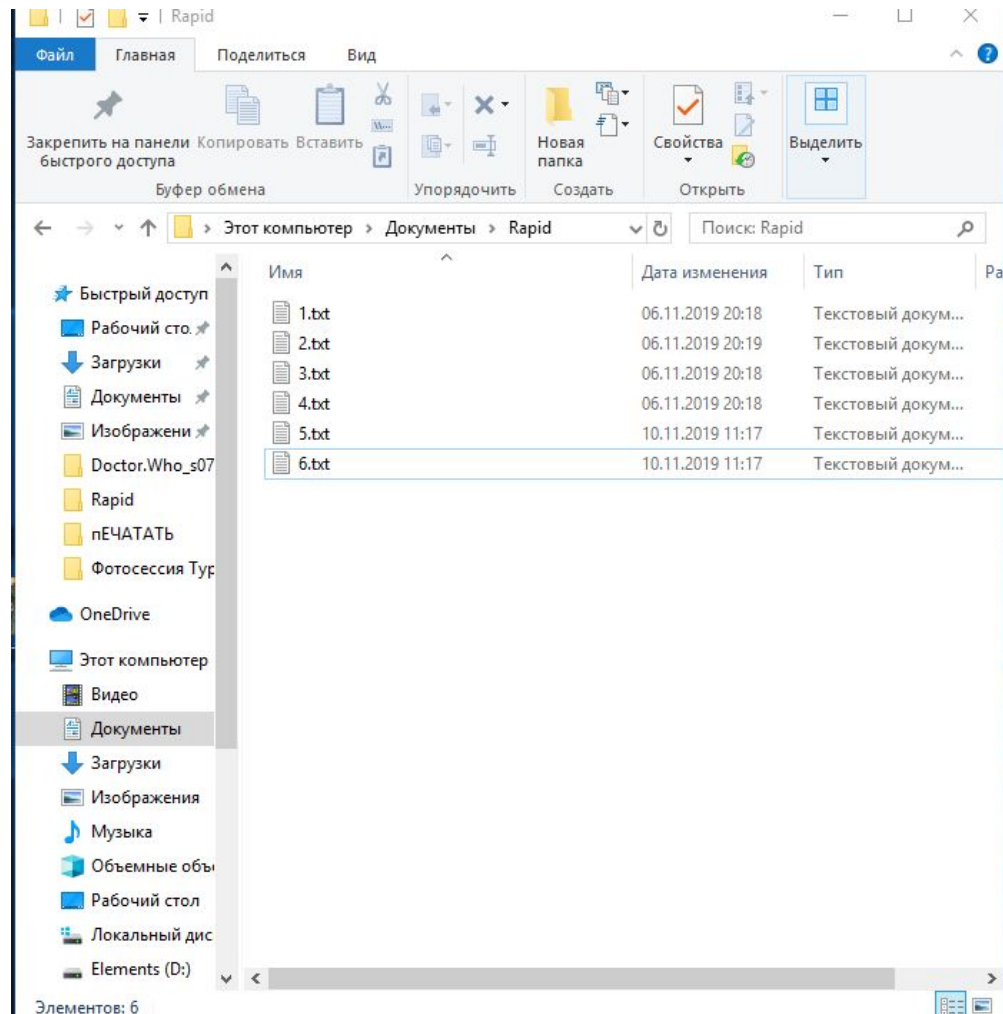


Практическое задание



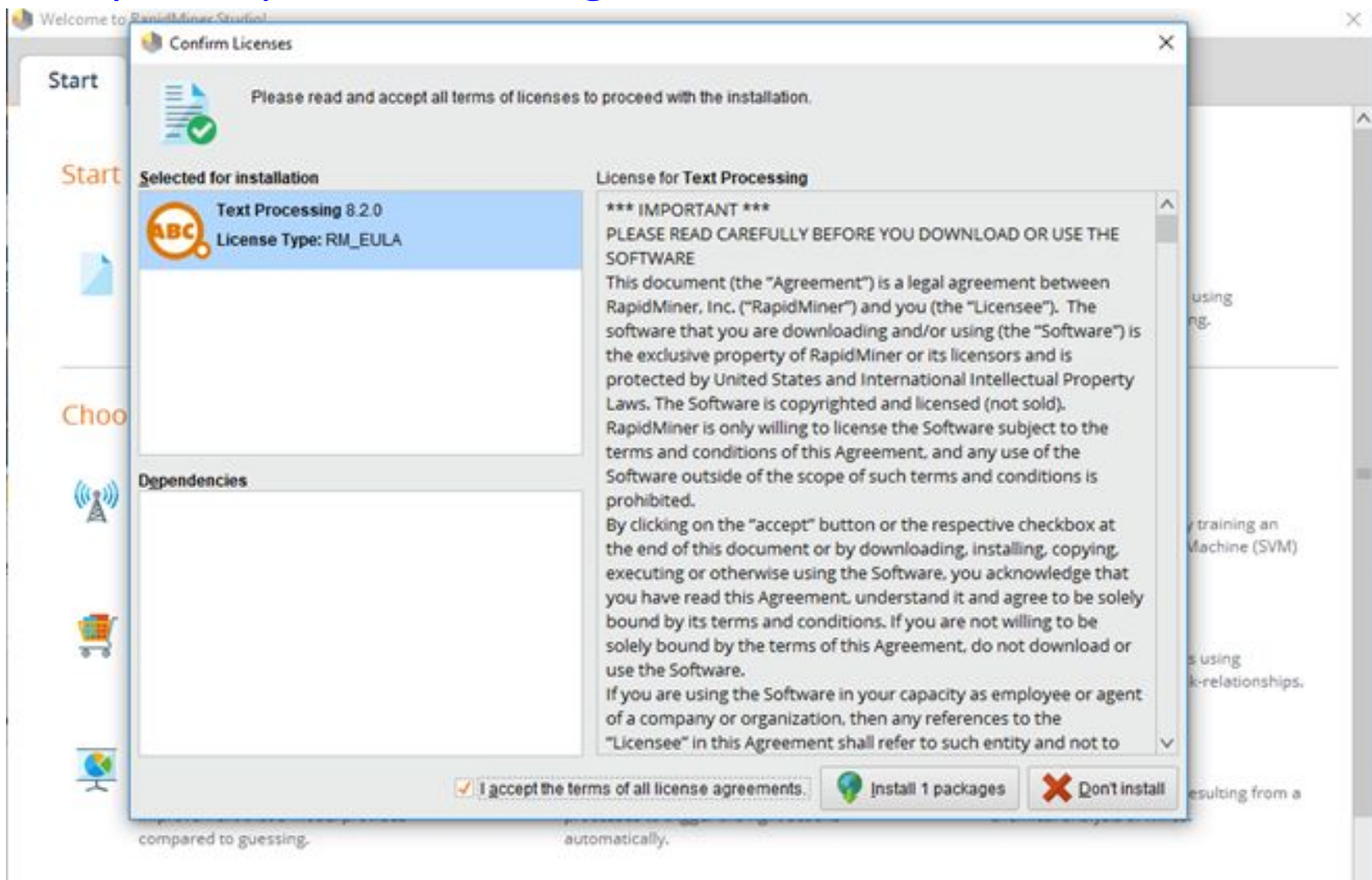
1) Подготовка данных



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 yellow 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 results, including "Process Documents from Files" which is highlighted. A tooltip at the bottom of the operators panel reads: "We found 'Text Analysis by AYLIEN', 'Edda - Extensions for Binomin...' and 6 more results in the Marketplace. [Show me!](#)"

The "Help" panel at the bottom right provides a summary for the "Process Documents from Files" operator, including tags: [Text Processing](#).

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".

At the very bottom, a footer note reads: "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

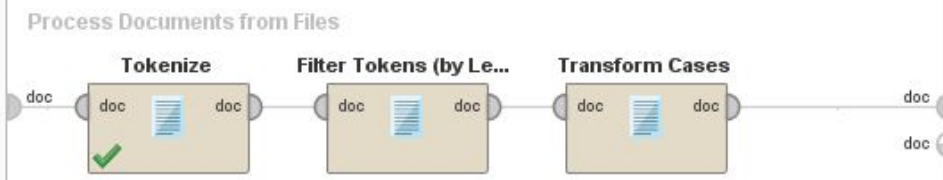
transfo

- 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 interface with several panels:

- Repository:** Shows data sources including Training Resources, Samples, Community Samples, DB (Legacy), and Local Repository (Оксана).
- Operators:** A search for 'cl' has been performed, showing various clustering algorithms under the 'Segmentation (14)' category, including 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 two operators: 'Process Documents...' and 'Clustering'. The 'Process Documents...' operator has two input ports ('inp') and two output ports ('wor', 'exa'). The 'Clustering' operator has two input ports ('exa', 'clu') and two output ports ('clu', 'res').
- Parameters:** The 'Clustering (k-Means)' operator is configured with the following settings:
 - 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:** Provides information about the 'k-Means' operator, including tags (Unsupervised, Clustering, Segmentation, Grouping, Similarity, Similarities, Euclidean, Distances, Centroids, K Means, K means, Kmeans) and a synopsis: 'This Operator performs clustering using the k-means algorithm.'

At the bottom of the interface, there is a notification: 'Leverage the Wisdom of Crowds to get operator recommendations based on your process design!' with an 'Activate Wisdom of Crowds' button.

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

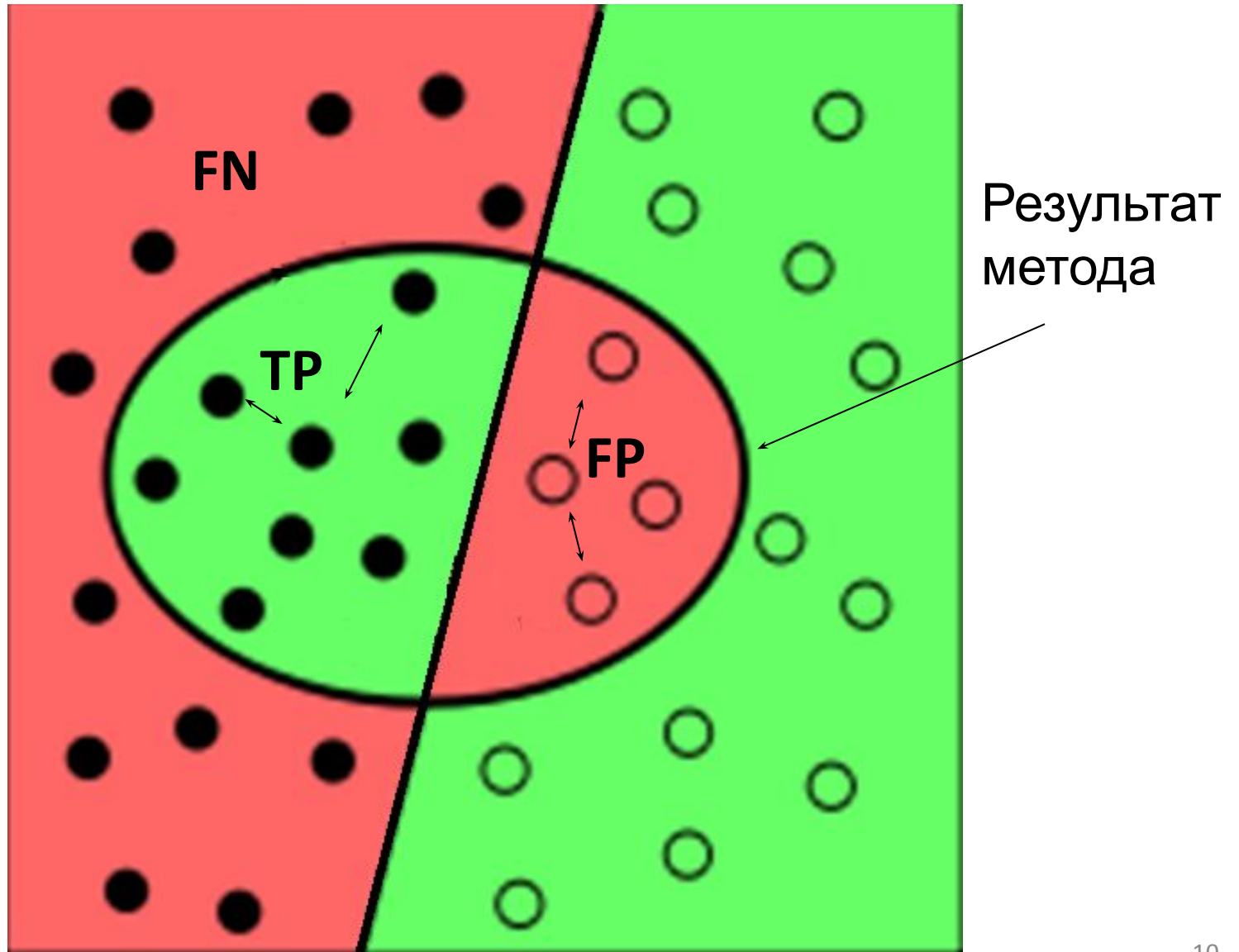
(точность, полнота, F-мера)

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

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

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

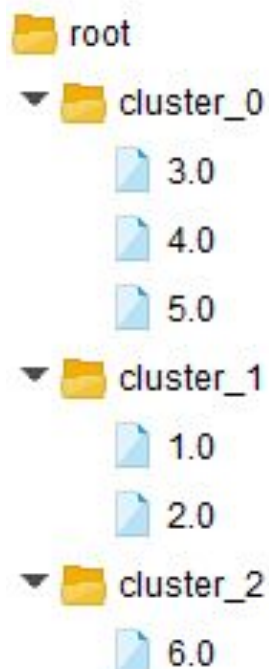
F-мера

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

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

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