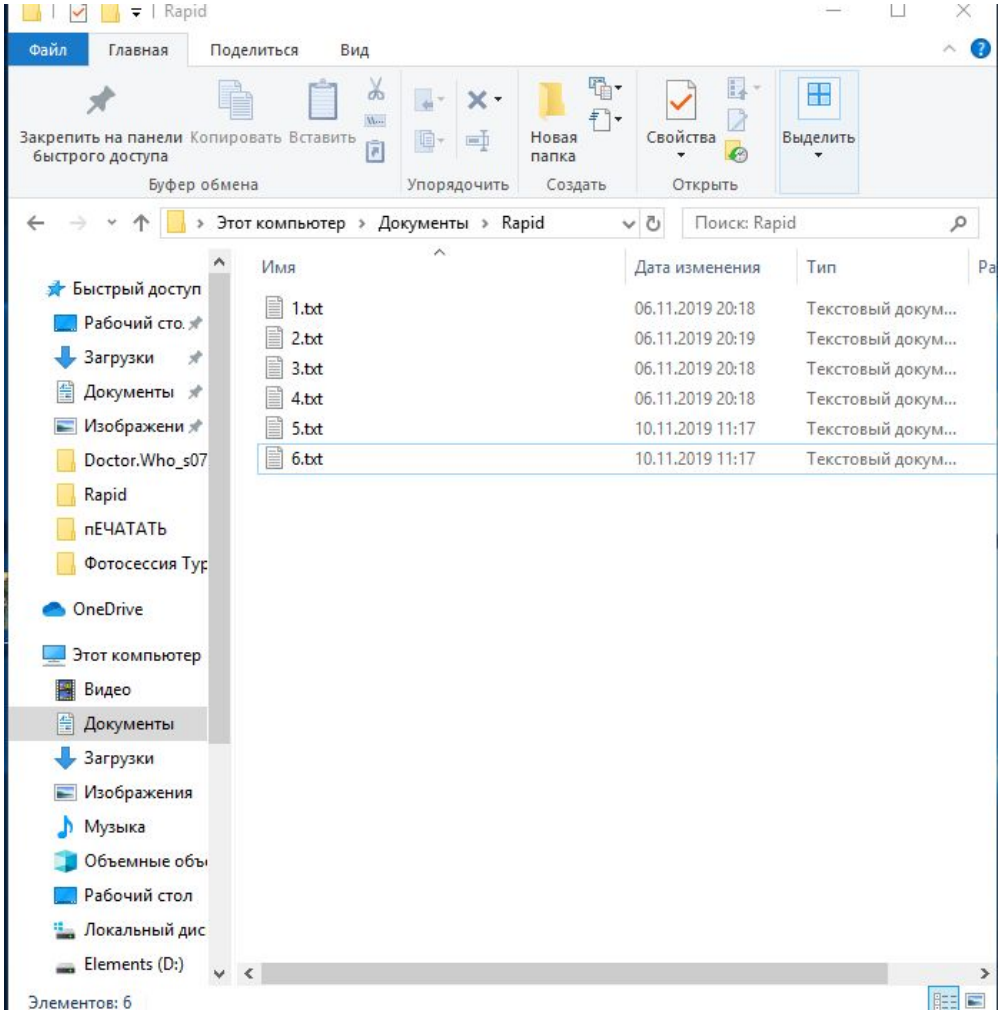
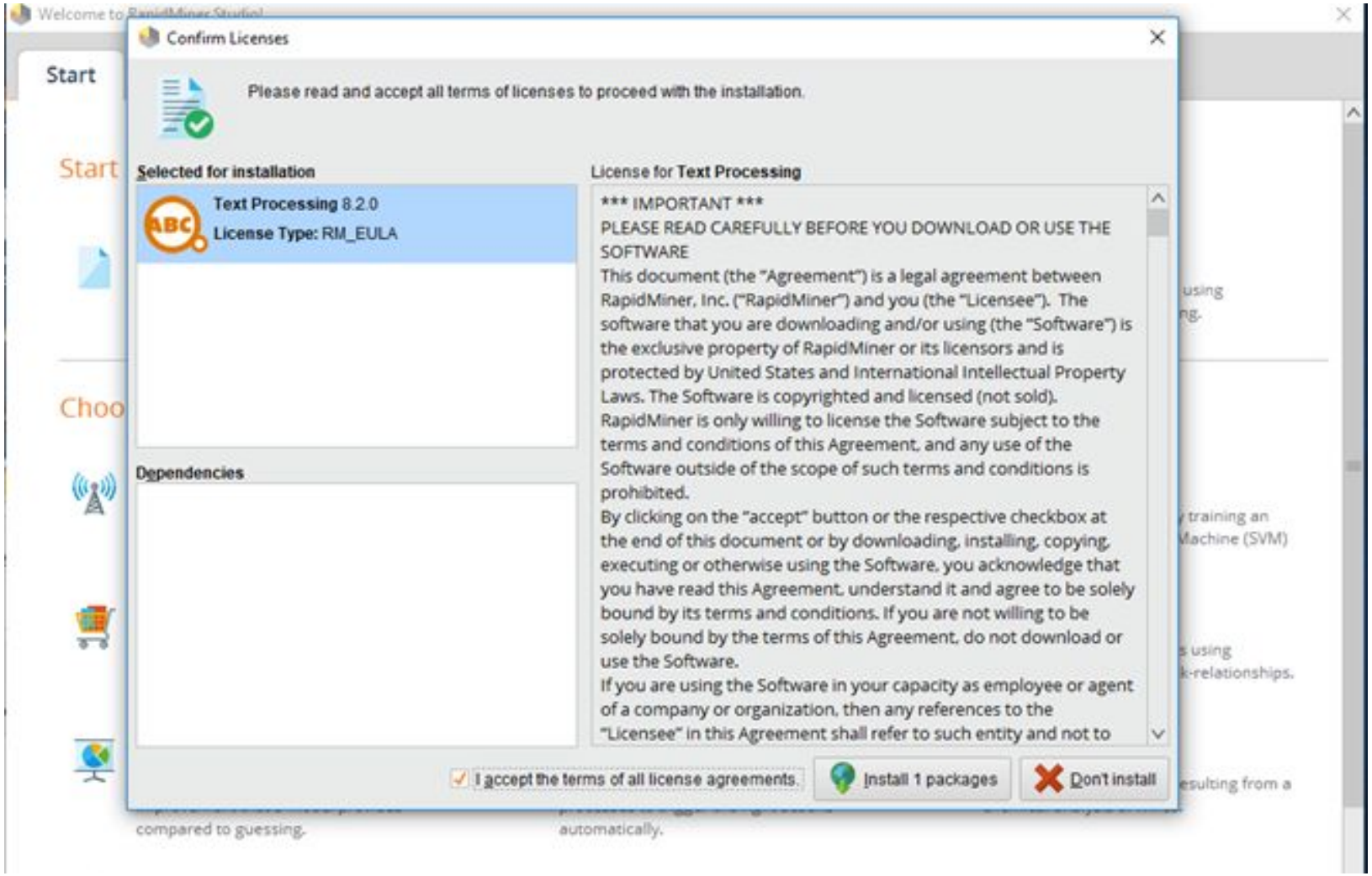
Практическое задание

1)Подготовка данных

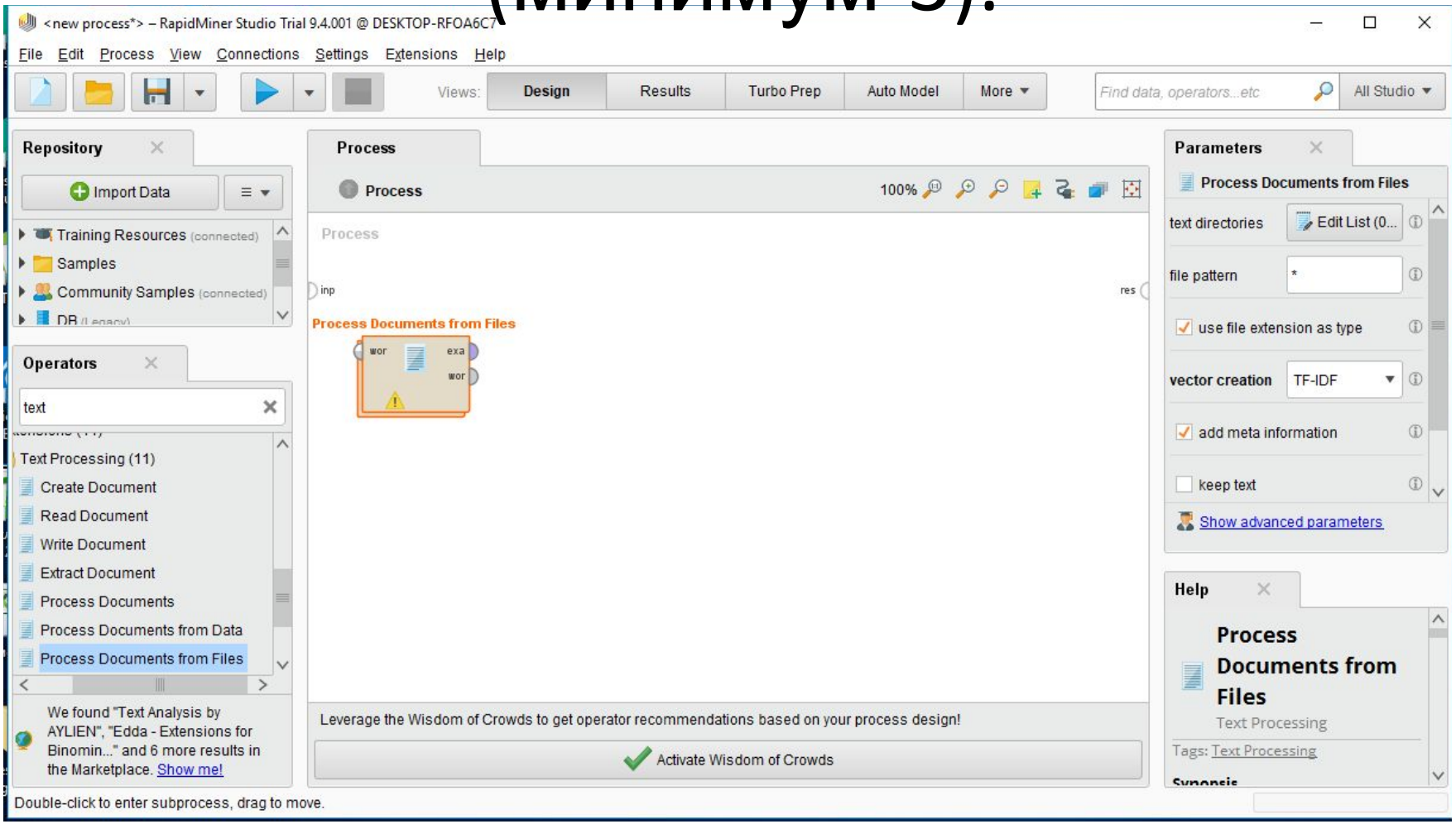


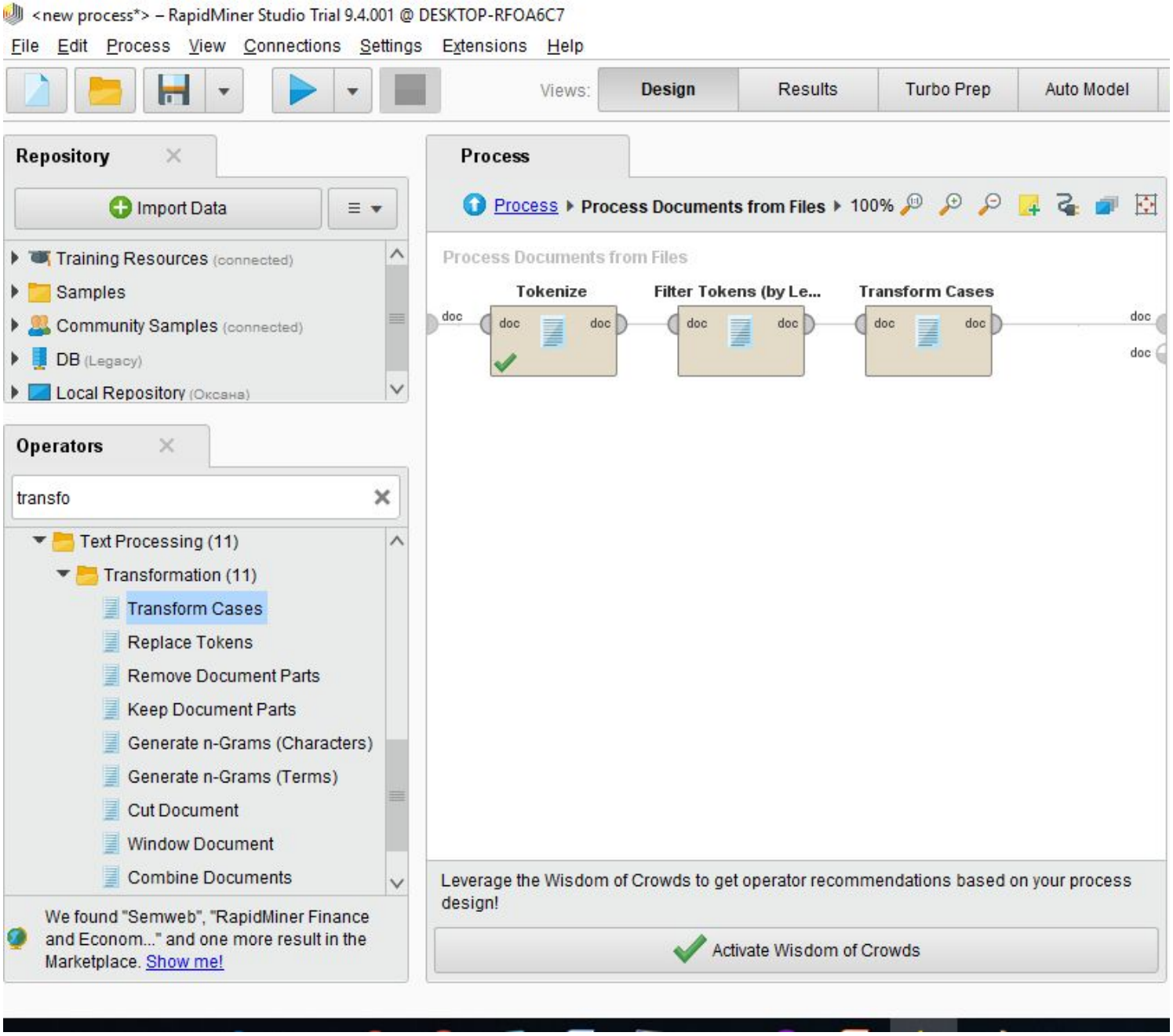
2)
Установка RapidMiner. Установка компонентов textMining

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

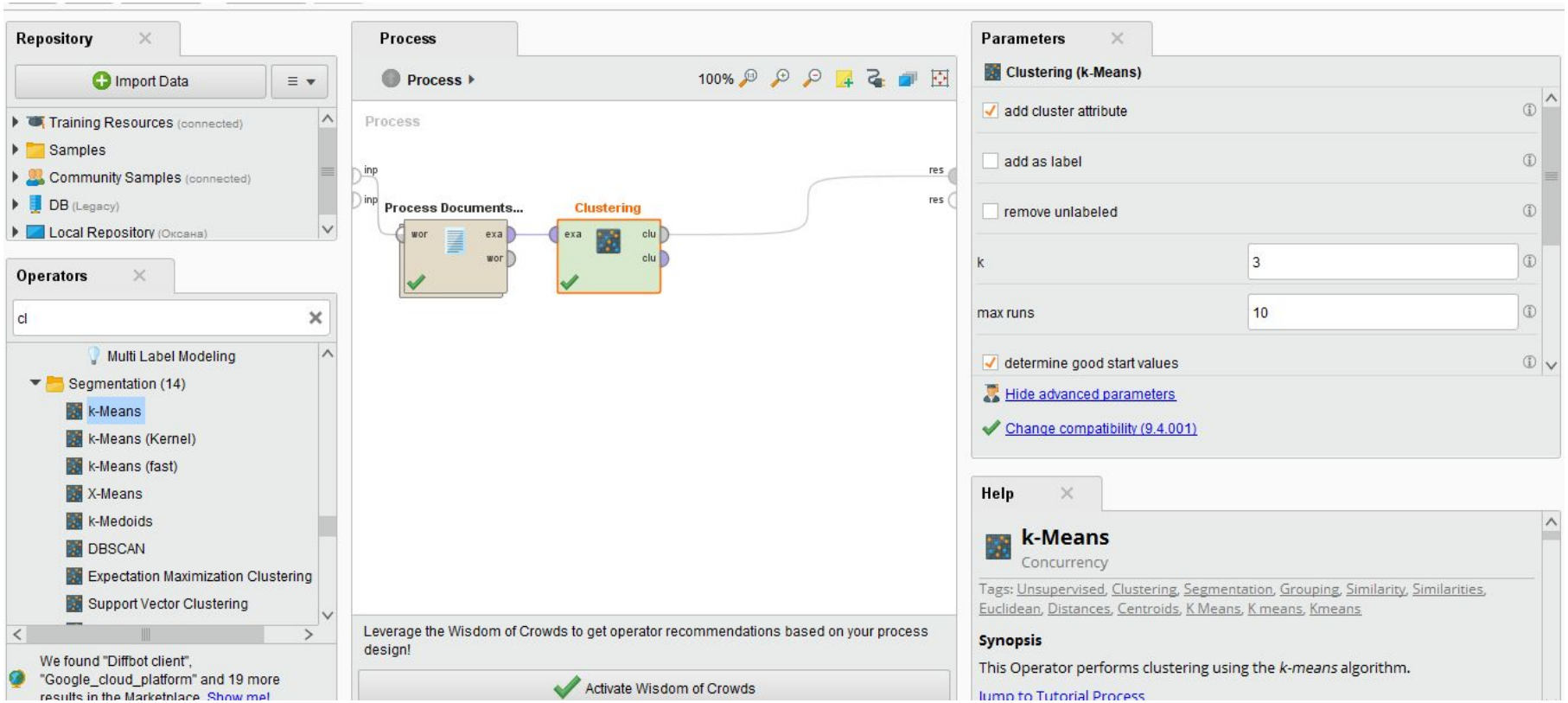


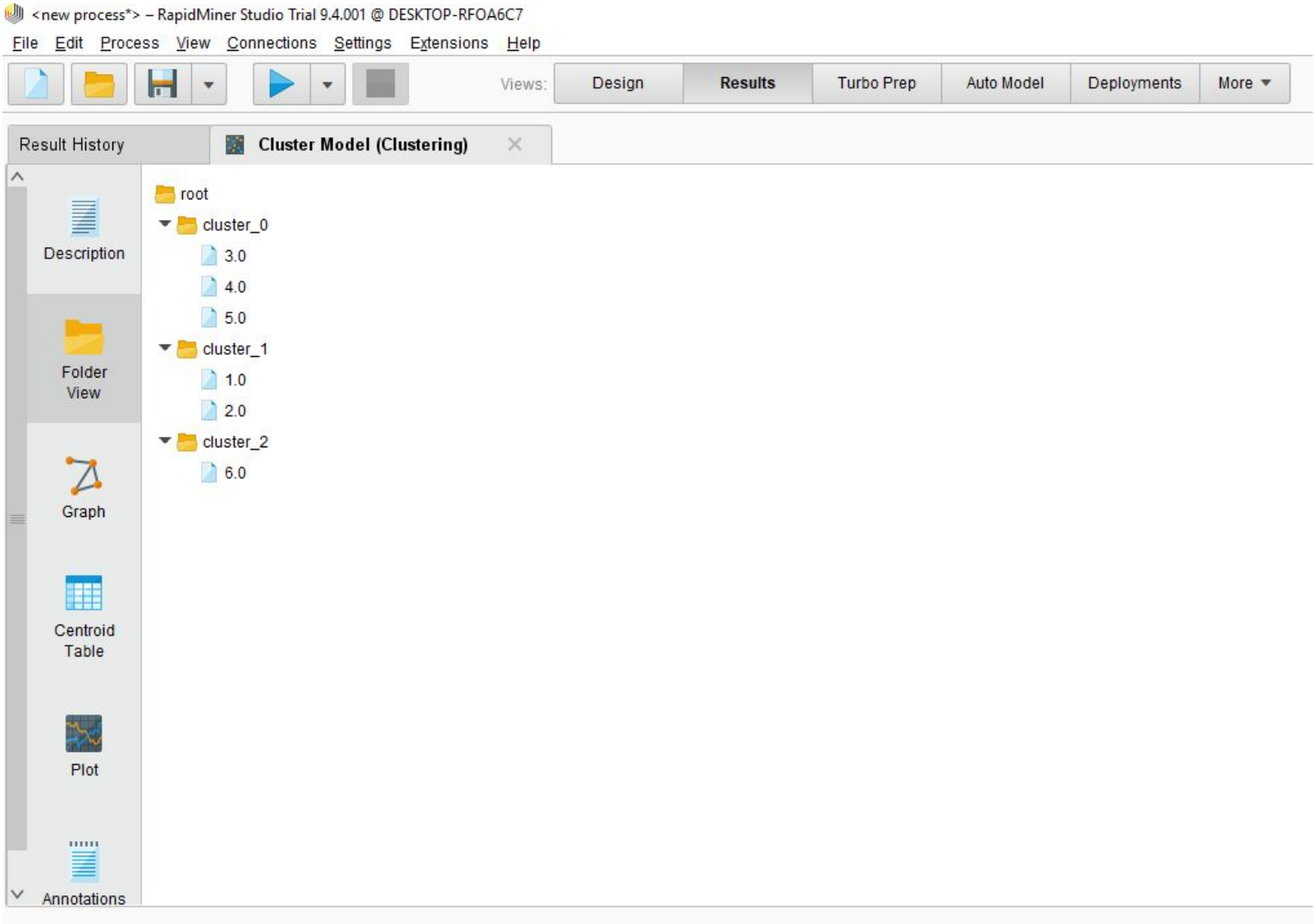
3) Использование компоненты- Process Dociment from files и различных фильтров (минимум-3).




<new process*> – RapidMiner Studio Trial 9.4.001 @ DESKTOP-RFOA6C7
File Edit Process View Connections Settings Extensions Help
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 ▸ Process Documents from Files ▸ 100%
Process Documents from Files
Tokenize
Filter Tokens (by Le...
Transform Cases
doc
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

4) Проведение кластеризации документов




<new process*> – RapidMiner Studio Trial 9.4.001 @ DESKTOP-RFOA6C7
File Edit Process View Connections Settings Extensions Help
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

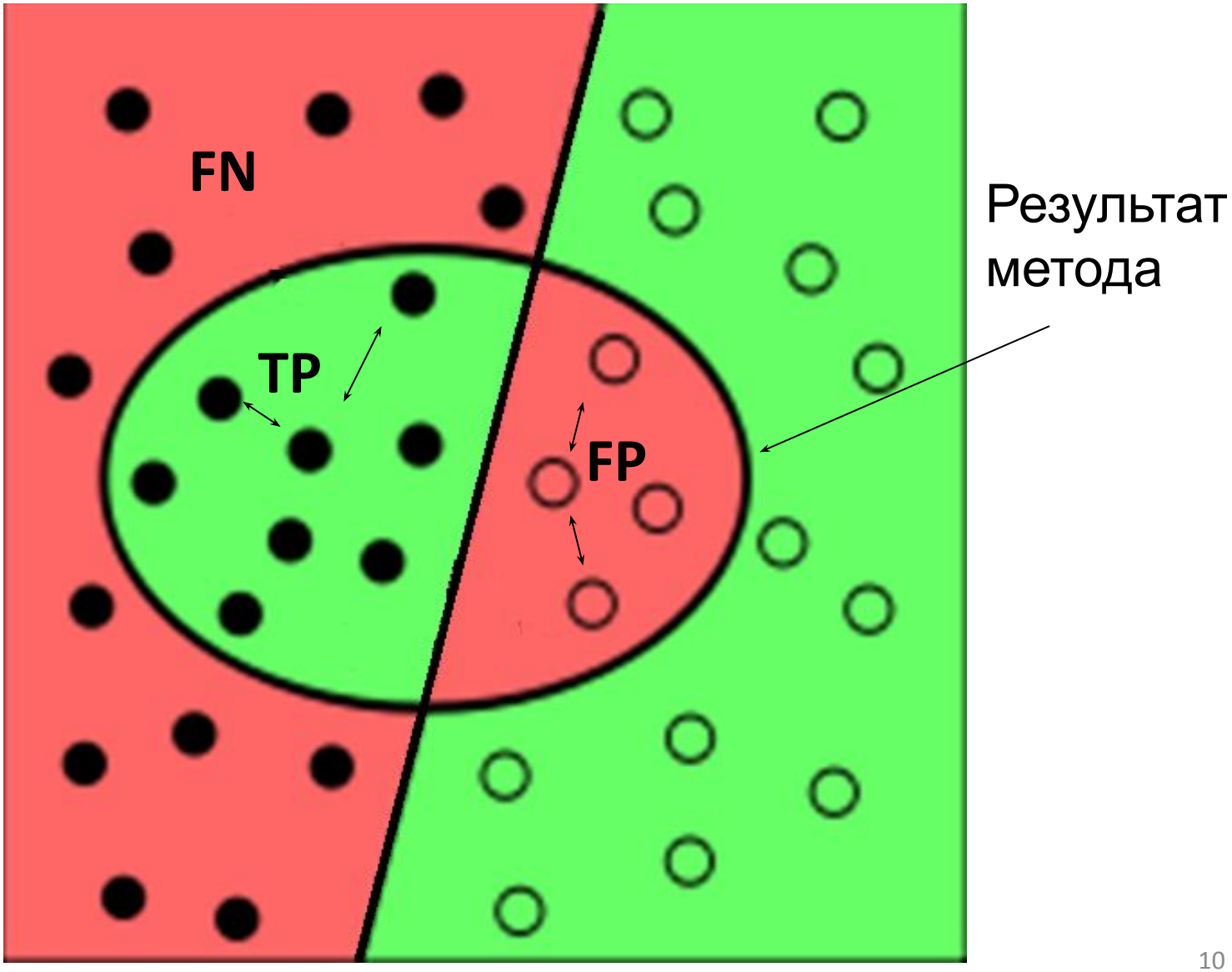
TP— истино-положительное решение;

TN — истино-отрицательное решение;

FP — ложно-положительное решение;

FN — ложно-отрицательное решение.

Пример (наглядность)



Точность и полнота

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

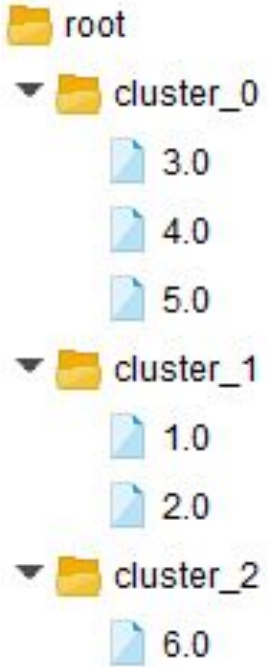
$$Recall = \frac{TP}{TP+FN}$$

F-мера

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

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

Где β принимает значения в диапазоне 0 <β <1, если Вы хотите отдать приоритет точности, а при β> 1 приоритет отдается полноте.
При β=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

Precision= 5/6= 0.83

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