НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Classification and routing of incoming correspondence with specific vocabularies.

PREPARED BY

YAKUNINA KSENIIA

TERM PAPER SUPERVISOR:

A.V. HELVAS

Content

- Work Description
- Target and Objectives
- Data collection
- Data pre-processing
- Cluster Analysis
- Results

Work Description

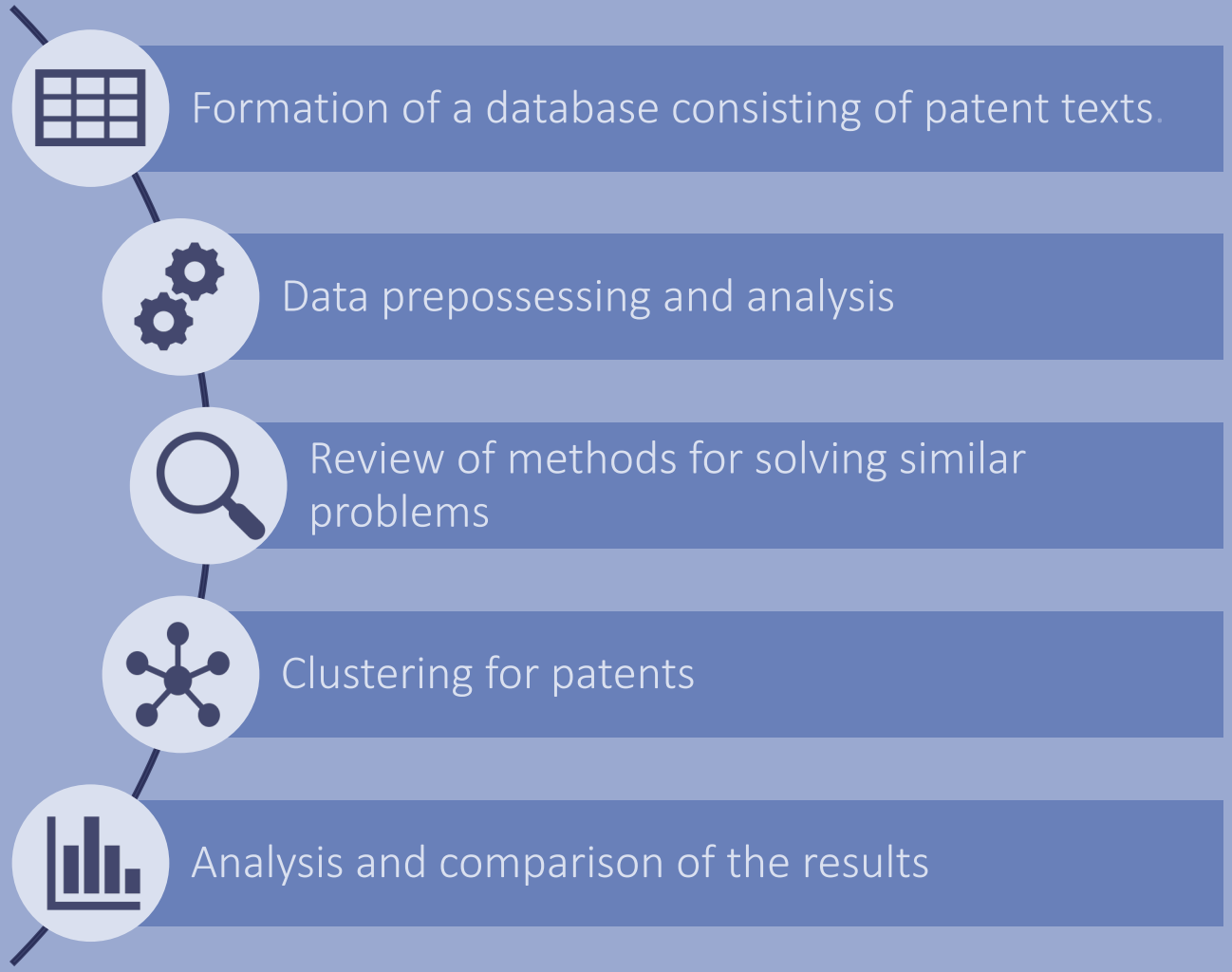
This work represents a project for the clustering of text documents by keywords in Russian, in particular patent abstracts.

Relevance

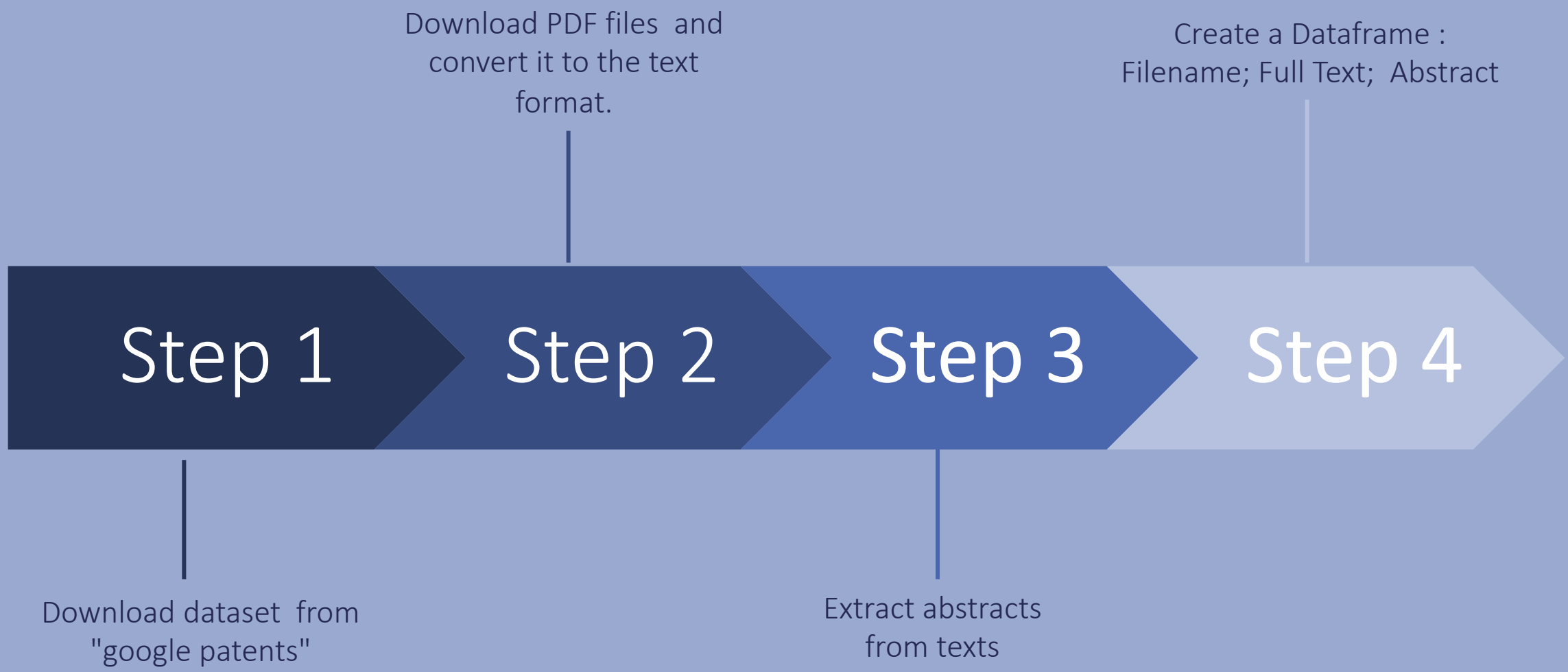
For any company, the process of sorting the incoming flow of documents can be costly in terms of time and human resources. Therefore, it is much more profitable to use automatic sorting of documents, for example, to immediately send a document to the necessary department for work. This will significantly reduce the cost of manual processing of documents and allow more efficient allocation of resources.

Target and Objectives

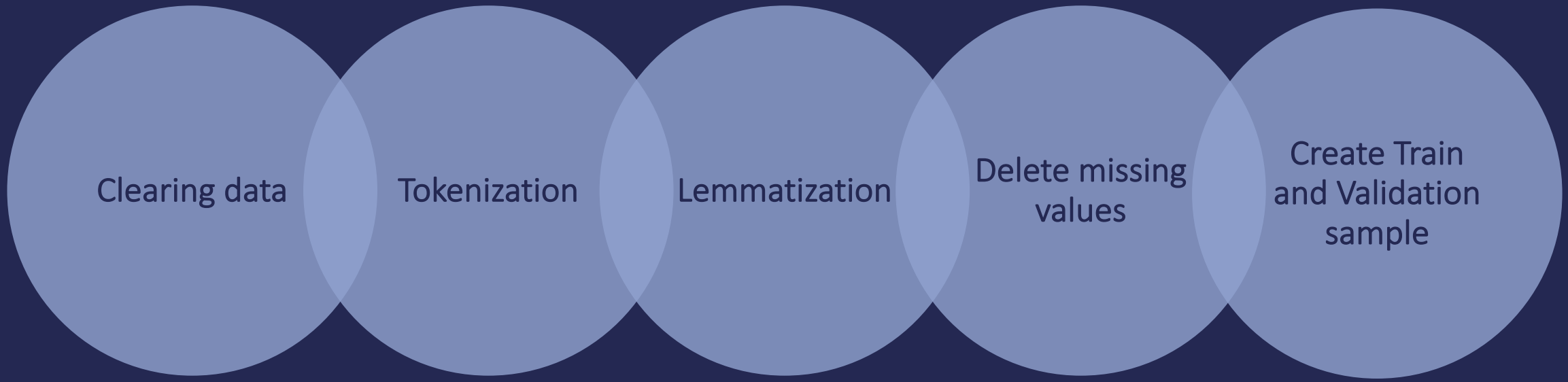
Create a training sample for a thematic area containing a large amount of special terminology and to create a solution for the clustering of documents.



Data Collection



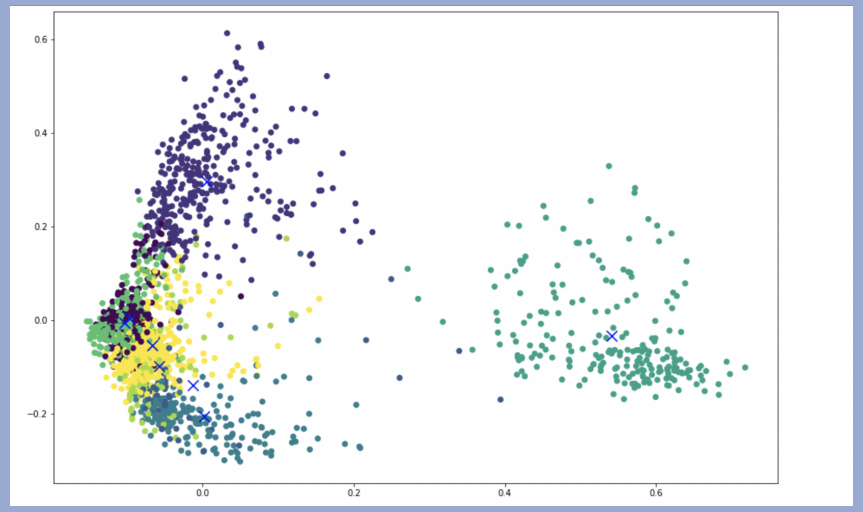
Data pre-processing



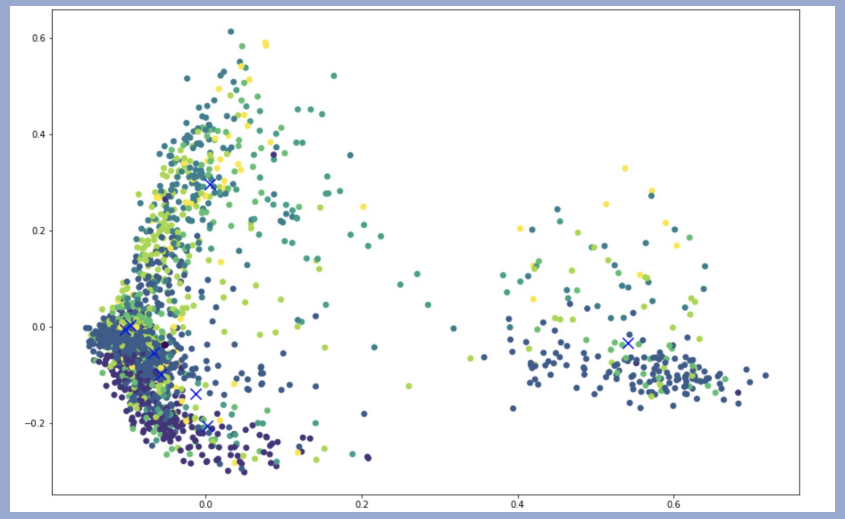
Cluster Analysis

- Data with encoding by Tf-idf algorithm show better performance than without it.

Clustering with Tf-idf



Clustering without Tf-idf

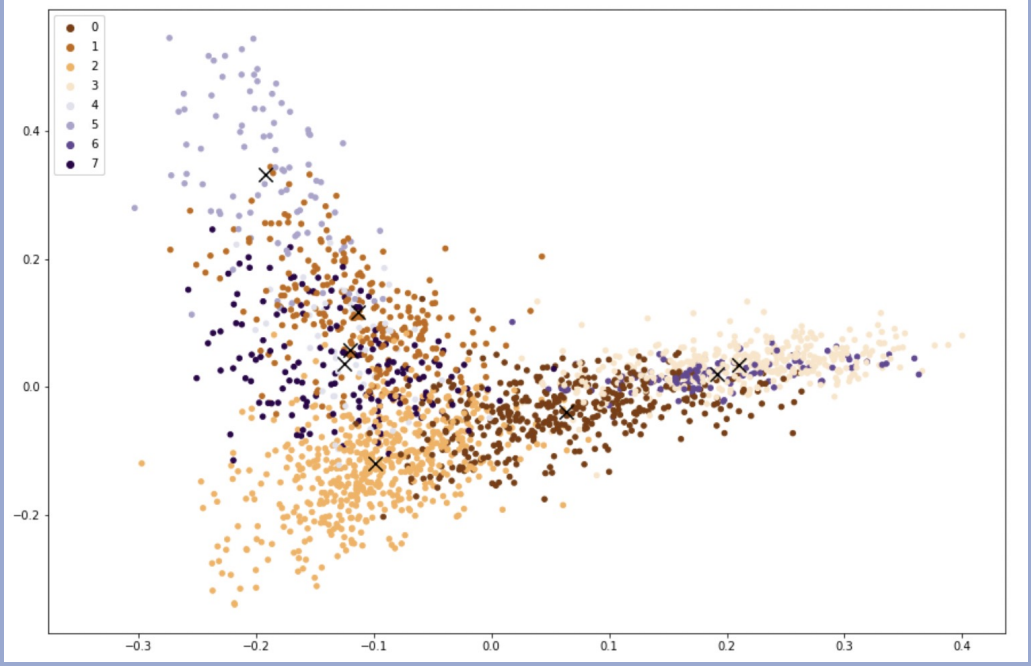


- Optimal Number of clusters — 8

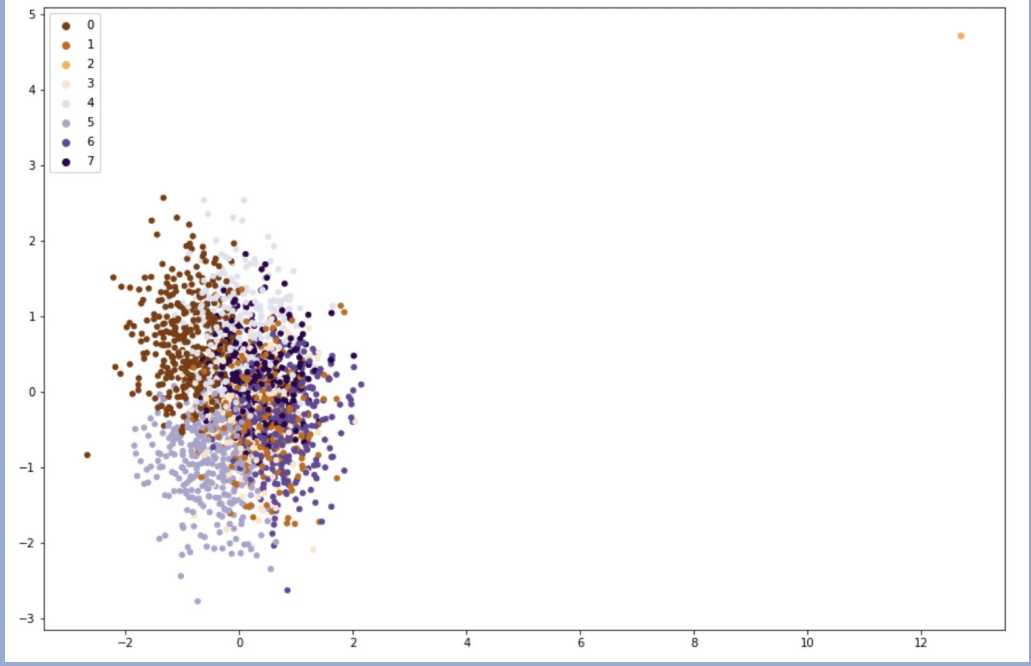
Cluster Analysis

Only K-mean method show better performance than K-means with the pre-trained Bert model.

K-means



K-means + Bert



Cluster Results

Cluster 0: получение, композиции, кислоты, композиция, масса, основе, способ, материал, изготовление, газ, смеси, слой;

Cluster 1: данные, изображения, информация, блок, кодирование, декодирование, множество, основа, сигнал, первый, устройство, передача;

Cluster 2: элемент, часть, устройство, поверхность, аэрозоль, узел, направление, корпус, элемент, отверстие;

Cluster 3: лечение, антитело, клетки, рак, человек, заболевание, применение, последовательность, композиции, против;

Cluster 4: транспортный, средство, система, панель, управление, камеры, достигается, множество;

Cluster 5: связь, беспроводной, передача, линия связи, доступ, управления, сети, ресурсы, устройство, канал, данные, информация;

Cluster 6: алкил, формула, группа, соединение, выбран, состоящий, каждый, галоген;

Cluster 7: управления, двигателя, система, сгорания, клапана, системы, датчика, давления, мощности, датчик, устройства, выполнен, внутреннего, определения, сигнала;

— Cluster 0 - Manufacturing of materials, composite materials, mixtures.

— Cluster 1 - Data and image processing.

— Cluster 2 - Creation of aerosols for surfaces and devices.

— Cluster 3 - Medicine, cancer and disease treatment

— Cluster 4 - Transport, transport management, cameras and systems for transport.

— Cluster 5 - Communication, access and data transmission devices.

— Cluster 6 - Chemistry, compounds, halogen group and alkyl group.

— Cluster 7 - Creation of motors and sensors for them.

Results

Classes

Clusters	S1	S2	S3	…	Sn
K1	a_{11}	a_{12}	a_{13}	…	a_{1S}
K2	a_{21}	a_{22}	a_{23}	…	a_{2S}
K3	a_{31}	a_{32}	a_{33}	…	a_{3S}
…	…	…	…	…	…
Kn	a_{K1}	a_{K2}	a_{K3}	…	a_{KS}

$$Recall = \frac{\sum_s max_k\{a_{ks}\}}{(\sum_k \sum_s a_{ks} + U)}$$

$$Precision = \frac{\sum_k max_s\{a_{ks}\}}{\sum_k \sum_s a_{ks}}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Results

Validation sample

Cluster	True Data	Cluster Data
Cluster 0	194	213
Cluster 1	208	199
Cluster 2	102	234
Cluster 3	229	231
Cluster 4	44	44
Cluster 5	228	190
Cluster 6	330	269
Cluster 7	348	273

F1-score = 74.9%

As a result of the work, an algorithm was developed that clusters patent texts by keywords was developed and trained.

Thanks for attention