VIII-я Международная конференция по компьютерной обработке тюркских языков «TurkLang 2020»

Кластерный анализ текстов поурочного планирования системы «Электронное образование Республики Татарстан»

Музафарова А. И. , Минуллин Д. А.

Казанский федеральный университет

Гафарова В.Р.

Институт прикладной семиотики АН РТ

Цель работы

- Исследование возможности использования методов BigData в образовательной аналитике
- Разработка системы кластеризация текстов поурочного планирования для определения принадлежности их к соответствующему УМК
- Анализ средней успеваемости учеников для различных УМК

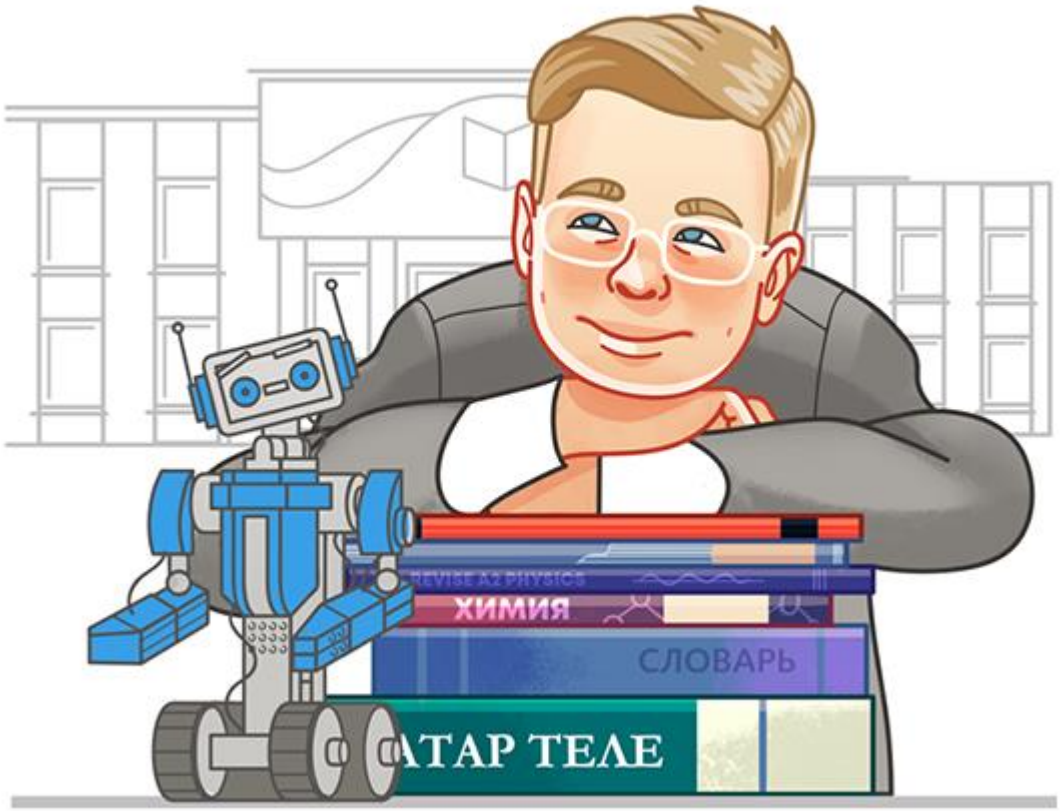
Большие данные в образовании

Государственная информационная система «Электронное образование в Республике Татарстан».

Данные об образовательном процессе:

https://edu.tatar.ru/

- Уроки (~18 Гб, >90 000 000)
- Оценки (~71 Гб, >1 000 000 000)
- Задания (3 Гб)
- Ученики (>505 000)
- Педагоги (>150 000)
- Классы
- Предметы
- ……..



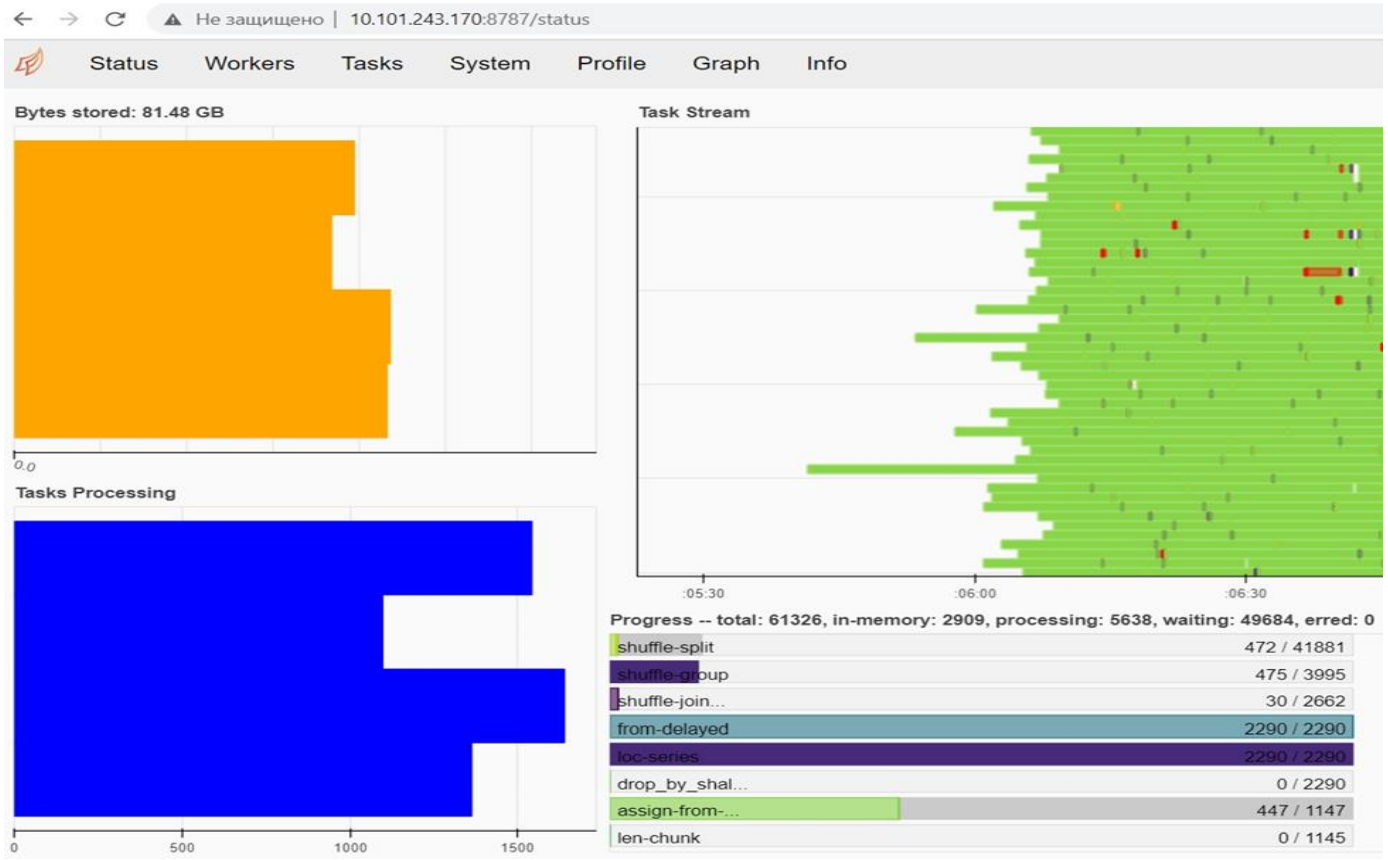
Проблема

- В системе электронное образование Республики Татарстан» данные о занятиях и успеваемости учеников не имеют привязки к УМК.
- Тексты тематического планирования (длиной 100-200 символов) учителя заполняют в личных кабинетах на основе учебников, по которым ведут уроки. Учебники имеют привязку у конкурентным УМК.
- Эти тексты не совпадают полностью с темами занятий в учебниках, поэтому невозможно выявить используемый УМК простым сравнением текстов. Также могут быть и орфографические ошибки, и тд

Примеры исходных текстов:
259187961, Первые кругосветные путешествия. Значение Солнца для жизни на Земле, 2019-02-11 16:01:19
259187962, Механизмы регулировки швейной машины, Механизмы регулировки швейной машины,2019-02-05 21:23:30
259187963, РР7 Сочинение № 1 по коллективно составленному плану на материале экскурсий, личных наблюдений, практической деятельности «Был такой случай»",2019-02-05 21:40:20
259187965Снежная королева. По Х.-К. Андерсену,Р.р. Снежная королева. По Х.-К. Андерсену,2019-02-06 19:20:52
259187967, Без чыршы бәйрәменә барабыз.“Кыш” темасын йомгаклау.,2019-02-05 21:36:57
259187970, Построение треугольников по длинам сторон. Высота в треугольнике, 2019-02-11 15:56:23

Методы Big Data для первичной обработки данных

Dask -гибкая библиотека параллельных вычислений для аналитики, предназначенная главным образом для обеспечения масштабируемости и расширения возможностей существующих пакетов и библиотек



Вычислительный кластер:

4-виртуальных машин, по 1ТБ HDD, 32 ГБ ОЗУ, 16 вычислительных ядер

Из большого объёма данных (18 Гб) выделены тексты поурочного планирования занятий с привязкой к педагогам по предмету математики отдельно по каждому классу

Результаты после обработки в Dask

Id педагога	Тема занятия
594697	Связи между скоростью временем и расстоянием
594697	Письменное умножение двузначного числа на двузначное
594697	Письменное умножение двузначного числа на двузначное Закрепление
594697	Виды треугольников
594697	Виды треугольников Решение задач

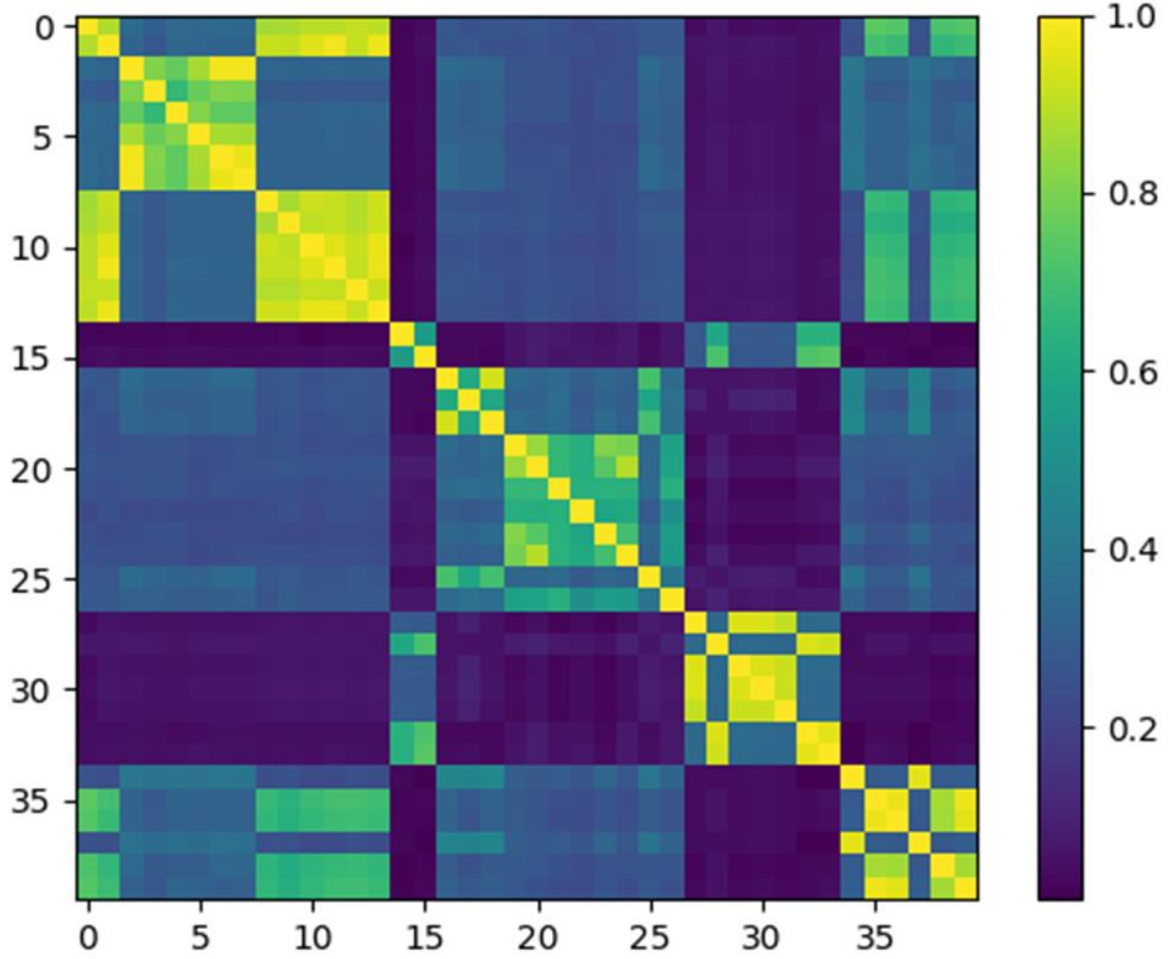
Тексты поурочного планирования каждого преподавателя, полученные на первом этапе, соединяются в единую строку, которая подвергается следующей обработке:

- Строка разбивается на токены.
- Из массива токенов удаляются: знаки препинания, пустые строки и стоп-слова, которые не придают особого значения предложению.

Это необходимо для повышения точности сравнения.

Определение схожести текстов

- Косинусное сходство — это мера сходства между двумя векторами пространства внутренних произведений, которое измеряет косинус угла между ними.



Матрица расстояний на основе матрицы схожести

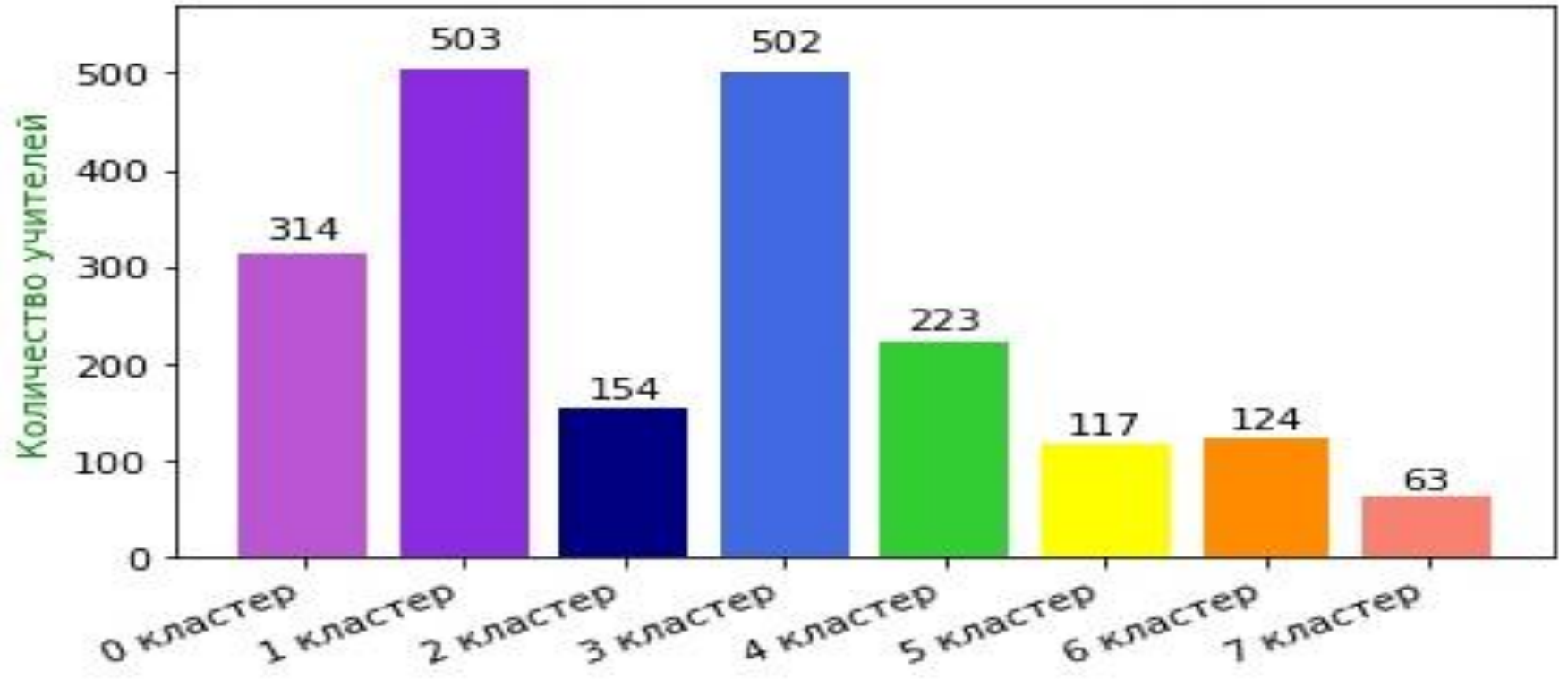
distance_matrix=1-similarity_matrix

0	0.11570488	0.64698573	0.69548666	0.65554492	0.6590529
0.11570488	0	0.66931716	0.71853311	0.66244267	0.67039118
0.64698573	0.66931716	0	0.18808854	0.23903464	0.12948035
0.69548666	0.71853311	0.18808854	0	0.32660805	0.23145201
0.65554492	0.66244267	0.23903464	0.32660805	0	0.18070803
0.6590529	0.67039118	0.12948035	0.23145201	0.18070803	0

Кластерный анализ

Проведена агломеративная кластеризация на основе матрицы расстояний. Использована библиотека **sklearn**, метод **AgglomerativeClustering.**

```
cluster = AgglomerativeClustering(affinity='euclidean', linkage='ward',n_clusters=clusters_count)
cluster.fit(distance_matrix)
```



Предполагаемые УМК кластеров и средние оценки учеников

№ кластера	Количество учителей	Язык кластера	Средняя оценка учеников
0	314	Русский	4.01
1	503	Русский	3.95
2	154	Русский	3.97
3	502	Русский	3.90
4	223	Русский	3.97
5	117	Русский	4.03
6	124	Татарский	3.93
7	63	Татарский	4.00

Выводы

- Разработан программный комплекс для автоматической обработки текстов с целью определения соответствующего УМК на основе технологий Big Data.
- По предмету «математика» выделены 8 кластеров, два из которых на татарском языке преподавания
- Для каждого кластера найден предполагаемый учебно-методический комплекс
- Проведён сравнительный анализ успеваемости учеников в зависимости от УМК.
- Разработанный программный комплекс можно использовать для автоматической обработки текстов поурочного планирования в целях определения УМК для всех предметов и классов.

Спасибо за внимание!