Министерство образования и науки Российской Федерации

**федеральное государственное автономное**

**образовательное учреждение**

**высшего образования**

**«Казанский (Приволжский) федеральный университет»**

Институт Вычислительной математики и информационных

технологий

Кафедра анализа данных и исследования операций

**Отчет о прохождении производственной практики**

**Работу выполнил :** студент 2 курса магистратуры

группы 09-515

**Специальность:** Прикладная математика и информатика

Бадыков Айрат Сагитович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(подпись)*

**Научный руководитель:**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*(подпись)*

Казань – 2017

Тема дипломной работы

Кластеризация населенных пунктов республики Татарстан на говоры на основе морфологических атрибутов.

Описание дипломной работы

В 2012 году коллективом ученых АН РТ и КФУ был создан электронный атлас татарских народных говоров. Атлас базировался на материалах печатной версии, изданной в 1989 году ИЯЛИ АН РТ и неопубликованных материалах по сибирскому диалекту.

В атласе была предложена экспертная классификация татарских говоров, созданная лингвистами ИЯЛИ АН РТ. Цель данной работы заключалась в исследовании возможности применения алгоритмов многомерного статистического анализа для классификации говоров татарского языка, проведения классификации говоров для населенных пунктов РТ, анализ полученных результатов и их сравнение с экспертным вариантом, предложенным лингвистами ИЯЛИ АН РТ.

На практику было дано задание изучения языковых явлений, представленных в электронном атласе татарских говоров, а также изучения различных алгоритмов кластеризации, описанных в научных работах и знакомство со статистическими пакетами для многомерного статистического анализа данных.

Результаты

За время практики было изучена обширная литература по кластеризации данных применительно к лингвистическим и экономическим задачам, проанализированы различные алгоритмы кластеризации данных, проанализированные различные варианты выбора метрик в пространстве языковых признаков и приняты решения по возможности использования тех или иных подходов для решения поставленной задачи.

# Описание алгоритма CLOPE.

В последнее десятилетие ведутся активные исследования в области разработки масштабируемых алгоритмов кластеризации категорийных и транзакционных данных.

Одним из эффективных считается алгоритм CLOPE, который основан на оптимизации некоторого глобального критерия. Этот глобальный критерий использует параметр поддержки.

CLOPE предложен в 2002 году группой китайских ученых

В основе алгоритма кластеризации CLOPE лежит идея максимизации глобальной функции стоимости, которая повышает близость транзакций в кластерах при помощи увеличения параметра кластерной гистограммы. Рассмотрим простой пример из 5 транзакций: {(a,b), (a,b,c), (a,c,d), (d,e), (d,e,f)}. Представим себе, что мы хотим сравнить между собой следующие два разбиения на кластеры:

(1) {{ab, abc, acd}, {de, def}}

(2) {{ab, abc}, {acd, de, def}}.

Для первого и второго вариантов разбиения в каждом кластере рассчитаем количество вхождений в него каждого элемента транзакции, а затем вычислим высоту (H) и ширину (W) кластера. Например, кластер {ab, abc, acd} имеет вхождения a:3, b:2, c:2 с H=2 и W=4.

Качество двух разбиений оценим, проанализировав их высоту H и ширину W. Кластеры {de, def} и {ab, abc} имеют одинаковые гистограммы, следовательно, равноценны. Гистограмма для кластера {ab, abc, acd} содержит 4 различных элемента и имеет площадь 8 блоков (H=2.0, H/W=0.5), а кластер {acd, de, def} – 5 различных элементов с такой же площадью (H=1.6, H/W=0.32). Очевидно, что разбиение (1) лучше, поскольку обеспечивает большее наложение транзакций друг на друга (соответственно, параметр H там выше).

На основе такой очевидной и простой идеи геометрических гистограмм и работает алгоритм CLOPE (англ.: Clustering with sLOPE).

# Реализация.

Была произведена имплементация данного алгоритма на функциональном языке программирования Elixir по оригинальной статье авторов алгоритма Yiling Yang, Xudong Guan и Jinyuan You (<http://www.inf.ufrgs.br/~alvares/CMP259DCBD/clope.pdf>).

Были написаны тесты для валидации корректности выполнения (в том числе интеграционные). В качестве сервиса для непрерывной интеграции (CI) был использован<https://semaphoreci.com/>.

Интерфейс реализации имеет следующий вид:

iex> input = [  
 {"transaction1", ["object1", "object2", "object3"]},  
 {"transaction2", ["object1", "object5"]},  
 {"transaction3", ["object2", "object3"]},  
 {"transaction4", ["object1", "object5"]}  
 ]  
 iex> result = input |> Clope.clusterize(2)  
 [  
 [  
 {"transaction1", ["object1", "object2", "object3"]},  
 {"transaction3", ["object2", "object3"]}  
 ],  
 [  
 {"transaction2", ["object1", "object5"]},  
 {"transaction4", ["object1", "object5"]}  
 ]  
 ]

На вход принимается массив кортежей вида {название транзакции, [объекты транзакции] } и вспомогательный параметр repulsion, необходимый для функции качества и определяющий близость кластеров.

На выход получаем массив кластеров (классов) из тех же кортежей.

Исходный код, реализованного мною алгоритма, находится в открытом доступе и доступен по адресу<https://github.com/ayrat555/clope> .

# Результаты.

В исходных данных содержатся 49 морфологических атрибутов (карт), каждый атрибут содержит от 2 до 5 возможный значений. Каждое значение атрибута рассматривалось как отдельная категория (транзакция). Таким образом, населенный пункт может иметь около 160 различных категорий.

## Определение диалектов.

### Экспертные данные

В Татарстане всего 3 диалекта:

1. Средный – 205 нас. пунктов

2. Западный – 66 нас. пунктов

3. Переходный – 2 нас. пункта

### Полученные результаты

Параметр repulsion = 1

1. 1-ый кластер – 220 нас. пунктов

2. 2-ой кластер - 51 нас. пунктов

### Предварительный анализ

Отличия населенных пунктов Среднего диалекта от 1-ого кластера 17 населенных пунктов, причем все населенные пункты из Среднего диалекта присутствуют в 1-ом кластере, также 2 населенных пункта из Переходного говора попали в 1-ый кластер.

Нужно отметить, что отличие западного диалекта от полученного первого кластера только в населенных пунктах Чистопольского говора. В основном совпадение полное.

Во втором полученном кластере присутствуют населенные пункты из Западного диалекта и нет населенных пунктов из других классов экспертного разбиения, т.е они полностью пересекаются.

## Определение говоров.

### Экспертные данные

В Татарстане всего 18 говоров.

1. Диалект: Средний. Говор: Балтачевский. 21 нп

2. Диалект: Средний. Говор: Заказанско-кряшенский. 7 нп

3. Диалект: Средний. Говор: Нижнекамско-кряшенский. 21 на

4. Диалект: Средний. Говор: Мамадышский. 30 нп

5. Диалект: Средний. Говор: Мензелинский. 71 нп

6. Диалект: Средний. Говор: Дубъязский. 17 нп

7. Диалект: Средний. Говор: Лаишевский. 12 нп

8. Диалект: Средний. Говор: Нурлатский. 12 нп

9. Диалект: Средний. Говор: Камско-устьинский. 6 на

10. Диалект: Средний. Говор: Камско-устьинский. 8 нп

11. Диалект: Западный. Говор: Подберезинский. 2

12. Диалект: Западный. Говор: Чистопольский. 47 нп

13. Диалект: Западный. Говор: Мордва-каратаевский. 1 нп

14. Диалект: Западный. Говор: Чистопольско-кряшенский. 2 нп

15. Диалект: Западный. Говор: Стерлитамакский. 2 нп

16. Диалект: Западный. Говор: Дрожжановский. 12 нп

17. Диалект: Группа переходных говоров. Говор: Тарханский-камско-устьинский. 1 нп

18. Диалект: Группа переходных говоров. Говор: Мамадышский-лаишский. 1 нп

### Полученные результаты (repulsion = 2)

1. 1-ый кластер 2 нп

Салихово, Новые Чути

2. 2-ой кластер 2 нпи

Хозесаново, Старое Тябердино

3. 3-ий кластер 5 нп

Старый Юмралы, Татарские Саралы, Ташкирмень, Мордовский Каратай, Азимово-Курлебаш

4. 4-ый кластер 11 нп

Большие Ачасыры, Бурундуки, Молвино, Сатламышево, Кугеево, Мамадыш-Акилово, Большие Кушманы, Акзигитово, Старые Чечкабы, Шамбулыхчы, Бакрче

5. 5-ый кластер 5 нп

Большой Битаман, Тавель, Нижний Шандер, Шеморбаш, Козяково-Челны

6. 6-ый кластер 173 нп

Кугарчино, Янга-Сала, Татарское Ходяшево, Кибячи, Старые Бугады, Тюгеевка, Урмышла, Ильтень-Бута, Чубуклы, Сарапала, Маметьево, Савалеево, Светлое Озеро, Калейкино, Васильевка, Сармашбаш, Абдрахманово, Верхние Шипки, Старое Утяганово, Девятерня, Исенбаево, Ямурзино, Крынды, Янга-Аул, Биктово, Табарле, Тойгузино, Камаево, Татарская Тумбарла, Новое Алимово, Кзыл-Яр, Сарабикулово, Федотовка, Акбаш, Дым-Тамак, Нижние Чершилы, Иж-Бобья, Янчиково, Кряш-Шуран, Старое Шугурово, Сугушла, Татар. Шуган, Старый Каразерик, Салауз-Мухан, Учалле, Нижний Табын, Баюково, Старые Карамалы, Байряки-Тамак, Верхнее Стярле, Сарлы, Каракашлы, Карамалы, Асеево, Большой Сухояш, Тойгильдино, Балтачево, Масягутово, Кавзияково, Мальбагуш, Иляксаз, Сулеево, Бишмунча, Новый Имян, Старое Байсарово, Старое Сафарово, Татарские Суксы, Старое Курмашево, Атясево, Исансупово, Семяково, Старый Иркеняш, Аю, Ст.Мелькен, Верхние Юшады, Сарсаз Такермень, Ляки, Средний Багряж, Кабан-Бастрык, Кадырово, Бурды, Средние Пинячи, Мелекес, Верхний Такермен, Калмия, Мунайка, Рысово, Верхнее Абдулово, Янга-Аул, Нижние Гордалы, Штырь, Псеево, Бизяки, Старое Гришкино, Абалачи, Морты, Большой Шурняк, Альвидино, Большой Салтан, Большой Ошняк, Татарский Янтык, Большая Елга, Ямашурма, Шали, Пелево, Именьково, Кирби, Курманаково, Татарский Кабан, Нижний Искубаш, Кугунур, Арбор, Нуринер, Верхний Субаш, Янгулово, Мамашир, Бурбаш, Норма, Карадуван, Туембаш, Байлянгар, Сиза, Верхняя Корса, Измя, Евлаштау, Псяк, Урясьбаш, Кильдебяк, Старый Мичан, Большие Верези, Сардебаш, Апазово, Средний Пшалым, Мендюш, Качелино, Байкал, Ашитбаш, Утар-Аты, Большой Менгер, Ключи-Сап, Кубян, Коморгузя, Ташлы-Ковали, Ислейтар, Малый Сулабаш, Альдермыш, Большие Ковали, Большие Кургузи, Березняк, Большой Сардек, Чишма, Куюк Ерыкса, Крещеный Пакшин, Верхняя Ошма, Нижний Таканыш, Средние Кирмени, Уразбахты// Татарский Омар, Тяберлины Челны, Янцевары, Большие Кибя-Кози, Кукча, Ишкеево, Верхняя Сунь, Нижняя Сунь, Ятмас Дусай, Татарская Икшурма, Албай, Тюбяк, Ядыгерь, Сатышево, Баландыш, Субаш, Новый Арыш

7. 7-ой кластер 75 нп

Ерыклинский, Мокрые Курнали, Татарские Шатрашаны, Старые Чукалы, Большая Цильна, Старая Задоровка, Мочалей, Старое Шаймурзино, Новые Ишли, Татарские Тюки, Энтуганы, Чутеево, Бикулово, Нижний Нурлат, Савиново, Кривое Озеро, Курманаево, Ст. Баран//Иске Рязап, Новый Баран, Нижняя Каменка, Лашманка, Татар. Волчье, Нов. Кадеево, Шахмайкино, Старое Ибрайкино, Старые Киязлы, Исляйкино, Средние Тиганы, Старое Алпарово, Старое Ямкино, Сухие Курнали, Новые Челны, Нижнее Алькеево, Татарское Ахметьево, Каргополь, Тяжбердино, Татарское Бурнаево, Ямбухтино, Старые Салманы, Татарское Тюгульбаево, Татарская Тахтала, Измери, Средний Юрткуль, Большие Кармалы, Большие Янгасала, Балтачево, Большие Болгояры, Большие Атряси, Большая Турма, Старое Барышево, Чуру-Барышево, Татарская Беденьга, Большая Карланга, Деушево, Черки-Кильдуразы, Старый Студенец, Сюндюково, Старые Тинчали, Бик-Утеево, Верхние Лащи, Новое Узеево, Тубулга, Татарский Елтан, Старая Киреметь, Новое Демкино, Сунчелеево, Муслюмкино, Акбулатово, Ст.Татар. Адам, Верхняя Татарская Майна, Новая Амзя, Степная Шентала, Кутлушкино, Кульбаево-Мараса, Кичкальня

### Предварительный анализ

1-ый кластер полностью совпадает со Стерлитамакским говором

2-ой кластер полность совпадает с Подберезинским говором

4-ый полностью совпадает с Нурлатским говором

6-ой кластер состоит из говоров Среднего диалекта

7-ой кластер в основном состоит из говоров Западного диалекта

### Полученные результаты (repulsion = 3)

1. 9 нп

Большие Ачасыры, Бурундуки, Молвино, Сатламышево, Кугеево, Мамадыш-Акилово, Большие Кушманы, Акзигитово, Старые Чечкабы

2. 2 нп

Салихово, Новые Чути

3. 8 нп

Татарская Тумбарла, Кзыл-Яр, Сарабикулово, Федотовка, Акбаш, Дым-Тамак, Нижние Чершилы, Сугушла

4. 65 нп

Старые Бугады, Тюгеевка, Урмышла, Ильтень-Бута, Чубуклы, Сарапала, Маметьево, Савалеево, Светлое Озеро, Калейкино, Васильевка, Сармашбаш, Абдрахманово, Верхние Шипки, Старое Утяганово, Ямурзино, Новое Алимово, Кряш-Шуран, Татар. Шуган, Старый Каразерик, Салауз-Мухан, Учалле, Нижний Табын, Баюково, Старые Карамалы, Байряки-Тамак, Верхнее Стярле, Сарлы, Каракашлы, Карамалы, Асеево, Большой Сухояш, Тойгильдино, Балтачево, Масягутово, Кавзияково, Иляксаз, Сулеево, Бишмунча, Новый Имян, Старое Байсарово, Старое Сафарово, Татарские Суксы, Старое Курмашево, Атясево, Исансупово, Семяково, Старый Иркеняш, Аю, Ст.Мелькен, Верхние Юшады, Сарсаз Такермень, Ляки, Средний Багряж, Кабан-Бастрык, Кадырово, Бурды, Средние Пинячи, Мелекес, Верхний Такермен, Калмия, Верхнее Абдулово, Нижние Гордалы, Морты, Большой Шурняк

5. 2 нп

Мальбагуш, Мокрые Курнали

6. 7 нп

Татарские Шатрашаны, Старые Чукалы, Большая Цильна, Мочалей, Старое Шаймурзино, Новые Ишли, Татарские Тюки

7. 2 нп

Хозесаново, Старое Тябердино

8. 1 нп

Тавель

9. 1 нп

Мордовский Каратай

10. 16 нп

Девятерня, Исенбаево, Крынды, Янга-Аул, Биктово, Табарле, Тойгузино, Камаево, Иж-Бобья, Мунайка, Рысово, Псеево, Бизяки, Старое Гришкино, Абалачи, Чишма

11. 25 нп

Старый Юмралы, Ст. Баран//Иске Рязап, Новый Баран, Татарское Бурнаево, Старые Салманы, Измери, Средний Юрткуль, Большие Кармалы, Азимово-Курлебаш, Большие Янгасала, Балтачево, Большие Болгояры, Большие Атряси, Большая Турма, Старое Барышево, Чуру-Барышево, Татарская Беденьга, Большая Карланга, Шамбулыхчы, Деушево, Черки-Кильдуразы, Сюндюково, Бакрче, Бик-Утеево, Верхние Лащи

12. 32 нп

Янга-Сала, Янга-Аул, Ямашурма, Кугунур, Арбор, Нуринер, Верхний Субаш, Янгулово, Мамашир, Бурбаш, Карадуван, Туембаш, Байлянгар, Большие Верези, Сардебаш, Апазово, Средний Пшалым, Байкал, Ашитбаш, Большой Менгер, Ключи-Сап, Кубян, Коморгузя, Ташлы-Ковали, Ислейтар, Малый Сулабаш, Альдермыш, Большой Битаман, Большие Ковали, Большие Кургузи, Большой Сардек, Ядыгерь

13. 56 нп

Кугарчино, Татарское Ходяшево, Кибячи, Янчиково, Штырь, Альвидино, Большой Салтан, Большой Ошняк, Татарский Янтык, Большая Елга, Шали, Пелево, Именьково, Кирби, Курманаково, Татарские Саралы, Татарский Кабан, Нижний Искубаш, Норма, Сиза, Верхняя Корса, Измя, Евлаштау, Псяк, Урясьбаш, Кильдебяк, Старый Мичан, Мендюш, Качелино, Утар-Аты, Березняк, Ташкирмень, Нижний Шандер, Куюк Ерыкса, Крещеный Пакшин, Верхняя Ошма, Нижний Таканыш, Средние Кирмени, Уразбахты// Татарский Омар, Тяберлины Челны, Янцевары, Большие Кибя-Кози, Кукча, Ишкеево, Верхняя Сунь, Нижняя Сунь, Ятмас Дусай, Татарская Икшурма, Албай, Тюбяк, Шеморбаш, Козяково-Челны, Сатышево, Баландыш, Субаш, Новый Арыш

14. 47 нп

Старое Шугурово, Ерыклинский, Старая Задоровка, Энтуганы, Чутеево, Бикулово, Нижний Нурлат, Савиново, Кривое Озеро, Курманаево, Нижняя Каменка, Лашманка, Татар. Волчье, Нов. Кадеево, Шахмайкино, Старое Ибрайкино, Старые Киязлы, Исляйкино, Средние Тиганы, Старое Алпарово, Старое Ямкино, Сухие Курнали, Новые Челны, Нижнее Алькеево, Татарское Ахметьево, Каргополь, Тяжбердино, Ямбухтино, Татарское Тюгульбаево, Татарская Тахтала, Старый Студенец, Старые Тинчали, Новое Узеево, Тубулга, Татарский Елтан, Старая Киреметь, Новое Демкино, Сунчелеево, Муслюмкино, Акбулатово, Ст.Татар. Адам, Верхняя Татарская Майна, Новая Амзя, Степная Шентала, Кутлушкино, Кульбаево-Мараса, Кичкальня

### Предварительный анализ

1-ый кластер совпадает с Нурлатским говором

2-ой кластер совпадает со Стерлитамакским говором

3-ий кластер состоит из населенных пунктов Мензилинского говора

4-ый кластер включает населенные пункты Нижнекамско-кряшенского говоры

Остальные кластеры отличны от экспертных говоров.