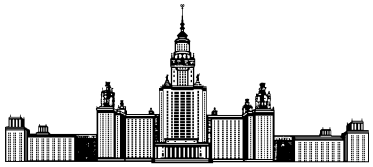
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

# ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

## «Методы обучения без учителя для выделения поляризаций в новостных потоках»
## «Unsupervised polarization detection methods in newsflows»

**Выполнил**:

студент 4 курса 417 группы

*Висков Василий Алексеевич*

**Научный руководитель**:

д.ф-м.н., и.о. заведующего кафедры

*Воронцов Константин Вячеславович*

Москва, 2022

# Содержание

**Аннотация**

В данной работе исследуется задача на стыке **Polarization Detection** и **Stance Detection**, предлагается алгоритм разбиения выборок документов, связанных одним общим событием, на непересекающиеся подмножества, описывающие мнения, с нефиксированным их количеством методами обучения без учителя и предлагается способ формирования поляризованных размеченных выборок.

# 1 Введение

Важные общественные события, связанные с политикой, спортом или другими насущными темами, освещаются рядом журналистских служб, и зачастую публикуемые новости описывают произошедшее с разных позиций, ссылаясь либо на одну из противоборствующих сторон, либо вообще субъективно оценивая ситуацию. В рамках новостной полемики вокруг некоторого события могут возникать различные мнения, зачастую выраженные не столько используемой лексикой, сколько эмоциональной окраской и степенью эксцентричности, с которой смысл текста раскрывается. Причем этих мнений может быть произвольное количество в зависимости от степени общественного резонанса в рамках новостного фона.

Рассмотрим несколько примеров, демонстрирующих поляризованность новостного потока вокруг некоторого события.

1. Тема: «Разрушение Израилем офисного здания в секторе Газа»

   - Полюс 1: «Взгляд одной из телевизионных компаний, базирующейся в этом здании»
     - «В «Аль-Джазире» прокомментировали уничтожение Израилем своего офиса в Газе. Там подчеркнули, что эти действия являются «варварским актом» который направлен на то, чтобы помешать «говорить правду».
   - Полюс 2: «Взгляд израильской стороны»
     - «Израиль нанес удары по второму высотному зданию в Газе, сообщил источник. Как позднее пояснили в Армии обороны Израиля, в этом доме находилось военное оборудование разведки ХАМАС, а офисы медиакомпаний движение использовало как живой щит».
   - Полюс 3: «Нейтральный вгляд/констатация факт»
     - «Израиль разрушил здание в Газе с офисами международных СМИ. Авиаудар, нанесенный Израилем по сектору Газа, привел к обрушению 15-этажного здания. Там, в частности, располагались офисы международных СМИ. Речь идет о высотном здании «аль-Джала».

2. Тема: «Инцидент с пермскими подростками и таксистом»

   - Полюс 1: «Произвол/негатив»
     - «Пермские подростки изрезали водителя такси».
   - Полюс 2: «Замалчивание»
     - «В Перми студенты колледжа ПГНИУ катались в машине с раненым таксистом».
   - Полюс 3: «Констатация факта»
     - «Трое студентов колледжа в Перми стали фигурантами дела о нападении с ножом на таксиста».

В данной работе будут рассмотрены способы построения кластеризации для произвольной выборки документов из новостного потока, объединенных описанием одного события, где кластер, или полюс, характеризует поляризованное или нейтральное мнение, а также шумовое подмножество документов не о событии выборки. Мнение должно быть связано с одним и только одним событием и может характеризоваться не только конкретной позицией о каком-то вопросе, но и степенью эксцентричности, с которой эта позиция раскрывается. Также будет построена совокупность размеченных валидационных выборок на русском языке, состоящих из подмножеств документов, объединенных одним событием и включающих тексты, описывающие в совокупности от 1 до 10 различных мнений.

## 1.1 Основные определения

**Разбиение множества** – это представление его в виде объединения произвольного количества попарно непересекающихся непустых подмножеств.

**Темой** будем называть множество документов, большинство из которых описывает одно и то же событие. Это событие будем называть **событием** темы.

**Поляризованное мнение** (**мнение**), или **полюс поляризации** – эмоционально, синтаксически или семантически однородное подмножество исходной темы, задающее некоторую позицию вокруг события. Будем считать, что нейтральное мнение также является полюсом поляризации.

Документы темы, не относящиеся к его событию, будем называть **шумовыми**. Совокупность мнений и шумовых документов образуют разбиение темы.

**Привилегированная информация** — дополнительная информация об объектах выборки, доступная только на этапе обучения. Таковой в данной работе выступает разметка подмножества генеральной совокупности тем, коими являются и предлагаемые в дальнейшем к разметке наборы документов. Эти темы будут использоваться исключительно для измерения качества разбиения их на полюсы поляризации.

**Тонально окрашенными** будем называть слова или словосочетания, выражающие эмоциональное отношение автора высказывания к некоторому объекту, выраженном в тексте.

**Тональностью** слова или словосочетания будем называть его степень тональной окраски и определять целочисленной величиной.

**Именованной сущность** называют слово или словосочетание, четко идентифицирующее один элемент из набора других элементов, имеющих аналогичные свойства. Примерами именованных сущностей могут быть ФИО, названия мест и компаний, т.д.

**Эмбеддинг** (*от англ. embedding*) – веществennознachный вектор некоторой размерности, выступающий признаковым описанием объекта (в данной работе – текста).

**Матрицей объектов-признаков** выборки - мощности # называется вещественная матрица / 2 R$^{\#}$ , состоящая по строкам из векторных представлений / 2 R объектов выборки.

## 1.2 Исходные данные

Предоставленными данными служит совокупность документов из новостного потока X. Внешним алгоритмом по отношению к этой работе они разбиты на совокупность тем f -$_{8}$g$_{8}^{\%}$, где % = 90. Каждая из тем – -$_{8}$ = f$G_{9}^{8}$ j $G_{9}^{8}$ = f 5$_{1}$-••• 5 g$_{1}^{j-8j}$, где 5$_{8}$ - один из признаков документа. Темы описывают события из следующих новостных рубрик:

- происшествия (кол-во: 42);
- наука и техника (кол-во: 12;
- политика (кол-во: 27);
- культура (кол-во: 3);

ˆ ñèëîâûå ñòðóêòóðû (êîë-âî: 3); ˆ ýêîíîìèêà/ôèíàíñû (êîë-âî: 3);

Ñîáûòèÿìè ìîãóò ñëóæèòü, íàïðèìåð, ¾àñòðîíîìû NASA âïåðâûå îáíàðóæèëè ðåíòãåíîâñêèå ëó÷è îò óðàíà¿ èëè ¾Áàéäåí îòêàçàëñÿ îò âñòðå÷è ñ Çåëåíêîì ïåðåä ðàçãîâîðîì ñ Ïóòèíûì¿.

Ðèñ. 1: Èññëåäîâàíèå äàííûõ

Ñðåäíÿÿ äëèíà äîêóìåíòà ñ ó÷åòîì çàãîëîâêà è òåëà ñîñòàâëÿò ïðèìåðíî 1388 ñèìâîëîâ èëè 188 ñëîâ. Èçâåñòíî [1], ÷òî ñðåäíÿÿ äëèíà ïóáëèêàöèé â ñåòè Twitter ñîñòàâëÿåò 28 ñèìâîëîâ, ìàêñèìàëüíàÿ äëèíà ïðèìåðíî 140 ñèìâîëîâ, ÷òî íå èäåò íè â êàêîå ñðàâíåíèå ñ èìåþùèìñÿ äàííûìè.

Êîëè÷åñòâî äîêóìåíòîâ â òåìàõ ñèëüíî ñìåùåíî â ñòîðîíó íåáîëüøèõ âåëè÷èí, äëÿ íåáîëüøèõ âûáîðîê ïðàâèëüíîé êëàñòåðèçàöèè ìîæåò áûòü ñëîæíîé çàäà÷åé.

Ïðèçíàêàìè äîêóìåíòîâ âûñòóïàþò, ïîìèìî çàãîëîâêà è òåëà äîêóìåíòà, è ðåçóëüòàòû îáðàáîòêè ñîâîêóïíîñòüþ ìîäåëåé ïîíèìàíèÿ åñòåñòâåííîãî ÿçûêà, âíåøíèõ ïî îòíîøåíèþ ê ýòîé ðàáîòå, à èìåííî: ðàçìåòêà òåëà äîêóìåíòà íà èìåíîâàí-

[1] https://smk.co/the-average-tweet-length-is-28-characters-long-and-other-interesting-facts/

íûå ñóùíîñòè ñ ïðîñòàâëåíèåì êàòåãîðèé ¾ëè÷íîñòü¿, ¾ãåîãðàôè÷åñêîå ïîëîæåíèå¿ èëè ¾îðãàíèçàöèÿ¿, ðàçìåòêà òîíàëüíîñòè äëÿ èìåíîâàííûõ ñóùíîñòåé è ñîöèàëüíî-äåìîãðàôè÷åñêèå ïðèçíàêè, ñâÿçàííûå ñ ôîðìèðîâàíèåì öåëåâîé àóäèòîðèè äëÿ êàæäîãî èç äîêóìåíòîâ. Ïîñëåäíèå ïðèçíàêè ìîæíî ðàçáèòü íà ãðóïïû:

- âåðîÿòíîñòü òîãî, ÷òî öåëåâîé àóäèòîðèåé äîêóìåíòà ÿâëÿåòñÿ íåêîòîðàÿ âîçðàñòíàÿ ãðóïïà (âûäåëÿþòñÿ 4 ãðóïïû: îò 0 äî 21, îò 22 äî 39, îò 40 äî 59, îò 60 è ñòàðøå);
- âåðîÿòíîñòü òîãî, ÷òî öåëåâîé àóäèòîðèåé äîêóìåíòà ÿâëÿåòñÿ ãðóïïà ëþäåé, èìåþùèõ îïðåäåëåííóþ àêàäåìè÷åñêóþ ñòåïåíü èëè îáðàçîâàíèå (âûäåëÿþòñÿ 5 ãðóïï: äîêòîðà íàóê, êàíäèäàòû íàóê, âûñøåå, ñðåäíåå îáùåå è ñðåäíåå ñïåöèàëüíîå îáðàçîâàíèå);
- âåðîÿòíîñòü òîãî, ÷òî öåëåâîé àóäèòîðèåé äîêóìåíòà ÿâëÿåòñÿ ãðóïïà ëþäåé ñ îïðåäåëåííûì åæåìåñÿ÷íûì äîñòàòêîì, âûðàæåííî â íåêîòîðûõ óñëîâíûõ åäèíèöàõ (âûäåëÿþòñÿ 4 ãðóïïû: äîñòàòîê îò 0 äî 50, îò 50 äî 100, îò 100 äî 150 è îò 150 äî 200);
- âåðîÿòíîñòü òîãî, ÷òî öåëåâîé àóäèòîðèåé ÿâëÿþòñÿ æåíùèíû èëè ìóæ÷èíû;

Òîíàëüíîñòü ÿâëÿåòñÿ ïîðÿäêîâûì ïðèçíàêîì íà äîìåíå öåëûõ ÷èñåë, ãäå 0 ñîîòâåòñòâóåò íåéòðàëüíîìó îòíîøåíèþ, ìäóëü çíà÷åíèÿ îòáðàæàåò ñòåïåíü òîíàëüíîé îêðàøåííîñòè äëÿ êàæäîé èç èìåíîâàííûõ ñóùíîñòåé, à çíàê  îòðèöàòåëüíîå èëè ïîëîæèòåëüíîå îòíîøåíèå ê îáúåêòó ïðèâÿçêè.

Êîëè÷åñòâî ìíåíèé â êàæäîé èç òåì íåèçâåñòíî, êàê ñðåäè èìåþùèõñÿ äàííûõ, òàê è âî âñåé ãåíåðàëüíîé ñîâîêóïíîñòè òåì, ïîýòîìó ðåøàòü çàäà÷ó ìåòîäàìè, êîòîðûå êàêèì-ëèáî îáðàçîì çàäåéñòâóþò ðàçìåòêó, ò.å ìåòîäàìè îáó÷åíèÿ ñ ó÷èòåëåì èëè ìåòîäàìè ÷àñòè÷íîãî îáó÷åíèÿ, íå ïðåäñòàâëÿåòñÿ âîçìîæíûì.

## 1.3 Îáçîð îáëàñòè

Â ðàáîòàõ, ïóíêòùèõñÿ ê îáëàñòè  Polarization Detection è áëèçêîé ê íåé  Stance Detection, àâòîðû èñïîëüçóþò ðàçëè÷íûå ïîñòàíîâêè ðåøàåìîé çàäà÷è: ïîñòàíîâêó ñ ó÷èòåëåì [13][6], ïîñòàíîâêó ÷àñòè÷íîãî îáó÷åíèÿ [8] è ðåëåâàíòíóþ äàííóé ðàáîòå

ïîñòàíîâêó áåç ó÷èòåëÿ  ïîñëåäíþþ ðàñøèðåíèå ïîäõîäîâ. Ðàçìåòêà âûáîðî  ðåñóðñîçàòðàòíûå äåéñòâèå. Áîëåå òîãî, áîëüøèíñòâî ñóùåñòâóþùèõ ðåøåíèé çàäà÷è Polarization Detection íà îñíîâå ðàçìå÷åííûõ íàáîðîâ äîêóìåíòîâ Twitter èñïîëüçóþ êîíòðîëèðóåìûå ïîäõîäû, êîòîðûå, êàê áûëî ïîêàçàíî, íå ïîääàþòñÿ îáîáùåíèþ è ñòðàäàþò îò îãðàíè÷åííîé äîñòóïíîñòè ê ãåíåðàëüíîé ñîâîêóïíîñòè è ê íàäåæíûì íàáîðàì äàííûõ, ïîäòâåðæäàþùèõ äîñòîâåðíîñòü ïîëó÷åííûõ ðåçóëüòàòîâ [4].

Îñíîâíîé èäååé ðàáîò, â êîòîðûõ ïðèìåíÿëîñü îáó÷åíèå áåç ó÷èòåëÿ, ìîæíî ñ÷èòàòü ðåøåíèå çàäà÷è ïîèñêà ëàòåíòíîãî âåêòîðíîãî ïðîñòðàíñòâà ìàëîé ðàçìåðíîñòè, â êîòîðîì âûãîäíåå ïðîèçâîäèòü ïîèñê áëèçêèõ ñîñåäåé ñåòè Twitter è âåêòîð ïîëèòè÷åñêèõ ñòîðîí ïîñðåäñòâîì èñïîëüçîâàíèÿ êîíèðîâàíèÿ ïîëüçîâàòåëåé ñòîðîííèõ ïóáëèêàöèé ñ óêàçàíèåì ññûëêè íà íèõ êàê îòêëèêîâ è ñîîîñòàâëåíèÿ èõ ïóáëèêàöèé ñ ïóáëèêàöèÿìè ïîëèòè÷åñêèõ ñòîðîí. Â ðàáîòå [16] ïðèìåíÿåòñÿ çàñëóæèâàþùèé âíèìàíèÿ ïîäîä ñ îáó÷åíèåì ïðåäñòàâëåíèé âçãëÿäîâ ìåòîäîì áåç ó÷èòåëÿ (unsupervised belief representation learning, êàê ýòî ïîñòðîåíèå ëàòåíòíîãî ïðîñòðàíñòâà ñàìè èìåþò àâòîðû), ñ ïîìîùüþ âàðèàöèîííîãî àâòîêîäèðîâùèêà íàä äâóäîëüíûì íåîðèåíòèðîâàííûì ãðàôîì ïîëüçîâàòåëåé è êîíòåíòà ñåòè Twitter, ãðàô ïîëüçîâàòåëüñêèõ ìàòðèöå ñìåæíîñòè, âåñà êîòîðîé ñòðîèëèñü áëî÷íûì îáðàçîì. Â ñðàâíåíèè ñî ìíîãèìè àëãîðèòìàìè, íåêîòîðûå èç êîòîðûõ ñïåöèôè÷íû äëÿ äàííûõ Twitter, ýòà ìîäåëü ïîêàçûâàåò íàèëó÷øåå ðåçóëüòàòû ñðåäè ïîäõîäîâ îáó÷åíèÿ áåç ó÷èòåëÿ.

Âñå ýòè ðàáîòû òàê èëè èíà÷å îáúåäèíÿåò íàëè÷èå íà ýòàïå îáó÷åíèÿ íåêîòîðîãî ðàçìå÷åííîãî âàëèäàöèîííîãî íàáîðà äîêóìåíòîâ ñ àïðèîðíî çàäàííûì èç äàííûõ ÷èñëî öåíòðîâ ïîëÿðèçàöèè, êîòîðîå âûòóïàåò â êà÷åñòâå åäèíñòâåííî âåðíîãî ãèïåðïàðàìåòðà êëàñòåðèçàöèè èëè îò êîòîðîãî ôóíêöèîíàëüíî çàâèñÿò ÿâíûå ãèïåðïàðàìåòðû àëãîðèòìîâ. Ïîñëåäíèå äâå ðàññìîòðåííûå ðàáîòû îæíî ïîñòàâèòü â òåðìèíàõ îáó÷åíèÿ áåç ó÷èòåëÿ ñ ïðèâëåãèðîâàííîé èíôîðìàöèåé. Çà÷àñòóþ àâòîðû ñòàòåé äëÿ âàëèäàöèè ìîäåëåé èñïîëüçóþò îòêðûòûå äàííûå èç ñåòè Twitter, â îäíîé èç ðàáîò èñïîëüçóþò ÿâíûå ìîäåëèðîâàíèå îòíîøåíèÿ ê ïîëèòè÷åñêîé ñòîðîíå â òðåõïëÿðíîé ïîëèòè÷åñêîé ñèñòåìå.

Çàäà÷àPolarization Detection äëÿ ðóññêîãî ÿçûêà â ïîñòàíîâêå áåç ó÷èòåëÿ ñ ïðèâëåãåðîâàííîé èíôîðìàöèåé ðåøàëàñü â ðàáîòå [26], â ýòîé ñòàòüå ïðèâåäåíà ê ïîñòðîåíèþ ïðèçíàêîâîãî ïðîñòðàíñòâà ñ ïîìùüþ ìîäåëè ìåøêà ñëîâ íàä òðèïëåòàìè ¾ñóáúåêò-ïðåäèêàò-îáúåêò¿ äëÿ ñèíòàêñè÷åñêèõ ðîëåé ñëîâ è èõ ÷àñòåé ðå÷è, òîíàëüíîñòüìè è ñåìàíòè÷åñêèìè ðîëÿìè ñëîâ ïî Ôèëìîðó [12] è ïîñëåäóþùåãî ïîñòðîåíèÿ ìÿãêîé êëàñòåðèçàöèè ïîñðåäñòâîì òåìàòè÷åñêîé ìîäåëè ARTM [25]. Äëÿ ïîëó÷åíûõ òåìàòè÷åñêèõ âåêòîðîâ ðàçìåðíîñòè 2 (êîìïîíåíòà ñîòâåòñòâóåò âåðîÿòíîñòè ïðèíàäëåæíîñòè êàêîìó-òî ïîëþñó) ïðîèçâîäèëñÿ ïåðåõîä ê æåñòêîé êëàñòåðèçàöèè ïîñðåäñòâîì îïåðàòîðà Argmax. Àâòîðû ñîñòàâèëè äâå ñîáñòâåííûå òåìû ñ äâóìÿ ïîëÿðèçîâàííûìè ìíåíèÿìè â êàæäîì è ïîëó÷èëè F1-ìåðó â ðàéîíå 0.85 äëÿ îáåèõ òåì. Åñëè èñïîëüçîâàòü íîòàöèþ äëÿ áèíàðíîé ìàòðèöû îøèáîê [11], ôîðìóëó ýòîé ìåòðèêè ìîæíî âûïèñàòü òàê:

$$_1 = \frac{)\%}{)\%\ ,\ \frac{)\%,\ \#}{2}}$$

Ðåøåíèå çàäà÷èPolarization Detection ìåòîäàìè îáó÷åíèÿ áåç ó÷èòåëÿ ñ ïðèâëåãåðîâàííîé èíôîðìàöèåé äëÿ èçíà÷àëüíî íåòèêñèðîâàííîãî ÷èñëà ìíåíèé â òåìàõ íå ðåøàëàñü íè äëÿ àíãëèéñêîãî ÿçûêà, íè äëÿ ðóññêîãî.

## 1.4 Öåëü ðàáîòû

Â äàííîé ðàáîòå òðåáóåòñÿ ñðàâíèòü ðàçíûå ïîäõîäû ê ïîñòðîåíèþ êëàñòåðèçàöèè âûáðîñ ñ íåòèêñèðîâàííûì ÷èñëîì ïîëíûé ìíåíèé, äëÿ êîòîðîé êëàñòåðû âûñòóïàëè áû â êà÷åñòâå ïîëþñîâ ïîëÿðèçàöèè. Òðåáóåòñÿ òàêæå ñðàâíèòü ìåòîäû íà ñîîòíîøåíèè âûäåëÿòü øóì â òåìàõ.

Âòîðîé öåëüþ èññëåäîâàíèÿ ÿâëÿåòñÿ ïîñòðîåíèå ìåòîäà îöåíèâàíèÿ ïîëó÷åííûõ ìîäåëåé êëàñòåðèçàöèè íà ñîîòíîøåíèè ñòðîèòü ðàçáèåíèå òåì íà ìíîæåñòâà, îòíîñÿùèåñÿ ïîëþñû ïîëÿðèçàöèè, è íà ñîîòíîøåíèè âûÿâëÿòü øóìîâîé êëàñòåð.

## 1.5 Ïîñòàíîâêà çàäà÷è ìàøèííîãî îáó÷åíèÿ

Èìåþòñÿ ñîâîêóïíîñòü òåì $X = fX_8g_1^{\%}$, êàæäàÿ èç êîòîðûõ ïîêðûâàåòñÿ îáúåêòîâ-ïðèçíàêîâ, è èõ ðàçìåðêà $Y = fY_8g_8^{\%}$. Ñòàâÿòñÿ%íåçàâèñèìûõ çàäà÷ îáó-

÷åíèÿ áåç ó÷èòåëÿ ñ ïðèâèëåãèðîâàííîé èíôîðìàöèåé: íåîáõîäèìî ïîñòðîèòü îïäåëü, ïðèíèìàþùóþ íà âõîäå âûáîðêó $X_8 = f G_9^8 j G_9^8 2 R^{8-} 9= \overline{1-j-_8 j} g 2 X$, à íà âûõîäå ïðîãíîçèðóþùóþ åå êëàññèôèêàöèþ, ò.å. îïäåëü äîëæíà âûäàâàòü ìíîæåñòâî $Y_8 = f H_9^8 j H_9^8 = \overline{1-_8-} 9= \overline{1-jX_8 j} g$.

Äëÿ èçìåðåíèÿ êà÷åñòâà ïîëó÷åííîé ìîäåëè èñïîëüçóåòñÿ ìíîæåñòâî $Y_8$. Òàê êàê ñðàâíèâàòüñÿ áóäóò ðàçëè÷íûå àëãîðèòìû êëàññèôèêàöèè, òðåáóåòñÿ âûáðàòü êàêèå-òî âíåøíèå ìåòðèêè êà÷åñòâà ïî îòíîøåíèþ ê íèì. Èìè âûñòóïàþò BCubed-òî÷íîñòü (%), BCubed-ïîëíîòà (' ) è BCubed-F-ìåðà ( ) [1].

Ïóñòü êîëè÷åñòâî äîêóìåíòîâ íåêîòîðîé ðàçìå÷åííîé òåìû - = $f G_8 g$ ðàâíà #, êàæäîìó îáúåêòó $G_8$ïîñòàâëåíà ìåòêà êëàññà $H_8 = \overline{1-=}è$ ìåòêà êëàññåðà: $_8 = \overline{1-<}$. Ìíîæåñòâî êëàññîâ îáîçíà÷èì . = f. $_8 j 8= \overline{1-=} g$, ìíîæåñòâî êëàññåðîâ = f $_8 j 8= \overline{1-<} g$. Ââåäåì ôóíêöèþ $2^1 G_8 - G_9^o = 1 » H_8 = H_9 ¼ 1 »$: $_8 = : _9 ¼$ Òîãäà ìåòðèêè êà÷åñòâà âûðàæàþòñÿ ñëåäóþùèìè ôîðìóëàìè:

$$\% = \frac{1}{\#} \tilde{O}_{G_8 2-} \frac{1}{j \; :_8 j} \tilde{O}_{G_9 2 \; :_8} 2^1 G_8 - G_9^o$$

$$' = \frac{1}{\#} \tilde{O}_{G_8 2-} \frac{1}{j. H_8 j} \tilde{O}_{G_9 2. H_8} 2^1 G_8 - G_9^o$$

$$= 2 \; \frac{\% \; '}{\%_{,} \; '}$$

Ìåòðèêà BCubed-F-ìåðà óäîâëåòâîðÿåò ÷åòûðåì ñâîéñòâàì, ýâðèñòè÷åñêè õàðàêòåðèçóþùèì ¾õîðîøóþ¿ êëàñòåðèçàöèþ ïîñðåäñòâîì îòíîñèòåëüíîñòè ðåçóëüòàòîâ.

Îòäåëüíî ïðîêîììåíòèðóåì ñâîéñòâà ¾ìåøêà ñ ìóñîðîì¿ è ¾ðàçìåð êëàñòåðà ïðîòèâîêîëåæíûõ ¾ìóñîðíûõ¿ òî÷åê âûáîðêè â ïäèí êëàñòåð è ê ìèíèìàëüíîìó ðàçíàùåíèþ ðàçíîâèäíîñòåé ìåæäó îáúåêòàìè ïîëíîãî ìíîæåñòâà â ñëó÷àå, åñëè ýòî íå âðåäèò îáùåìó îæèäàíèþ îò ðåçóëüòàòîâ êëàñòåðèçàöèè, ÷òî ñõîæèå îáúåêòû ïîìåùàþòñÿ â îäíî ìíîæåñòâî, à íåïîõîæèå â ðàçíûå. Ìåòðèêó BCubed-F-ìåðà áóäåì ñ÷èòàòü îñíîâíîé ìåòðèêîé èç âûáðàííûõ.

Ìåòðèêà BCubed-òî÷íîñòü êîëåáëåòñÿ â ïðîìåæóòêå îò $\frac{<}{\#}$ äî 1, BCubed-ïîëíîòà îò $\frac{=}{\#}$ äî 1, BCubed-F-ìåðà îò $\frac{2 \frac{=<}{=_{,} <}}{\#}$ äî 1.

Ðèñ. 2: Ñâîéñòâà ìàòðèöû êà÷åñòâà êëàñòåðèçàöèè

Îáîçíà÷èì $\gamma$ çà ¾øóìîâîé¿ êëàññ. Äëÿ ñðàâíåíèÿ ìîäåëåé íà ñïîñîáíîñòü âûäåëÿòü øóì áóäåì èñïîëüçîâàòü âèäîèçìåíåííûå èñõîäíûå ìàòðèöû. Ââåäåì $MSE_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_{ij} - \mu_{ij})^2$, $MSE_2 = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_{ij} - \mu_{ij})^2$. Îöåíèì ñïîñîáíîñòè àëãîðèòìà êëàñòåðèçàöèè âûäåëÿòü øóì íàçîâåì BCubedNoise-F-ìåðà ($BCubedNoise_F$):

$$BCubedNoise_F = 2 \frac{BCubedNoise_P \cdot BCubedNoise_R}{BCubedNoise_P + BCubedNoise_R}$$

Ìåòðèêà ïðèíèìàåò òåì áîëüøèå çíà÷åíèÿ, ÷åì ëó÷øå óäàëîñü ïîìåñòèòü øóìîâûå îáúåêòû â îäèí êëàñòåð, ïðè ýòîì íå îòíåñÿ ê íåìó íåøóìîâûå îáúåêòû. Ìàêñèìàëüíîå çíà÷åíèå 1 ñîîòâåòñòâóåò èäåàëüíîé êëàñòåðèçàöèè äëÿ øóìîâûõ äîêóìåíòîâ.

Êà÷åñòâî ìîäåëè, êàê â ñìûñëå ñïîñîáíîñòè ñòðîèòü êëàñòåðèçàöèþ òåìû, òàê è ñïîñîáíîñòè âûÿâëÿòü â íèõ øóì, áóäåì ñ÷èòàòü êàê ñðåäíåå êà÷åñòâî â êàæäîé òåìå. Ïðè ýòîì äîïîëíèòåëüíî áóäåì ðàññìàòðèâàòü îòäåëüíî ñðåäíåå êà÷åñòâî â ñðåçàõ ÷èñëà ìíåíèé â òåìàõ:

- 1 ìíåíèå;
- 2 ìíåíèÿ;
- 3 ìíåíèÿ;
- 4 ìíåíèÿ è áîëüøå;

# 2 Ðåøåíèå

Â ýòîé ñåêöèè ðàçáåðåì ïîñòàíîâêó çàäà÷è êëàñòåðèçàöèè äàííûõ âûáîðîê è ïîñòðîåíèå ïîñòàâëåííîé çàäà÷è ìàøèííîãî îáó÷åíèÿ.

## 2.1 Êëàñòåðèçàöèÿ

Ïðèíöèïèàëüíîå îòëè÷èå ïîñòàâëåííîé çàäà÷è îò òåõ, ÷òî ðåøàëèñü â ñõîæèõ ðàáîòàõ, ÿâëÿåòñÿ íàëè÷èå íåîèçâåñòíîâàííîãî ÷èñëà ïîëÿðèçîâàííûõ ìíåíèé â òåìàõ.

Áóäåì ñ÷èòàòü, ÷òî êàæäûé ýëåìåíò èñõîäíîé âûáîðêè îïèñûâàåòñÿ íåêîòîðûì âåùåñòâåííûì âåêòîðîì ðàçìåðíîñòè . Â ýòîé ðàáîòå ðàññìîòðèì áàçîâûå àëãîðèòìû âåêòîðíîé êëàñòåðèçàöèè.

### 2.1.1 Àëãîðèòìû, îñíîâàííûå íà àíàëèçå ïëîòíîñòíûõ õàðàêòåðèñòèê âûáîðêè

Ïåðâûé ðàññìàòðèâàåìûé àëãîðèòì ¾ Îñíîâàííàÿ íà ïëîòíîñòè ïðîñòðàíñòâåííàÿ êëàñòåðèçàöèÿ äëÿ ïðèëîæåíèé ñ øóìàìè ¿ (DBSCAN ) [9]. Èìååò òðè êëþ÷åâûõ ãèïåðïàðàìåòðà:

- eps: ðàññòîÿíèå, ïîðåäåëÿþùåå ñîñåäñòâî òî÷åê (äâå òî÷êè îáúÿâëÿþòñÿ ñîñåäíèìè, åñëè ðàññòîÿíèå ìåæäó íèìè ìåíüøå eps);
- min_samples: ìèíèìàëüíîå êîëè÷åñòâî òî÷åê äëÿ îïðåäåëåíèÿ êëàñòåðà;
- ôóíêöèÿ ðàññòîÿíèÿ/ìåòðèêà : ëþáàÿ ôóíêöèÿ, óäîâëåòâîðÿþùàÿ àêñèîìàì òîæäåñòâà, ñèììåòðèè è òðåóãîëüíèêà;

Íà èõ îñíîâå òî÷êà ïðîñòðàíſòâà ìîæåò áûòü ïîðåäåëåíà êàê:

- ÿäðîâàÿ òî÷êà : âîêðóã íåå â ðàäèóñå eps ïî çàäàííîé ìåòðèêå íàõîäèòñÿ min_samples òî÷åê, ó÷èòûâàÿ ñàìó ÿäðîâóþ òî÷êó;
- ãðàíè÷íàÿ òî÷êà : â ðàäèóñå eps îò íåå åñòü õîòÿ áû îäíà ÿäðîâàÿ òî÷êà, íî â òîì æå ðàäèóſå íåò ðàñſïëîæåíî ìåíüøå min_samples òî÷åê ſ ó÷åòîì ſàìîé ſãðàíè÷íîé òî÷êè;
- âûáðſſ: âſ âſſõ ſſòàëüſûſ ſëó÷àÿſ;

Êëàñòåðû ôîðìèðóþòñÿ ÿäðîâûìè òî÷êè, êîòîðûå ÿâëÿþòñÿ ñîñåäíèìè (ò.å. äîñòèæèìûìè äðóã èç äðóãà), è âñå ãðàíè÷íûå òî÷êè ýòèõ ÿäåðâûõ òî÷åê. Íåîáõîäèìûì óñëîâèåì äëÿ ôîðìèðîâàíèÿ êëàñòåðà ÿâëÿåòñÿ íàëè÷èå ïî êðàéíåé ìåðå îäíîé ÿäðîâîé òî÷êè. Õîòÿ ýòî î÷åíü ìàëîâåðîÿòíî, ó íàñ ìîæåò áûòü êëàñòåð òîëüêî ñ îäíîé îñíîâíîé òî÷êîé è åå ãðàíè÷íûìè òî÷êàìè. Àëãîðèòì ïîçâîëÿåò ìîäåëèðîâàòü ¾øóìîâîé¿ êëàñòåð â ïðîöåññå ïîñòðîåíèÿ êëàñòåðèçàöèè.

Íàáîð ðàñïîàòðàâàåìûõ ãèïåðïàðàìåòðîâ:

- eps: » 10– 1¼ìî ëîãàðèôìè÷åñêîé ñåòêå ïî îñíîâàíèþ 4;
- min_samples: »2–8¼
- ôóíêöèÿ ðàññòîÿíèÿ/ìåòðèêà : åâêëèäîâî è êîñèíóñíîå ðàññòîÿíèå;

Âòîðîé ðàñïîàòðàâàåìûé àëãîðèòì ¾ Ñäâèã ñðåäíåãî çíà÷åíèÿ ¿ (Mean shift ) [5]. Àëãîðèòì ñäâèãà ñðåäíåãî çíà÷åíèÿ ÿâëÿåòñÿ ïðîöåäóðíé äëÿ ïîñëåäîâàòåëüíîãî ìåñòîïîëîæåíèÿ ìàêñèìóìîâ (ìîä) ïëîòíîñòè âåðîÿòíîñòè, çàäàâàåìîé äèñêðåòíîé âûáîðêîé ïî ýòîé ôóíêöèè. Ìåòîä ÿâëÿåòñÿ èòåðàòèâíûì äëÿ êàæäîé òî÷êè íåêîòîðîãî èíèöèàëèçèðîâàííîãî íàáîðà öåíòðîèäîâ è çàêëþ÷àåòñÿ â ïîñëåäîâàòåëüíîì ïîèñêå ëîêàëüíîãî ñðåäíåãî íà îñíîâå ïëîòíîñòè òî÷åê â íåêîòîðîì îêíå â êàæäûé ìîìåíò âðåìåíè. Â êà÷åñòâå ãèïåðïàðàìåòðîâ âûñòóïàþò:

- ðàçìåð îêíà h;
- ìíîæåñòâî öåíòðîèäîâ , èç êîòîðûõ ñòàðòóþò èòåðàòèâíûå ïðîöåññû ïîèñêà öåíòðîâ êëàñòåðîâ;
- ôóíêöèÿ ÿäðà : îïðåäåëÿåò âåñ áëèæàéøèõ òî÷åê äëÿ ïåðåîöåíêè ñðåäíåãî;

Ëîêàëüíîå ñðåäíåå îïðåäåëÿåòñÿ êàê âçâåøåííîå ñðåäíåå òî÷åê â ðàäèóñå h ïî âåêëèäîâîé ìåòðèêå îò ðàññìàòðèâàåìîé íà òåêóùåé èòåðàöèè Còî÷êè G(ìíîæåñòâî òàêèõ òî÷åê îáîçíà÷èì çà # ¹G):

$$< {}^{1}G^{o} = \frac{\sum_{G_8 2\#_{,}{}^{1}G^{o}} {}^{1}G_8 \quad G_{\text{Ç}}{}^{o}G_8}{\sum_{G_8 2\#\,{}^{1}G^{o}} {}^{1}G_8 \quad G_{\text{Ç}}{}^{o}}$$

Àëãîðèòì ñäâèãà ñðåäíåãî çíà÷åíèÿ òåïåðü íàçíà÷àåò $G_{\text{Ç}1} = < {}^{1}G^{o}$ è ïîâòîðÿåò îïåíêó, ïîêà $< {}^{1}G^{o}$ íå ñîéä¼òñÿ.

Õîòÿ àëãîðèòì ñäâèãà ñðåäíåãî çíà÷åíèÿ øèðîêî èñïîëüçóåòñÿ âî ìíîãèõ ïðèëîæåíèÿõ, ñòðîãîå äîêàçàòåëüñòâî ñõîäèìîñòè àëãîðèòìà, èñïîëüçóþùåãî ÿäðî íáùåãî âèäà â ïðîñòðàíñòâàõ âûñîêîé ðàçìåðíîñòè, îòñóòñòâóåò.

Íàáîð ðàññìàòðèâàåìûõ ãèïåðïàðàìåòðîâ:

- ðàçìåð îêíà h: » 20–0¼ïî ëîãàðèôìè÷åñêîé ñåòêå ïî îñíîâàíèþ 4;
- ìíîæåñòâî öåíòðîèäîâ : âñå òî÷êè âûáîðêè;
- ôóíêöèÿ ÿäðà : ¾ïëîñêîå¿ ÿäðî, çàäàâàåìîå ôîðìóëîé ¹G³ = 1»jƑjj2 ¼

2.1.2 Àëãîðèòì, ÿâíî ïàðàìåòðèçóåìûå êîëè÷åñòâîì êëàñòåðîâ

Ïåðâûé ðàññìàòðèâàåìûé àëãîðèòì ñìåñü íîðìàëüíûõ ðàñïðåäåëåíèé . Ñòðîèòñÿ àïïðîêñèìàöèÿ ðàñïðåäåëåíèÿ èñõîäíîé âûáîðêè - ìîùíîñòè # íåêîòîðûì ïîäëåæàùèì ðàñïðåäåëåíèåì, ñìåñüþ ìíîãîìåðíûõ íîðìàëüíûõ ðàñïðåäåëåíèé ñ K êîìïîíåíòàìè è àïðèîðíûìè âåñàìè c: , çàäàâàåìîé ñëåäóþùèìè ôîðìóëàìè:

$$?^{1}\text{-}^{o} = \ddot{Y}^{*} \tilde{O} c_{:} N^{1} G_{8} j \grave{}_{:} - {}_{:}{}^{o} - \tilde{O} c_{:} = 1 - c_{.} \quad 0$$

$8=1 \quad :=1 \qquad\qquad :=1$

Âåäåì îáîçíà÷åíèÿ $c = f c_1 - \bullet\bullet\bullet - c g$, $\grave{} = f \grave{}_1 - \bullet\bullet\bullet - \grave{}_: g$, $= f\ _1 - \bullet\bullet\bullet - g$. Ïðåäïîëàãàåòñÿ, ÷òî êàæäûé îáúåêò Gâûáîðêè ïðèíàäëåæèò ïäíîé è òîëüêî îäíîé êîìïîíåíòå ñìåñè. Ââîäèòñÿ ãðóïïà ñêðûòûõ ïåðåìåííûõ / = f I 1 – • • • – # Ig, ïîêûâàþùèõ ïðèíàäëåæíîñòü êàæäîãî îáúåêòà íåêîòîðîé êîìïîíåíòå, $I_8\ 2\ \overline{1-}$ - íîìåð êîìïîíåíòû ñìåñè, êîòîðîé ïðèíàäëåæèò îáúåêòà G8. Àïðèîðíûå ðàñïðåäåëåíèå âåêòîðà I 8 2 / : ?¹I8jcº = Cat¹8j cº = $\hat{I}$ c:$^{I:}_{8}$. Ïàðàìåòðû ìîäåëè f c: – ` : – : g: =1 íàñòðàèâàþòñÿ ïðè ïîìîùè EM-àëãîðèòìà [7] Ãèïåðïàðàìåòðàìè àëãîðèòìà âûñòóïàþò:

- ÷èñëî êîìïîíåíò ñìåñè;
- f c– ` – g íà÷àëüíàÿ èíèöèàëèçàöèÿ ïàðàìåòðîâ ìîäåëè;

Áóäåì èñïîëüçîâàòü ñëó÷àéíóþ èíèöèàëèçàöèþ ïàðàìåòðîâ ìîäåëè, ëó÷øàÿ ñðåäè îáó÷åííûõ EM-àëãîðèòìîì ìîäåëé âûáèðàåòñÿ ïî âåëè÷èíå âàðèàöèîííîé íèæíåé îöåíêè.

Íàáîð ðàññìàòðèâàåìûõ ãèïåðïàðàìåòðîâ:

ˆ : »1–minf10–j- jg¼

ˆ fc– `– g: ñëó÷àéíàÿ èíèöèàëèçàöèÿ (40 ïîâòîðåíèé);

Âòîðîé ðàññìàòðèâàåìûé àëãîðèòì ìåòîä K-ñðåäíèõ (K-means ) [22]. Ïðåäñòàâëÿåò ñîáîé ïðàâèëüíûé ñëó÷àé ïîñòðîåíèÿ ñìåñè íîðìàëüíûõ ðàñïðåäåëåíèé ïðè âçÿòèè ðàâíûì àïðèîðíûõ âåðîÿòíîñòåé êîìïîíåíò, ôèêñèðîâàíèè äèàãîíàëüíîãî âèäà ìàòðèö êîâàðèàöèè : è ñòðåìëåíèè åå êîìïîíåíò ê áåñêîíå÷íîñòè. Ãèïåðïàðàìåòðàìè àëãîðèòìà ÿâëÿþòñÿ:

ˆ K æåëàåìîå êîëè÷åñòâî ìîäåëèðóåìûõ êëàñòåðîâ;

ˆ f`1–•••–:g ìíîæåñòâî öåíòðîèäîâ, èç êîòîðûõ ñòàðòóþò èòåðàöèîííûå ïðîöåññû ïîèñêà öåíòðîâ êëàñòåðîâ;

Èñïîëüçóÿ ââåäåííóþ äëÿ ñìåñè íîðìàëüíûõ ðàñïðåäåëåíèé ôîðìóëàìè:

$$I_8 = \arg\min_{:} jjG_8 \quad `_: jj_2^2 - \quad 8 = \overline{1-\#}$$

$$`_: = \frac{1}{j(:j} \tilde{O}_{8 2(:} G_8 - (: = f8 = \overline{1-\#} j I_8 = : g - \quad : 2\, f1 - \bullet\bullet\bullet - g$$

Íàáîð ðàññìàòðèâàåìûõ ãèïåðïàðàìåòðîâ:

ˆ K : »1–minf10–j- jg¼

ˆ f`1–•••–:g: èíèöèàëèçàöèÿ K-ñðåäíèõ++ [2];

Ïîñëåäíèé ðàññìàòðèâàåìûé àëãîðèòì áàéåñîâñêàÿ íîðìàëüíàÿ ñìåñü [18]. Îñíîâíîé èäåå ëþáîé áàéåñîâñêîé ìîäåëè ÿâëÿåòñÿ ââåäåíèå àïðèîðíîãî ðàñïðåäåëåíèÿ íà ïàðàìåòðû ìîäåëè. Ââåäåì îáîçíà÷åíèå = f 1–•••– g, ãäå : = :1. Îáðàòíàÿ ìàòðèöà ñóùåñòâóåò â ñèëó íåâûðîæäåííîñòè ìàòðèöû êîâàðèàöèè íîðìàëüíîãî ðàñïðåäåëåíèÿ. Áóäåì èñïîëüçîâàòü ïàðàìåòðû âìåñòî ïàðàìåòðîâ , ââåä ïî ïåðâûì ïàðàìåòðàì âçàèìíî-îäíîçíà÷íî ïîëó÷àþòñÿ âòîðûå.

Ïåðâ òåì, êàê ñòðîèòü âåðîÿòíîñòíóþ ìîäåëü áàéåñîâñêîé íîðìàëüíîé ñìåñè, ââåäåì íåñêîëüêî ðàñïðåäåëåíèé. Ââåäåì ðàñïðåäåëåíèå íàä ñèììåòðè÷íûìè ïîëîæèòåëüíî ïîëóíåïðÿìûìè ìàòðèöàìè.

Ïóñòü  2 R3 3– =  ) < 0, ,  = , ) < 0, a  3  1. Òîãäà ðàñïðåäåëåíèå

$$?^1\ \ j\ `-,\ ^o = \ ^1,-a^o{}^1\det\ ^{o\frac{a\ 3\ 1}{2}}\exp^1\ \frac{1}{2}atr,\ \ ^1o-$$

ãäå ¹,– a º  íîðìèðîâî÷íàÿ êîíñòàíòà ðàñïðåäåëåíèÿ, áóäåì íàçûâàòü ðàñïðåäåëåíèåì Óèøàðòà.

Ïóñòü ` 2 R3, < 2 R3, V ¡ 0. Òîãäà ðàñïðåäåëåíèå

$$?^1`-\_^o = NW^1\ `-\_j<-V-a-^o; = N^1`\ j<-^1V\ ^o\ ^1oW^1\ \ ja-,^o$$

áóäåì íàçûâàòü Óèøàðò-íîðìàëüíûì ðàñïðåäåëåíèåì.

Ïóñòü c 2 R – c8  0– Í 8c8 = 1, U 2 R – U8 ¡ 0. Òîãäà ðàñïðåäåëåíèå

$$?^1c\ j\ U^o = Dir^1c\ j\ U^o = \frac{^1{}^{Í}_{8=1}U_8^{oo}Ö}{Î\ \ ^1U_8^o\ \ _{8=1}}\ c_8^{U_8\ 1}$$

íàçûâàåòñÿ ðàñïðåäåëåíèåì Äèðèõëå.

Ñ ó÷åòîì ââåäåíèÿ àïðèîðíûõ ðàñïðåäåëåíèé íà ïàðàìåòðû íîðìàëüíîé ñìåñè f c: – `: – : g: =1 âåðîÿòíîñòíóþ ìîäåëü ìîæíî çàïèñàòü ñëåäóþùèì îáðàçîì:

$$?^1-–/–c–`-^o = \ ^{Ö^{\#}_{8=1}}?^1G_8j\ l_8-`-^o?^1l_8j\ c^o?^1`-\ ^o?^1c^o$$

$$?^1G_8j\ l_8-`-^o = N^1`\ l_8-\ l_8^{1o}$$

$$?^1l_8j\ c^o = Cat^1l_8j\ c^o$$

$$?^1`\ l_8-\ l_8^o = NW^1\ `\ l_8-\ l_8j<_0-V_0-a_0-,_0^o$$

Ðàññìàòðèâàþòñÿ 2 òèïà áàéåñîâñêèé íîðìàëüíûõ ñìåñåé, îò ïîñòàíîâêè çàâèñèò çàäàíèå àïðèîðíîãî ðàñïðåäåëåíèÿ íà  c:

^ êîíå÷íàÿ:  ?¹cjU0º = Dir¹c j f U1–•••–Ukº;

^ áåñêîíå÷íàÿ:  ?¹cjU0º = Dir¹c j f U–•••Uº, ãäå  ! 1  ;

Äëÿ êîíå÷íîé ñìåñè ïðèìåíÿþòñÿ ñõåìà Ãèááñà äëÿ ãåíåðàöèè ñëó÷àéíîãî âåêòîðà èç ðàñïðåäåëåíèÿ Äèðèõëå [14], äëÿ áåñêîíå÷íîé ñìåñè  ñêîëëàïñèðîâàííàÿ ñõåìà Ãèááñà [17].

Â äàííîé ðàáîòå áóäåò èñïîëüçîâàòüñÿ ðåëèçàöèÿ áåñêîíå÷íîé ñìåñè íîðìàëüíûõ ðàñïðåäåëåíèé ÷åðåç àïïðîêñèìàöèþ ñõåìû ãåíåðàöèè ñ ïîìîùüþ ïðîöåññà ¾ëîìêè ïàëêè¿ (Stick-breaking) [3]. Êëþ÷åâûì ãèïåðïàðàìåòðîì àëãîðèòìà âûñòóïàåò ïàðàìåòð $U_0$ ðàñïðåäåëåíèÿ $?^1 c\ j U_0 º$, ñîñòîÿùèì èç K îäíîâêîâûõ ïîëîæèòåëüíûõ âåëè÷èí. Çàäàíèå íèçêîãî çíà÷åíèÿ ïðèâîäèò ê êîíöåíòðèðîâàíèþ âåðîÿòíîñòíîé ìàññû â íåáîëüøîé ÷èñëå êîìïîíåíò ñìåñè è ðàçðåæèâàíèþ ìîäåëè ÷åðåç ïðèáëèæåíèå ÷àñòè âåñîâ êîìïîíåíò ê 0. Óâåëè÷åíèå çíà÷åíèÿ ãèïåðïàðàìåòðà ïîçâîëÿåò îñòàâèòü â ìîäåëè áîëüøåå ÷èñëî êîìïîíåíò.

Äëÿ èíèöèàëèçàöèè ïàðàìåòðîâ áóäåì èñïîëüçîâàòü ñëåäóþùèå âåëè÷èíû:

- $< _0$: ñðåäíèé âåêòîð öåíòðîâ êëàñòåðîâ, ïîëó÷åííûõ ñ ïîìîùüþ àëãîðèòìà Ê-ñðåäíèõ;
- $V_0$: 1;
- $a_0$: ;
- , $_0$: ìàòðèöà êîâàðèàöèè äëÿ ìàòðèöû îáúåêòîâ-ïðèçíàêîâ âûáîðêè;

Íàáîð ðàñ÷èòàòâàåìûõ ãèïåðïàðàìåòðîâ:

- : »1–minf 25–j - jg¼4
- êîìïîíåíòû âåêòîðà $U_0$: » 6– 1¼4î ëîãàðèôì÷åñêîé ñòåïåíè ïî îñíîâàíèþ 4;

### 2.1.3 Âûâîäû

Àëãîðèòì ïîñòðîåíèÿ áàéåñîâñêîé íîðìàëüíîé ñìåñè ðàáîòàåò äîëüøå îñòàëüíûõ, íî ïîçâîëÿåò ìîäåëèðîâàòü ìåíüøåå èçíà÷àëüíî çàäàííîãî ÷èñëî êîìïîíåíò ñìåñè çà ñ÷åò çàíóëåíèÿ âåðîÿòíîñòåé ñîîòâåòñòâóþùèõ âåñîâ êîìïîíåíò. Íî âñå åùå òðåáóåò òîíêîãî ïîäáîðà ïàðàìåòðà ðàñïðåäåëåíèÿ Äèðèõëå.

Êàæäûé àëãîðèòì êëàñòåðèçàöèè òðåáóåò íàñòðîéêè íåêîòîðîãî êîëè÷åñòâà ãèïåðïàðàìåòðîâ íåîáõîäèìî ñïîíëüçîâàòü êàêîé-òî êðèòåðèé îòáîðà ìîäåëè äëÿ èõ ïîäáîðà. Â êà÷åñòâå òàêîâîãî áóäåì èñïîëüçîâàòü êîýôôèöèåíò ñèëóýòà [21]. Ðàñ-ñìîòðèì âûáîðêó - ìîùíîñòè # , ñîñòîÿùóþ èç îáúåêòîâ G2 R . Êàæäîìó îáúåêòó ïðèïèñàíà ìåòêà êëàñòåðà H2 1– , ïóñòü = f $_1$–•••– g ìíîæåñòâî êëàñòåðîâ,

îáðàçóþùèõ ðàçáèåíèå âûáîðêè. Çàôèêñèðóåì íåêîòîðóþ ìåòðèêó $d^1 - ^o$. Äëÿ êàæäîãî îáúåêòà $G_8 2 - $, äëÿ êîòîðîãî $j\ _{H_8} j\ ¡\ 1$, îïðåäåëèì:

- $\hat{}\ 0_8 = \frac{1}{j\ _{H_8} j\ 1} \sum_{G_9 2\ H_8 - G_9 < G_9} d^1 G_8 - G^o$ ñðåäíåå ðàññòîÿíèå îò èñõîäíîé òî÷êè äî âñåõ òî÷åê òîãî æå êëàñòåðà;
- $\hat{}\ 1_8 = \min_{H < H_8} \frac{1}{j\ _H j} \sum_{G 2\ H} d^1 G_8 - G$ íàèìåíüøåå ñðåäíåå ðàññòîÿíèå îò èñõîäíîé òî÷êè äî âñåõ òî÷åê äðóãîãî êëàñòåðà;

Êîýôôèöèåíò ñèëóýòà îáúåêòà $G_8$ âûáîðêè ïðåäëîæèì êàê

$$B_8 = 1 \text{»} j\ _{H_8} j\ ¡\ 1¼ \frac{1_8\ \ 0_8}{\max f 0_8 - 1_8 g}$$

Ýòà âåëè÷èíà íàõîäèòñÿ â îòðåçêå » 1–1¼è ïîêàçûâàåò, íàñêîëüêî õîðîøî óäàëîñü êëàñòåðèçîâàòü êîíêðåòíûé îáúåêò âûáîðêè. Áîëåå êà÷åñòâåííàÿ êëàñòåðèçàöèÿ èìååò áîëüøåå çíà÷åíèå êîýôôèöèåíòà ñèëóýòà, ïðè ýòîì íóëåâîå çíà÷åíèå âûñòóïàåò èíäèêàòîðîì íàñëîåíèÿ êëàñòåðîâ, à îòðèöàòåëüíûå çíà÷åíèÿ èíäèêàòîðîì òîãî, ÷òî áîëüøèíñòâî îáúåêòîâ âûáîðêè èìåþò íåâåðíóþ êëàñòåðèçàöèþ, ò.å. äëÿ áîëüøèíñòâà îáúåêòîâ êàêîé-òî äðóãîé êëàñòåð èç èìåþùèõñÿ áûë áû âûáîðêè ëó÷øå, ÷òî åùå îæàò ãîâîðèòü î ñëèøêîì áîëüøîì èëè ñëèøêîì ìàëîì ÷èñëå êëàñòåðîâ.

Êîýôôèöèåíò ñèëóýòà äëÿ êëàñòåðèçàöèè çàäàåòñÿ êàê $B = \frac{1}{\#} \sum_{8=1}^{\#} B_8$.

Èñõîäíûé êîýôôèöèåíò ñèëóýòà íå ïîäõîäèò äëÿ êëàñòåðèçàöèè ñ ïäíèì êëàñòåðîì äîïóñêàåòñÿ åãî äëÿ ýòîãî ñëó÷àÿ íóëåâûì çíà÷åíèåì. Òîãäà èòîãîâûé êîýôôèöèåíò ñèëóýòà ïðèìåò âèä:

$$B = 1 \text{»}\ ¡\ 1¼ \frac{1}{\#} \sum_{8=1}^{\#} B_8$$

Òàêèì îáðàçîì, ëó÷øèì íàáîðîì ãèïåðïàðàìåòðîâ àëãîðèòìà êëàñòåðèçàöèè áóäåì ñ÷èòàòü òàêîé, äëÿ êîòîðîãî êîýôôèöèåíò ñèëóýòà ìàêñèìàëåí.

Ìåòðèêó äëÿ êîýôôèöèåíòà ñèëóýòà áóäåì âûáèðàòü òó æå, êîòîðóþ èñïîëüçóåò àëãîðèòì êëàñòåðèçàöèè, åñëè ìåòðèêà âûñòóïàåò åãî ãèïåðïàðàìåòðîì, èíà÷å èñïîëüçóåì ââåäèòåëü ìåòðèêó.

## 2.2 Âåêòîðíîå ïðåäñòàâëåíèå

Âåêòîðíûå ïðåäñòàâëåíèÿ áóäåì ñòðîèòü íåçàâèñèìî äëÿ êàæäîé èç âûáðîê. Â êà÷åñòâå áàçîâîãî âåêòîðíîãî ïðåäñòàâëåíèÿ äîêóìåíòîâ ïðèìåíÿåòñÿ ñëåäóþùèé íàáîð ïðèçíàêîâ:

- êîëè÷åñòâî ïðèâåäåííûõ ê íà÷àëüíîé ôîðìå ìîðôîëîãè÷åñêèì àíàëèçàòîðîì pymorphy2 [15] èìåíîâàííûõ ñóùíîñòåé, èìåþùèõ ïîëîæèòåëüíàÿ, íåéòðàëüíàÿ è îòðèöàòåëüíàÿ òîíàëüíîñòè â äîêóìåíòå, áåç ó÷åòà ïðÿäêà ñëîâ â ïðåäîáðàáîòàííîé èìåíîâàííîé ñóùíîñòè (ðàçìåðíîñòü âàðüèðóåòñÿ îò 1 äî 3);
- ñóììàðíîå çíà÷åíèå òîíàëüíîñòè äëÿ êàæäîé èìåíîâàííîé ñóùíîñòè, ïðèâåäåííîé ê íà÷àëüíîé ôîðìå ìîðôîëîãè÷åñêèì àíàëèçàòîðîì, êîòîðàÿ âñòðå÷àòñÿ âî âñåõ òåêſòàõ âûáîðêè õîòÿ áû 2 ðàçà áåç ó÷åòà ïîðÿäêà ñëîâ â èìåíîâàííîé ñóùíîñòè (ðàçìåðíîñòü âàðüèðóåòñÿ îò 9 äî 55);
- êîëè÷åñòâî èìåíîâàííûõ ñóùíîñòåé ïîðåäëåííîé êàòåãîðèè (¾ëè÷íîñòü¿, ¾ååîãðàôè÷åñêîå ïîëîæåíèå¿, ¾îðãàíèçàöèÿ¿) (ðàçìåðíîñòü âàðüèðóåòñÿ îò 1 äî 3);
- ãðóïïà ñîöèàëüíî-äåìîãðàôè÷åñêèõ ïðèçíàêîâ èç èñõîäíîãî íàáîðà äîêóìåíòîâ (ðàçìåðíîñòü 15);
- ýìáåääèíã ïðåäîáðàáîòàííîãî îáúåäèíåíèÿ çàãîëîâêà è òåëà äîêóìåíòà, ïîëó÷åííûé ñ ïîìùüþ ïðåäîáó÷åííîãî SBERT [23], ñ îãðàíè÷åíèåì íà êîëè÷åñòâî âõîäíûõ òîêåíîâ 384, â ñëó÷àå åñëè äîêóìåíò èìåò áîëåå êîðîòêîå âíóòðåííåå äëÿ íåéðîííîé ñåòè ïðåäñòàâëåíèå, âõîäíîé âåêòîð äîçàïîëíÿåòñÿ ¾íóëåâûìè¿ òîêåíàìè (ðàçìåðíîñòü 1024);

Ïðèâåäåì ìîòèâàöèþ èñïîëüçîâàíèÿ òàêèõ ïðèçíàêîâ.

1. Ìíåíèå î ñîáûòèè, â êîòîðîì êëþ÷åâûìè ¾ó÷àñòíèêàìè¿ ìîãóò âûñòóïàòü ëè÷íîñòè èëè îðãàíèçàöèè, äåéſòâóþùèå â ïîðåäëåííîì ìåſòå, ôîðìóëèðîâàíî â âèäå òîíàëüíîſòè, ïðèíàäëåæíîé èìåíîâàííîé ſóùíîſòè ſſîòâåòſòâóþùåé êàòåãîðèè.

2. Ìíåíèå, ïðîäâèãàåìîå â ÑÌÈ íîâîñòíûè ðåñóðñàìè, íàïðàâëåíî íà îïðåäåëåííûé ñðåç ëþäåé, ãäå ñðåçàìè ìîãóò áûòü íåðàçîâàíèå, äîñòàòîê èëè ïîë.

3. Ìîäåëü SBERT îáó÷àëàñü ðåøàòü íåñêîëüêî çàäà÷ ïîíèìàíèÿ åñòåñòâåííîãî ÿçûêà îäíîâðåìåííî ( multitask-learning), à èìåííî:

   - çàäà÷à àâòîìàòè÷åñêîãî ïîðåäåëåíèÿ ëîãè÷åñêîé ñâÿçè ìåæäó òåêñòàìè (natural language inference): èìåòñÿ íåêîòîðàÿ òåêñòîâàÿ ïîñûëêà, îïåðóþùàÿ ñèòóàöèþ, íåêîòîðîå òåêñòîâîå ïðåäïîëîæåíèå î ñèòóàöèè, çàäà÷à ñîñòîèò â êëàññèôèêàöèè ïðåäïîëîæåíèÿ íà òðè êëàññà: ¾ïðàâäà¿, ¾ëîæü¿, ¾îïðåäåëèòü íåâîçìîæíî¿;
   - çàäà÷à âûäåëåíèÿ èìåíîâàííûõ ñóùíîñòåé;
   - çàäà÷à àíàëèçà òîíàëüíîñòè èìåíîâàííûõ ñóùíîñòåé;

   Ýòè çàäà÷è áëèçêè ê ðåøàåìîé â ýòîé ðàáîòå çàäà÷å ñ òî÷êè çðåíèÿ âûâèãàåìîé ìîòèâàöèè èñïîëüçîâàíèÿ ïðèçíàêîâ, ïîýòîìó åñòü ïîâîä ïîëàãàòü, ÷òî ïîâûøåíèå êà÷åñòâà ìîäåëü âêëþ÷åíèå ïðîñòðàíñòâî ýìáåääèíãîâ ïîçâîëèò ïîñòðîèòü êëàññíûé, îòíîñèòåëüíî íåáîëüøîé, ïî ðàçìåðó íåéðîííîé ñåòè ìîäåëü.

Ïðîèçâåäåì ñòàíäàðòèçàöèþ ìàòðèöû îáúåêòîâ-ïðèçíàêîâ: äëÿ êàæäîãî ïðèçíàêà íåçàâèñèìî âû÷òåì åãî ñðåäíåå àðèôìåòè÷åñêîå è ïîäåëèì ïîñëå ýòîãî íà ñòàíäàðòíîå îòêëîíåíèå.

## 2.3 Ñíèæåíèå ðàçìåðíîñòè

Èòîãîâàÿ ðàçìåðíîñòü âåêòîðà äîñòàòî÷íî âåëèêà, ïîýòîìó íåîáõîäèìî ïðèáåãíóòü ê ìåòîäàì ñíèæåíèÿ ðàçìåðíîñòè. Âîñïîëüçóåìñÿ ìåòîäîì ãëàâíûõ êîìïîíåíò (PCA) äëÿ ëèíåéíîãî ñíèæåíèÿ ðàçìåðíîñòè èñõîäíîãî âåêòîðíîãî ïðîñòðàíñòâà [19]. Ñóòü ìåòîäà ñîñòîèò â ïîèñêå òàêîãî ïðîñòðàíñòâà ìåíüøåé ðàçìåðíîñòè, êîòîðîå ìèíèìèçèðóåò ïîòåðþ èíôîðìàöèè î âûáîðêå ñ òî÷êè çðåíèÿ äèñïåðñèè. ×èñëî êîìïîíåíò áóäåì ïîäáèðàòü òàêîå, ÷òîáû ïîëó÷åííûå ïðèçíàêè ñîõðàíèëè 99% èñõîäíîé äèñïåðñèè. Ïóñòü $_1 \dots _$ ìîæåñòâî ñîáſòâåííûõ çíà÷åíèé ýìïèðè÷åñêîé êîâàðèàöèîííîé ìàòðèöû îáúåêòîâ-ïðèçíàêîâ. Îíè âñå íåîòðèöàòåëüíûå â ñëåó íåîòðè-

öàòåëüíîé îïðåäåëåííîñòè êîâàðèàöèîííîé ìàòðèöû. Òîãäà êðèòåðèé îòáîðà ÷èñëà êîìïîíåíò ñâîäèòñÿ ê ïîèñêó $\min_{:=f1-\bullet\bullet\bullet-g} : : \frac{\sum_{8=1}^{:} -_8}{\sum_{8=1}^{\ell} -_8} ¡ 0\bullet 99.$

Ðàññìîòðèì òàêæå ÿäðîâîé PCA ( kernel PCA, kPCA) äëÿ íåëèíåéíîãî ñíèæåíèÿ ðàçìåðíîñòè [24]. Â êà÷åñòâå ÿäðîâîé ôóíêöèè áóäåì èñïîëüçîâàòü êîñèíóñíóþ áëèçîñòü: $^1G-\P\vdash = \frac{G\,H}{jjGjj\,jjHjj}$. Îïòèìàëüíóþ ðàçìåðíîñòü ïðîñòðàíñòâà, êîòîðîå ñòðîèò ìîäåëü ÿäðîâîãî ìåòîäà ãëàâíûõ êîìïîíåíò, áóäåì èñêàòü â îòðåçêå »10–30½è ïîäáèðàòü ñîâìåñòíî ñ ïàðàìåòðàìè àëãîðèòìîâ êëàñòåðèçàöèè ïî êîýôôèöèåíòó ñèëóýòà.

## 2.4 Ðàçìåòêà äàííûõ

Ðàçìåòêà ïðîèçâîäèëàñü ïðè ïîìîùè èíñòðóìåíòà ßíäåêñ.Òîëîêà ïî èíñòðóêöèè, ïîëó÷åííîé â Ïðèëîæåíèè. Êàæäîìó äîêóìåíòó â âûáîðêå ïðîñòàâëÿëàñü ìåòêà ïîëÿðíà ïîëÿðèçàöèè pole_8– 8= 0–••, åñëè äîêóìåíò â ðàìêàõ îïèñûâàåìîãî â òåìå ñîáûòèÿ ÿâëÿëñÿ ïîëÿðèçîâàííûì, without_polarization, åñëè íå ÿâëÿëñÿ òàêîâûì, è other_topic, åñëè äîêóìåíò íå îòíîñèëñÿ ê ñîáûòèþ òåìû.

Ðèñ. 3: Èñõîäíàÿ ðàçìåòêè

Âèäèì, ÷òî îñíîâíàÿ ÷àñòü òåì ñîäåðæàò áîëüøå äâóõ ìíåíèé, ïðè ýòîì åñòü 4 òåìû, â êîòîðûõ ïðèñóòñòâóþò áîëüøå 5 ìíåíèé. Ñ ðîñòîì ÷èñëà ìíåíèé ðàñòåò è

êîëè÷åñòâî äîêóìåíòîâ â òåìàõ. Ñðåäíåå êîëè÷åñòâî øóìîâûõ äîêóìåíòîâ óâûâàåò ñ ðîñòîì ÷èñëà ìíåíèé â òåìàõ, ïðè ýòîì äëÿ òåì ñ îäíèì ìíåíèåì øóìîâûõ äîêóìåíòîâ áîëüøå, ÷åì ïîëÿðèçîâàííûõ.

# 3 Âû÷èñëèòåëüíûé ýêñïåðèìåíò

Äëÿ ýêñïåðèìåíòîâ èñïîëüçîâàëèñü ðåàëèçàöèè àëãîðèòìîâ êëàñòåðèçàöèè èç áèáëèîòåêè scikit learn [20].

Â êà÷åñòâå äîïîëíèòåëüíîé èíôîðìàöèè î êà÷åñòâå êëàñòåðèçàöèè áóäåì èñïîëüçîâàòü ìåòðèêó "8BB >D=C ðàçíîñòü ìåæäó êîëè÷åñòâîì êëàññîâ è êîëè÷åñòâîì êëàñòåðîâ, ïîñòðîåííûõ àëãîðèòìîì, óñðåäíåííàÿ ïî âñåì òåñòàì.

## 3.1 Ðåçóëüòàòû ýêñïåðèìåíòà

Ïðèâåäåì ðåçóëüòàòû â âèäå ñâîäíûõ òàáëèö ïî âñåì òåñòàì è ïî êàæäîìó èç îáîçíà÷åííûõ â ïîñòàíîâêå çàäà÷è ñðåçîâ.

| Ìîäåëü | | Ìåòðèêà êà÷åñòâà | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| DBSCAN | PCA | 0.565 | 0.919 | 0.674 | 0.428 | 1.69 |
| | kPCA | 0.584 | 0.869 | 0.668 | 0.428 | 1.36 |
| Ñðåäíåãî ñâÿçè | PCA | 0.959 | 0.267 | 0.397 | 0.611 | -9.86 |
| | kPCA | 0.897 | 0.371 | 0.479 | 0.670 | -6.20 |
| Íîðìàëüíàÿ ñìåñü | PCA | 0.637 | 0.572 | 0.563 | 0.588 | 0.33 |
| | kPCA | 0.692 | 0.553 | 0.564 | 0.588 | -0.38 |
| K-ñðåäíèõ | PCA | 0.679 | 0.774 | 0.681 | 0.646 | 0.36 |
| | kPCA | 0.764 | 0.510 | 0.569 | 0.600 | -1.73 |
| Áàéåñîâñêàÿ íîðìàëüíàÿ ñìåñü | PCA | 0.667 | 0.825 | 0.698 | 0.625 | 0.58 |
| | kPCA | 0.768 | 0.513 | 0.571 | 0.626 | -1.87 |

Òàáëèöà 1: Ðåçóëüòàòû ïî âñåì òåñòàì

| Модель | | Метрика качества | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| DBSCAN | PCA | 0.898 | **0.929** | **0.896** | **0.668** | -0.07 |
| | kPCA | 0.918 | 0.856 | 0.866 | **0.668** | -0.27 |
| Среднего сдвига | PCA | **1.000** | 0.150 | 0.244 | 0.359 | -11.13 |
| | kPCA | 0.992 | 0.221 | 0.347 | 0.416 | -6.93 |
| Нормальная смесь | PCA | 0.933 | 0.506 | 0.641 | 0.417 | -1.13 |
| | kPCA | 0.947 | 0.448 | 0.577 | 0.393 | -2.07 |
| К-средних | PCA | 0.966 | 0.667 | 0.765 | 0.428 | -1.40 |
| | kPCA | 0.984 | 0.429 | 0.569 | 0.435 | -2.60 |
| Байесовская нормальная смесь | PCA | 0.966 | 0.748 | 0.825 | 0.412 | -1.27 |
| | kPCA | 0.992 | 0.446 | 0.582 | 0.405 | -2.67 |

Таблица 2: Результаты по темам с 1 мнением

| Модель | | Метрика качества | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| DBSCAN | PCA | 0.574 | **0.936** | 0.700 | 0.377 | 1.15 |
| | kPCA | 0.583 | 0.908 | 0.693 | 0.377 | 0.96 |
| Среднего сдвига | PCA | **0.932** | 0.292 | 0.417 | 0.716 | -7.26 |
| | kPCA | 0.926 | 0.305 | 0.446 | 0.738 | -6.44 |
| Нормальная смесь | PCA | 0.671 | 0.584 | 0.601 | 0.695 | -0.33 |
| | kPCA | 0.733 | 0.547 | 0.596 | 0.728 | -0.93 |
| К-средних | PCA | 0.733 | 0.769 | 0.724 | **0.739** | -0.44 |
| | kPCA | 0.761 | 0.569 | 0.599 | 0.710 | -1.52 |
| Байесовская нормальная смесь | PCA | 0.722 | 0.808 | **0.739** | 0.732 | -0.26 |
| | kPCA | 0.766 | 0.565 | 0.595 | 0.655 | -1.89 |

Таблица 3: Результаты по темам с 2 мнениями

| Модель | | Метрика качества | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| DBSCAN | PCA | 0.493 | **0.892** | 0.627 | 0.445 | 1.76 |
| | kPCA | 0.497 | 0.865 | 0.621 | 0.445 | 1.69 |
| Среднего сдвига | PCA | **0.971** | 0.288 | 0.432 | 0.562 | -9.56 |
| | kPCA | 0.870 | 0.444 | 0.547 | **0.651** | -4.76 |
| Нормальная смесь | PCA | 0.583 | 0.571 | 0.547 | 0.567 | 0.14 |
| | kPCA | 0.651 | 0.573 | 0.575 | 0.539 | -0.34 |
| К-средних | PCA | 0.611 | 0.769 | 0.653 | 0.585 | 0.41 |
| | kPCA | 0.732 | 0.513 | 0.571 | 0.580 | -1.72 |
| Байесовская нормальная смесь | PCA | 0.613 | 0.810 | **0.671** | 0.564 | 0.52 |
| | kPCA | 0.728 | 0.513 | 0.572 | 0.595 | -1.66 |

Таблица 4: Результаты по темам с 3 мнениями

| Модель | | Метрика качества | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| DBSCAN | PCA | 0.401 | 0.930 | 0.539 | 0.352 | 3.74 |
| | kPCA | 0.460 | 0.834 | 0.551 | 0.352 | 2.68 |
| Среднего сдвига | PCA | **0.947** | 0.289 | 0.434 | 0.652 | -13.00 |
| | kPCA | 0.821 | 0.472 | 0.530 | **0.727** | -7.47 |
| Нормальная смесь | PCA | 0.434 | 0.607 | 0.476 | 0.535 | 2.74 |
| | kPCA | 0.492 | 0.616 | 0.495 | 0.536 | 1.68 |
| К-средних | PCA | 0.479 | 0.872 | **0.595** | 0.711 | 2.79 |
| | kPCA | 0.648 | 0.488 | 0.524 | 0.533 | -1.37 |
| Байесовская нормальная смесь | PCA | 0.434 | **0.936** | 0.581 | 0.660 | 3.32 |
| | kPCA | 0.655 | 0.491 | 0.526 | 0.659 | -1.53 |

Таблица 5: Результаты по темам с 4 мнениями

## 3.2 Выводы

Наиболее точным оказывается метод среднего сдвига с линейным снижением размерности, при этом он является худшим по метрике полноты, которую удается улучшить при использовании нелинейного снижения размерности.

Лучшим по *BCubed-полноте* оказывается DBSCAN.

По основной метрике *BCubed-F-мера* лучшим алгоритмом кластеризации является байесовская нормальная смесь.

Лучшим из рассмотренных способов снижения размерности почти всегда оказывается линейный метод главных компонент с отбором числа компонент по сохраненной дисперсии, но лучшим алгоритмом для построения шумового кластера оказывается метод среднего сдвига с нелинейным снижением размерности.

Стоит отметить, что более простой алгоритм K-средних показал наиболее близкое качество по метрикам кластеризации и выявления шума, как и байесовская нормальная смесь, при этом обучение модели K-средних проходит гораздо быстрее из-за меньшего количества параметров.

К применению в рамках задачи предлагается использование алгоритма кластеризации байесовской нормальной смесью.

# 4 Исследование абляции

Проведем дополнительный эксперимент, связанный с изучением вклада групп признаков в модель:

- признаки модели SBERT (*SBERT*);
- социально-демографические признаки (*соц.-дем.*);
- признаки, связанные с именованными сущностями и их тональностями (*NER*);

Так как при использовании метода главных компонент теряется интерпретируемость признаков, будем применять ее только для векторов SBERT, интерпретируемость которых и так отсутствует. В качестве алгоритмов кластеризации оставим только байесовскую нормальную смесь и K-средних как наилучшие по основной

метрике кластеризации, а в качестве метода снижения размерности оставим только линейный метод главных компонент с отбором числа компонент по сохраненной дисперсии. Способ подбора гиперпараметров и метрики качества будем использовать те же, рассматривать будем все темы совместно без срезов по числу мнений.

В качестве *базового* результата будем использовать кластеризацию обозначенным способом для всего векторного пространства, удаление соответствующих групп признаков будем обозначать *SBERT*, *соц.-дем.* и *NER*.

## 4.1 Результаты эксперимента

| Модель | | Метрика качества | | | | |
|---|---|---|---|---|---|---|
| | | % | ' | | #>8B4 | "8BB >D=C |
| Байесовская нормальная смесь | *базовый* | 0.665 | 0.825 | 0.696 | 0.624 | 0.58 |
| | *SBERT* | 0.607 (–0.058) | 0.819 (–0.006) | 0.670 (–0.026) | 0.600 (–0.025) | 0.98 (+0.40) |
| | *соц.-дем.* | 0.666 (+**0.001**) | 0.823 (+**0.004**) | 0.696 (+**0.000**) | 0.625 (+**0.001**) | 0.57 (–0.01) |
| | *NER* | 0.661 (-0.004) | 0.821 (-0.004) | 0.691 (-0.005) | 0.620 (-0.004) | 0.54 (–0.04) |
| К-средних | *базовый* | 0.685 | 0.764 | 0.680 | 0.640 | 0.24 |
| | *SBERT* | 0.642 (-0.043) | 0.648 (-0.116) | 0.609 (-0.071) | 0.589 (-0.051) | 0.13 (–0.11) |
| | *соц.-дем.* | 0.677 (-0.008) | 0.780 (+**0.016**) | 0.686 (+**0.006**) | 0.642 (+**0.002**) | 0.44 (+0.20) |
| | *NER* | 0.674 (-0.011) | 0.771 (+**0.007**) | 0.675 (-0.005) | 0.640 (+**0.000**) | 0.34 (+0.10) |

Таблица 6: Результаты по всем темам

## 4.2 Выводы

Признаки модели SBERT оказались наиболее важными для обоих алгоритмов кластеризации с точки зрения прироста по основной метрике кластеризации. Группа признаков, связанная с именованными сущностями и тональностями, также вносит положительный вклад в модель кластеризации. А вот группа социально-демографических признаков либо не влияет качество модели по *BCubed-F-мере*, как вышло для байесовской нормальной смеси, либо вовсе ухудшает модель, как вышло для алгоритма K-средних.

Заметим также, что кластеризация байесовской нормальной смесью в прошлом эксперименте лучше на 0.002 по *BCubed-F-мере*, чем кластеризация байесовской нормальной смесью в базовом варианте – лучше снижать размерность над всем признаковым пространством, а не только над пространством модели SBERT.

# 5 Заключение

В рамках выпускной квалификационной работы удалось:

1. Реализовать алгоритм кластеризации для имеющихся признаков, позволяющий достигать качества **0.698** по метрике *BCubed-F-мера*;

2. Сравнить модели на способность выявлять шум в темах по предложенной метрике;

3. Предложить методику оценивания моделей кластеризации тем для поляризованных выборок;

# Список литературы


[1] E. Amigó и др. “A comparison of extrinsic clustering evaluation metrics based on formal constraints.” В: (2009).

[2] D. Arthur и S. Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. В: (2007).

[3] D. Blei и M Jordan. “Variational inference for Dirichlet process mixtures”. В: (2006).

[4] R. Cohen и D. Ruths. “Classifying Political Orientation on Twitter: It’s Not Easy!” В: (2013).

[5] D. Comaniciu и P. Meer. “Mean Shift: A robust approach toward feature space analysis”. В: (2002).

[6] Kareem Darwish и др. “News Consumption in Time of Conflict: 2021 Palestinian-Israel War as an Example”. В: (2021).

[7] A.P. Dempster, N.M. Laird и D.B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. В: (1977).

[8] Subhabrata Dutta и др. “Semi-supervised Stance Detection of Tweets Via Distant Network Supervision”. В: (2022).

[9] M. Ester и др. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. В: (1996).

[10] Tiziano Fagni и Stefano Cresci. “Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning”. В: (2022).

[11] T. Fawcett. “An Introduction to ROC Analysis”. В: (2006).

[12] C.J. Filmore. *Some problems for case grammar*. 1971.

[13] Margherita Gambini и др. “Tweets2Stance: Users stance detection exploiting Zero-Shot Learning Algorithms on Tweets”. В: (2022).

[14] S. Geman и D. Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Image”. В: (1984).

[15] M. Korobov. “Morphological Analyzer and Generator for Russian and Ukrainian Languages”. English. В: *Analysis of Images, Social Networks and Texts*. Под ред. M. Khachay и др. Т. 542. Communications in Computer and Information Science. Springer International Publishing, 2015, с. 320—332. ISBN: 978-3-319-26122-5. DOI: 10.1007/978-3-319-26123-2_31. URL: http://dx.doi.org/10.1007/978-3-319-26123-2_31.

[16] Jinning Li и др. “Unsupervised Belief Representation Learning in Polarized Networks with Information-Theoretic Variational Graph Auto-Encoders”. В: (2022).

[17] Jun. S. Liu. “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Proble”. В: (1994).

[18] Jun Lu. “A survey on Bayesian inference for Gaussian mixture model ”. В: (2021).

[19] K. Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. В: (1901).

[20] F. Pedregosa и др. “Scikit-learn: Machine Learning in Python”. В: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.

[21] Peter J. Rousseeuw. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. В: (1987).

[22] Lloyd S. “Least square quantization in PCM’s.” В: (1957).

[23] SberDevices. “BERT large model multitask (cased) for Sentence Embeddings in Russian language”. В: (2021). URL: https://huggingface.co/sberbank-ai/sbert_large_mt_nlu_ru.

[24] B. Schölkopf, A. Smola и K. Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. В: (1998).

[25] К.В. Воронцов. *Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM*. 2021.

[26] Д. Фельдман и К.В. Воронцов. “Комбинирование фактов, семантических ролей и тональных слов в генеративной модели для поиска мнений”. В: (2020).

# 6 Приложение

## 6.1 Инструкция для асессоров для Яндекс.Толока

**Выявление поляризации текстов новостей внутри тем**.

На странице задания представлена группа текстов, которые принадлежат в основном какой-то одной теме. Ваша задача: прочитать все тексты, выделить мнения-полюсы, которые по Вашему мнению присутствуют в данной группе новостей, и определить, к какому из выделенных вами мнений относится каждый текст. Под полюсом тут подразумевается попытка авторов текста отразить некоторую смещенную точку зрения на определенное событие, например:

1. 
   - Полюс 1: «Мнение журналистов BBC»
     - «Журналист BBC заявил, что британский эсминец намеренно нарушил границы РФ в Черном море»;
     - «Корреспондент BBC раскрыл правду об инциденте с «Дефендером» в Черном море»;
   - Полюс 2: «Взгляд Российских властей»
     - «Россия заявила о рисках проведения США и их союзниками учений в Черном море»;
     - «Россия призвала США отказаться от военных маневров в Черном море»;
     - «Сенатор от Крыма оправдала действия российских бомбардировщиков в отношении британского эсминца»;

2. 
   - Полюс 1: «Произвол/негатив»
     - «В вытрезвитель теперь смогут забрать даже из дома, к тому же у клиента будет произведен досмотр вещей»;
   - Полюс 2: «Нейтральное/констатация факта»;
     - «МВД РФ согласовало правила помещения россиян в медицинские вытрезвители»;