Технология извлечения информации из последовательности текстов для определения их стиля

Волков Григорий, 371
Научный руководитель: О. Н. Граничин

Цель работы

1. Построить модель алгоритма, который векторизует текст так, что векторные представления текстов одного автора должны быть похожи
2. Придумать имплементацию этой модели и найти оптимальные параметры
3. Настроить тестирование моделей, описанного типа

Модель алгоритма

Модель состоит из трех больших частей:

1. Text preprocessing
2. Feature vectorization
3. Learning algorithm

Preprocessing

Представлен в виде композиции отображений таких, что в итоге мы получаем список токенов.

Например:

1. Перевод всех букв в нижний регистр
2. Разбиение текста на слова
3. Удаление стоп-слов

Feature vectorization

Преобразование списка текстов в список векторов

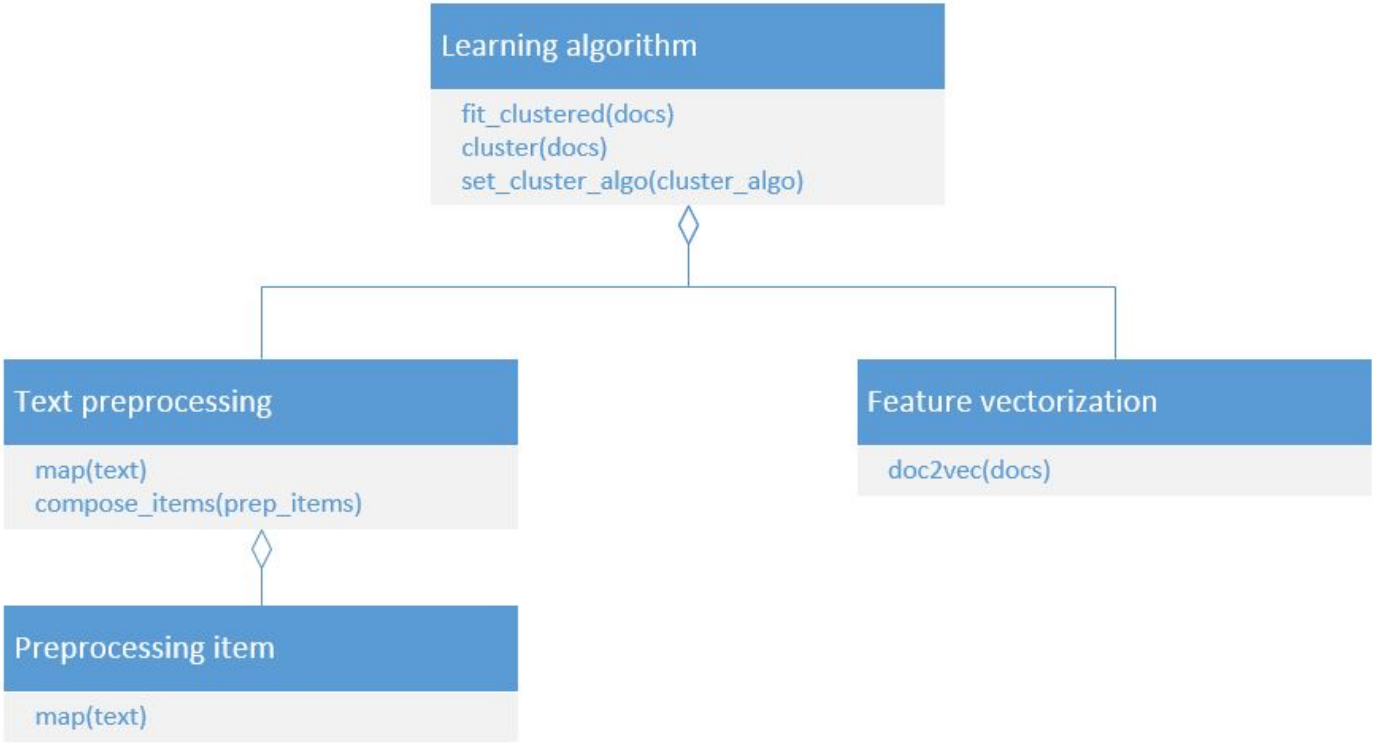
Например, подсчет n-grams (подряд идущих n букв)

Learning algorithm

Некий алгоритм:

1. Оперирует понятиями text preprocessing и feature vectorization
2. Поведение которого зависит от параметров, подбираемых в соответствие с training set

Диаграмма модели



Реализация прототипа модели

Был взят алгоритм из статьи “Patterning of Writing Style Evolution by means of Dynamic Similarity” и переформулирован в описанных ранее терминах.

Описание алгоритма

1. После препроцессинга, токены склеиваются
2. Каждый текст разбивается на чанки длины L
3. Все чанки векторизируются с помощью feature vectorizer
4. На основе расстояний между каким-то чанком i и T его предшественниками и чанком j с T его предшественниками считается элемент матрицы V[i, j]
5. Затем эта матрица интерпретируется как набор векторизованных чанков и передается алгоритму кластеризации
6. Исходный текст относят к тому кластеру, в котором оказалась бóльшая часть его чанков

Параметры

1. Параметры модели, подбираемые в соответствие с обучающей выборкой:
   - L - длина чанков
   - T - количество чанков-предшественников
2. Другие параметры:
   - Метрика расстояния между векторами
   - Конфигурация n-gram
   - Алгоритм кластеризации

Конфигурация n-gram

Назовем конфигурацией n-gram пару (n, amt), где

- n - размер gram
- amt - количество фитч (т.е. учитываются только amt часто встречаемых gram)

Расстояние между векторами

В качестве метрик расстояния были взяты:

1. Canberra
2. Euclidean
3. Cosine
4. Correlation

Алгоритмы кластеризации

Были выбраны следующие алгоритмы кластеризации:

1. K-medoids
2. K-means
3. Mini-Batch k-means
4. Agglomerative clustering
5. Birch

Инструменты

1. Python + numpy, sklearn, scipy, nltk

Выборка

По 5 произведений следующих авторов:

Русскоязычных:

- Шолохов
- Тургенев
- Пушкин
- Серафимович
- Толстой
- Достоевский

Англоязычных:

- Оруэлл
- Кинг
- Толкин
- Остин

Выбор нужной конфигурации алгоритма

Отправная точка - конфигурация из изученной статьи:

1. Алгоритм кластеризации - K-medoids
2. Метрика расстояния - canberra
3. Конфигурация n-gram - (2, 200)

Исходя из этой конфигурации были подобраны:

1. T = 15
2. L = 1000

Лучшие конфигурации

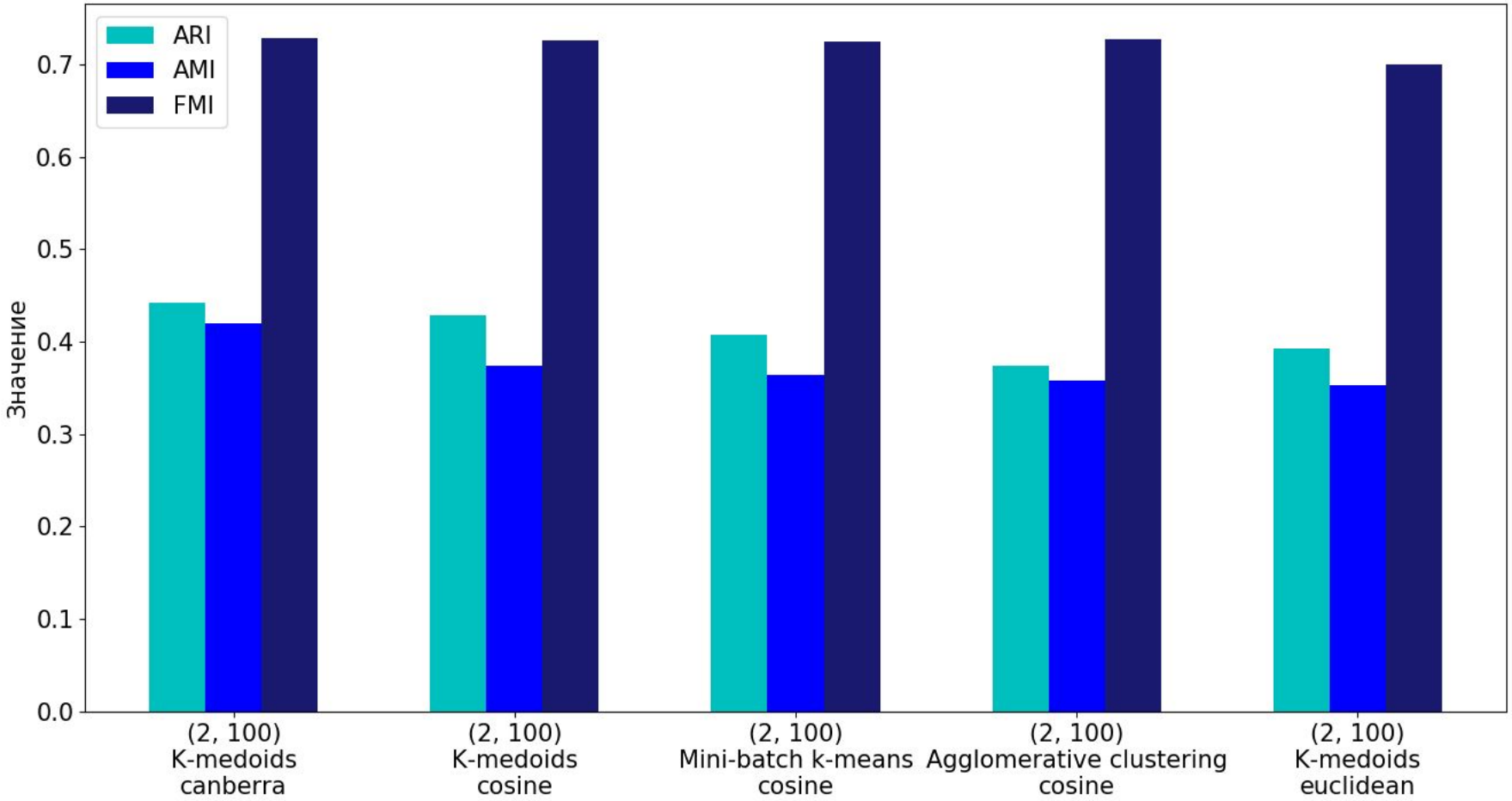
Исходя из оценок “правильности” кластеризации были выбраны следующие тройки параметров:

	n-grams	Алгоритм кластеризации	Метрика расстояния
1	(2, 100)	K-medoids	canberra
2	(2, 100)	K-medoids	cosine
3	(2, 100)	Mini-batch k-means	cosine
4	(2, 100)	Agglomerative clustering	cosine
5	(2, 100)	K-medoids	euclidean

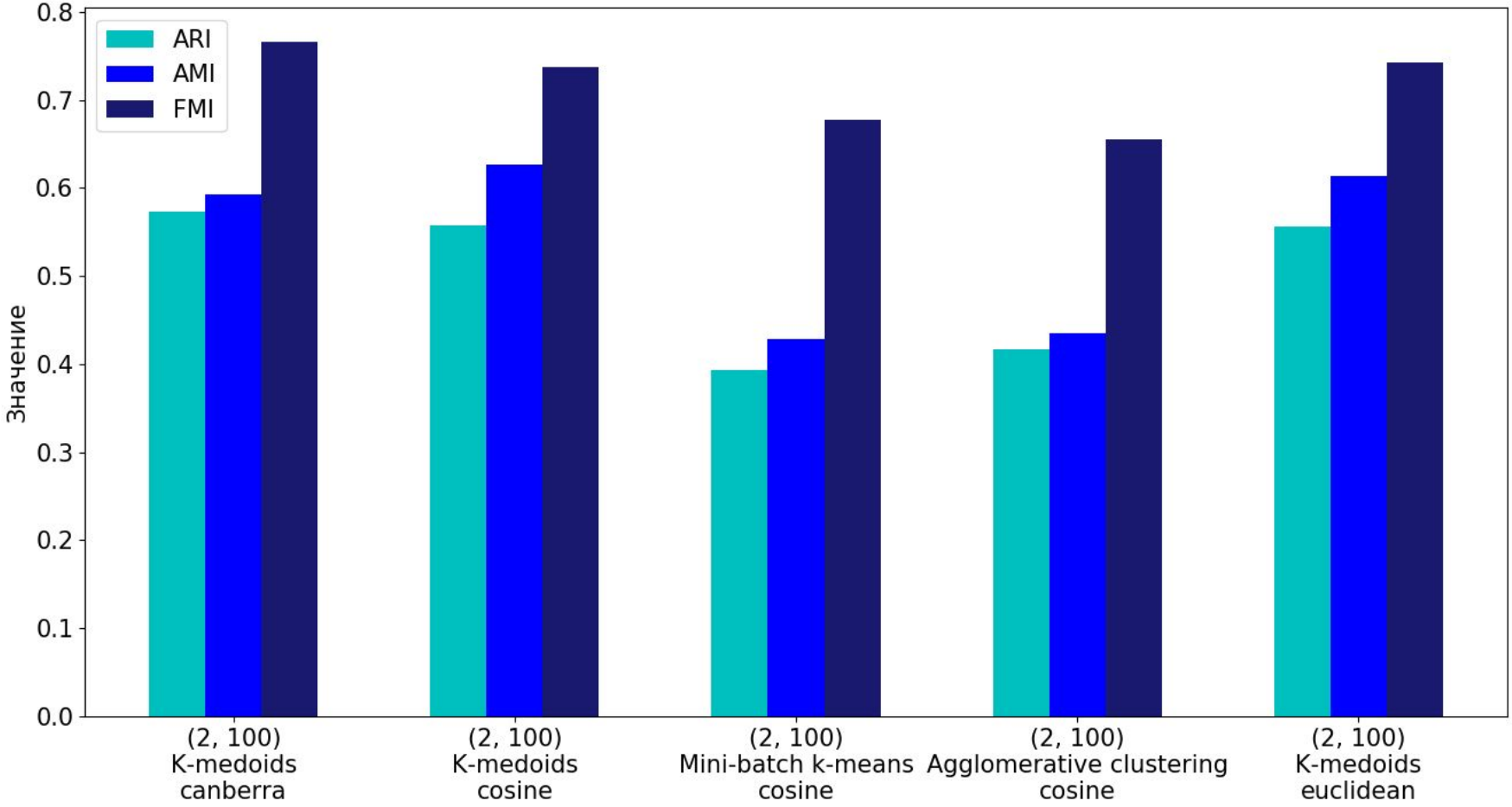
Оценка кластеризации

Для оценки были выбраны следующие метрики:

1. Adjusted rand index
2. Adjusted mutual information
3. Fowlkes–Mallows index



Оценка качества кластеризации чанков



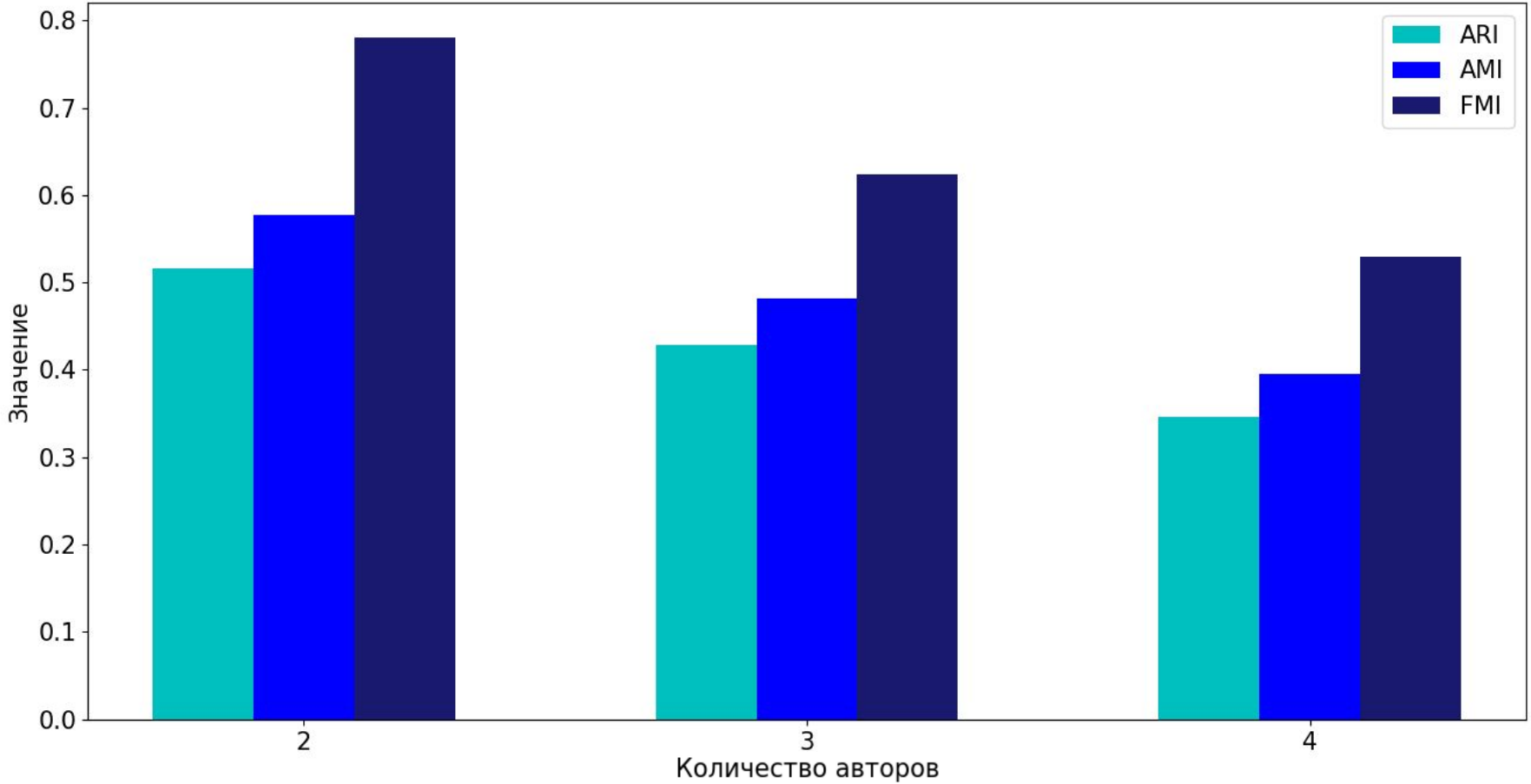
Оценка качества кластеризации текстов

Identity POS-граммы

Препроцессинг: разбиение на слова, удаление знаков препинания

Векторизация: n-граммы, где граммой является часть речи

Алгоритм: применения алгоритма кластеризации



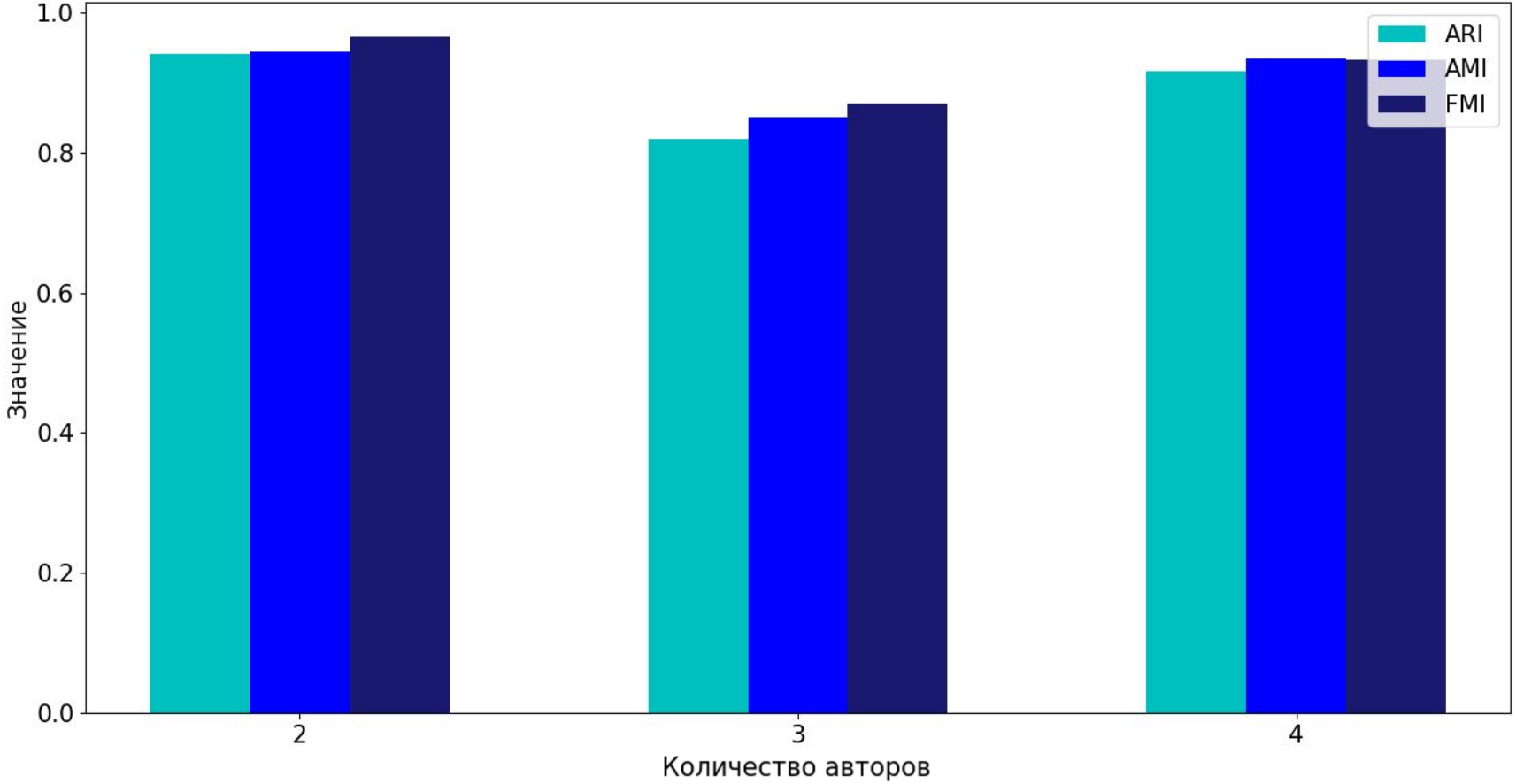
Оценка качества кластеризации текстов

Identity биграммы

Препроцессинг: разбиение на слова, удаление знаков препинания

Векторизация: буквенные n-граммы

Алгоритм: применения алгоритма кластеризации

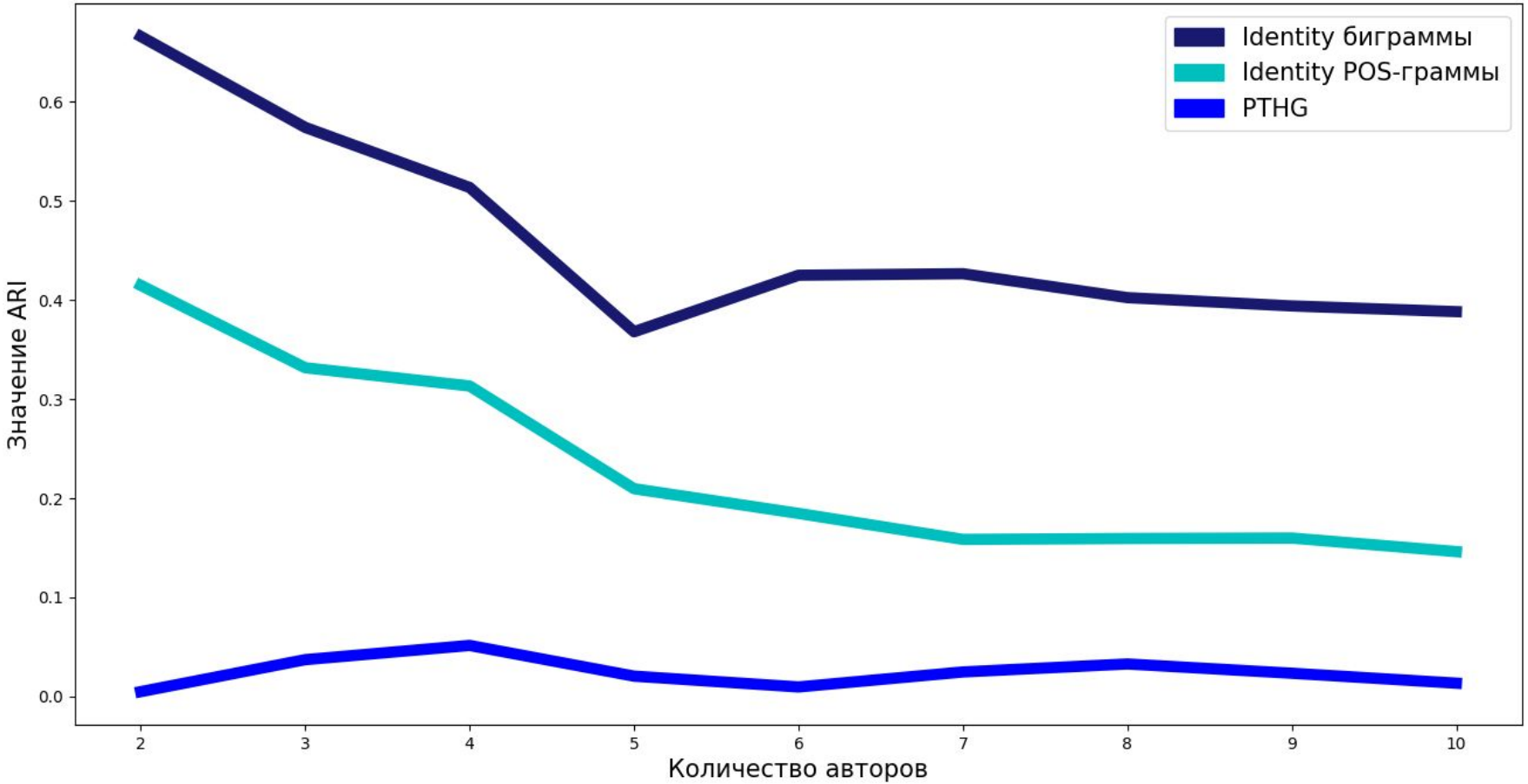


Оценка качества кластеризации текстов

Статьи новостного агентства “Рейтер”

Далее было проведено сравнение алгоритмов на выборке статей новостного агентства “Рейтер”

Выборка представляет из себя 2500 статей (50 авторов по 50 статей)



Оценка качества кластеризации текстов для различного количества кластеров