Рекомендательная система новостей на основе нечеткой кластеризации

Студент: Чалый Андрей Александрович

Группа: ИУ7-82Б

Руководитель: Русакова Зинаида Николаевна

Москва, 2022

Цель и задачи работы

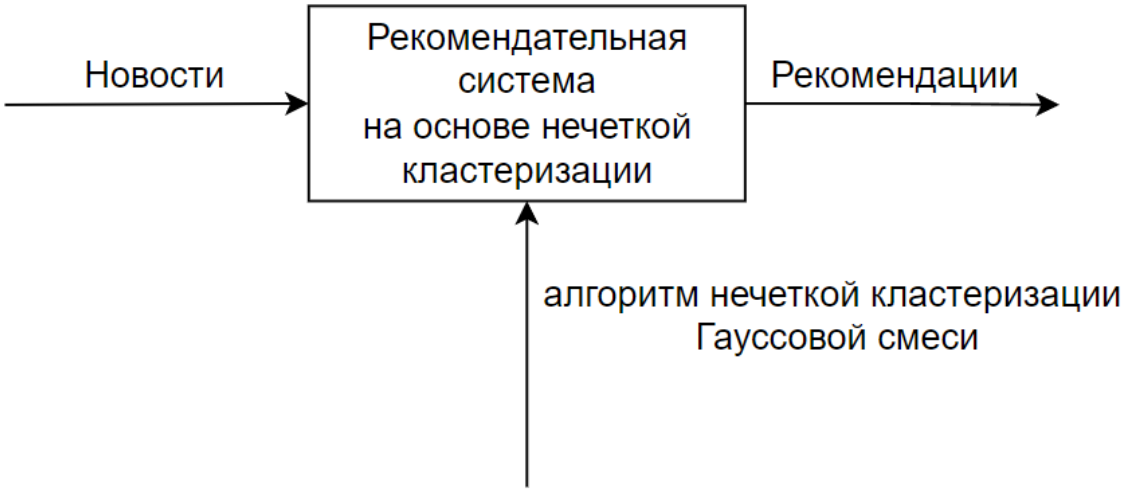
Цель работы — разработка и программная реализация рекомендательной системы новостей на основе нечеткой кластеризации.

Для достижения поставленной цели необходимо решить следующие задачи:

- провести анализ предметной области;
- проанализировать подходы к реализации рекомендательных систем ;
- в результате полученных во время анализа данных разработать рекомендательную систему на основе нечеткой кластеризации;
- разработать программное обеспечение для рекомендательной системы на основе нечеткой кластеризации;
- провести исследование работоспособности реализованной рекомендательной системы и алгоритма нечеткой кластеризации.

Постановка задачи

Ограничение на вход: новость должна быть на английском языке.



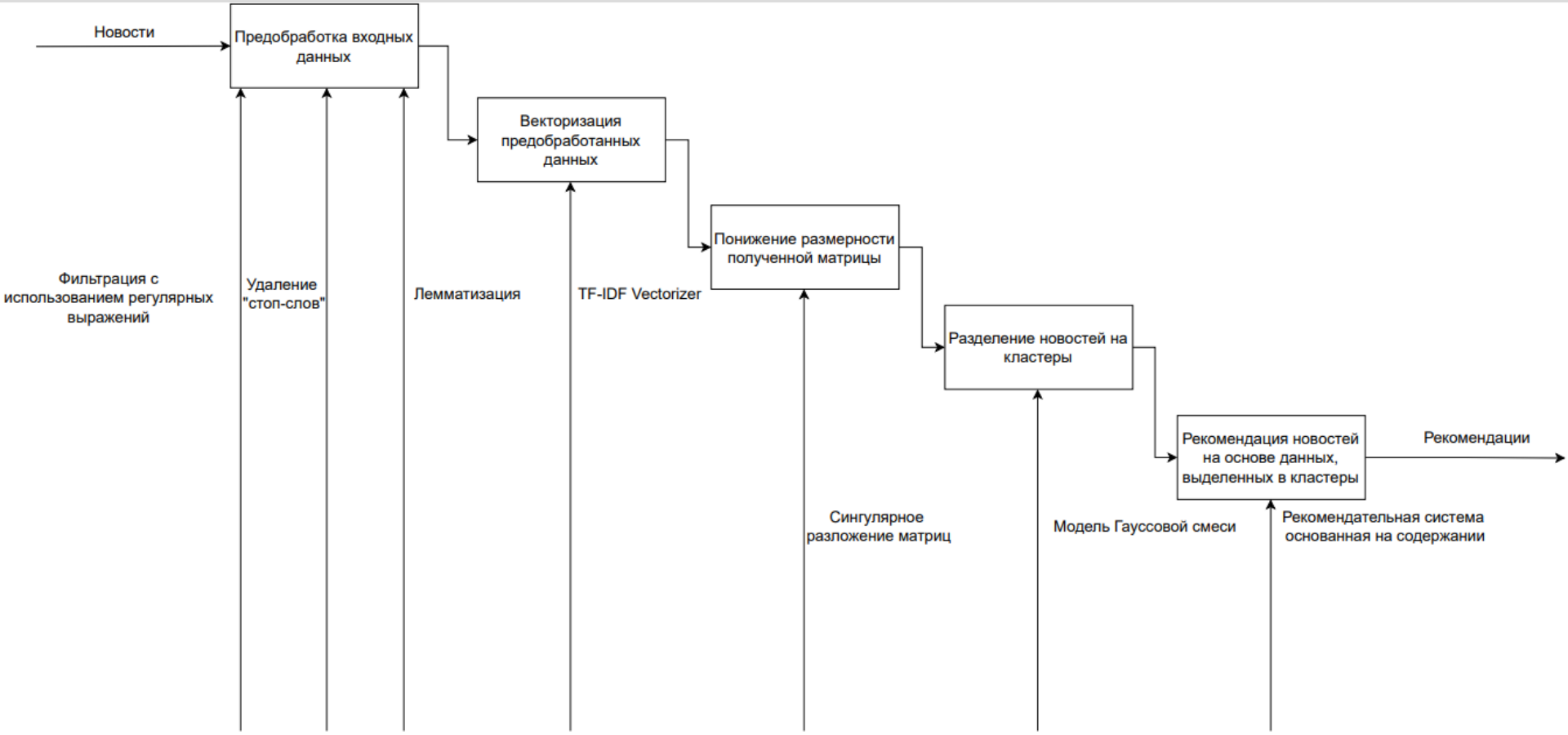
Подходы к построению рекомендательных систем

Подход	Преимущества	Недостатки
Неперсонализированный подход	Простота реализации	Не учитывает предпочтения отдельного пользователя
Коллаборативная фильтрация	Универсальность. Разнообразие рекомендации.	Проблема холодного старта. Разреженность. Смещения популярности.
Фильтрация на основе содержания	Учет необычных вкусов пользователей. Не нужна большая группа пользователей. Возможность рекомендовать непросмотренные элементы	Отсутствие разнообразия в рекомендациях.

Сравнение алгоритмов нечеткой кластеризации

Метод	Преимущества	Недостатки
Метод нечетких средних (FCM)	Скорость работы	Плохо работает с данными разных размеров и плотностей, чувствителен к выбросам.
Модель Гауссовой смеси (GMM)	Выявляет кластеры различных форм размеров и плотностей.	Работает медленнее чем FCM.

Декомпозиция рекомендательной системы



Предобработка входных данных

Этапы:

• объединить столбцы с заголовком, абстрактным описанием и самой новостью;

• удалить все символы, кроме символов из латинского алфавита;

• удалить все ”стоп-слова”;

• провести лемматизацию.

ID	Заголовок	Абстрактное описание	Новость	Дата публикации
N45436	Walmart Slashes Prices on Last-Generation iPads	Apple's new iPad releases bring big deals on last year's models...	This year, Walmart's not waiting until to offer steep deals on tech. Right now, you can save big on since new models for 2019...	10/29/2019
N23144	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding you back and keeping you from shedding that unwanted belly fat for good.	When you first start dieting and exercising, the pounds seem to melt off. But, we all hit that stagnant point where the last few pounds of belly fat just don't want to leave...	5/7/2019

Пример. 2 элемента входных данных

Векторизация предобработанных данных

Термин — это слово в начальной форме. Его вес рассчитывается как TF-IDF.

t – термин
D – коллекция документов
d – документ из коллекции D

$$TF - IDF(t, D) = TF(t, d) \times IDF(t_i, D)$$

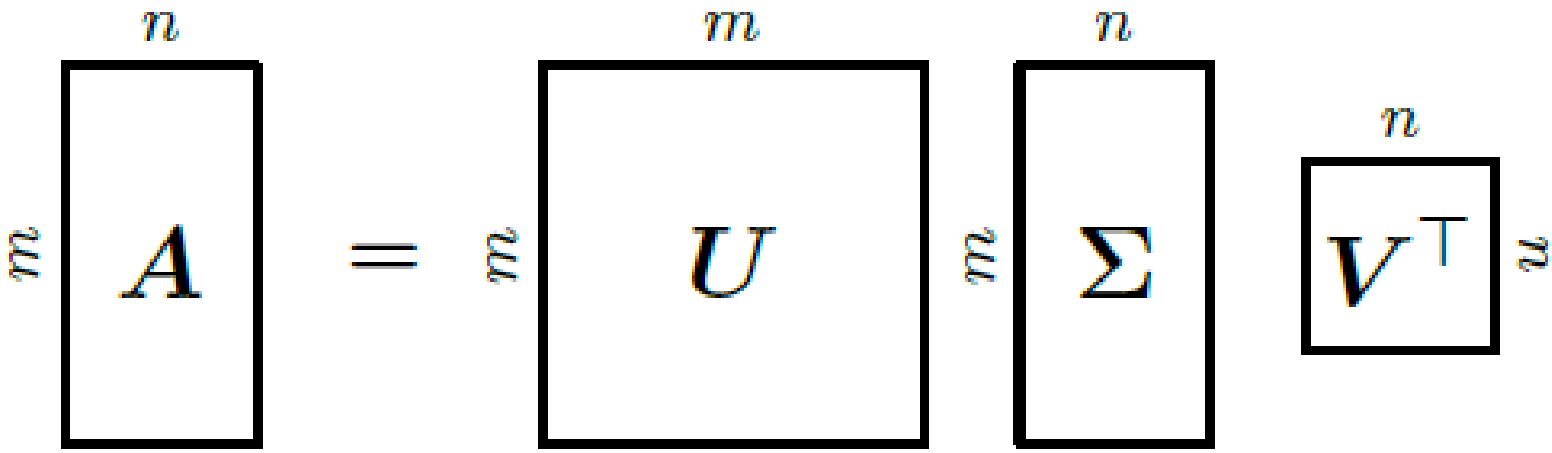
$$TF(t, d) = \frac{n_t}{\sum_k n_k}$$

n_t – количество вхождений термина t в документ
$\sum_k n_k$ – общее количество слов в документе

$$IDF(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right)$$

$|D|$ – количество документов
$|D_i \in D|$ – число документов, где t_i встретилось хотя бы один раз

Понижение размерности



A — входная матрица вещественных чисел mxn,
U и V^T ортогональные матрицы размеров mxm и nxn,
$\sum$ — матрица размера mxn с сингулярными числами на диагонали.

Разделение новостей на кластеры. Expectation – Maximization алгоритм.

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

где, μ — D-мерный вектор математических ожиданий x,
∑ — его ковариационная матрица, |∑| — определитель ковариационной матрицы, π — вероятность смешения.

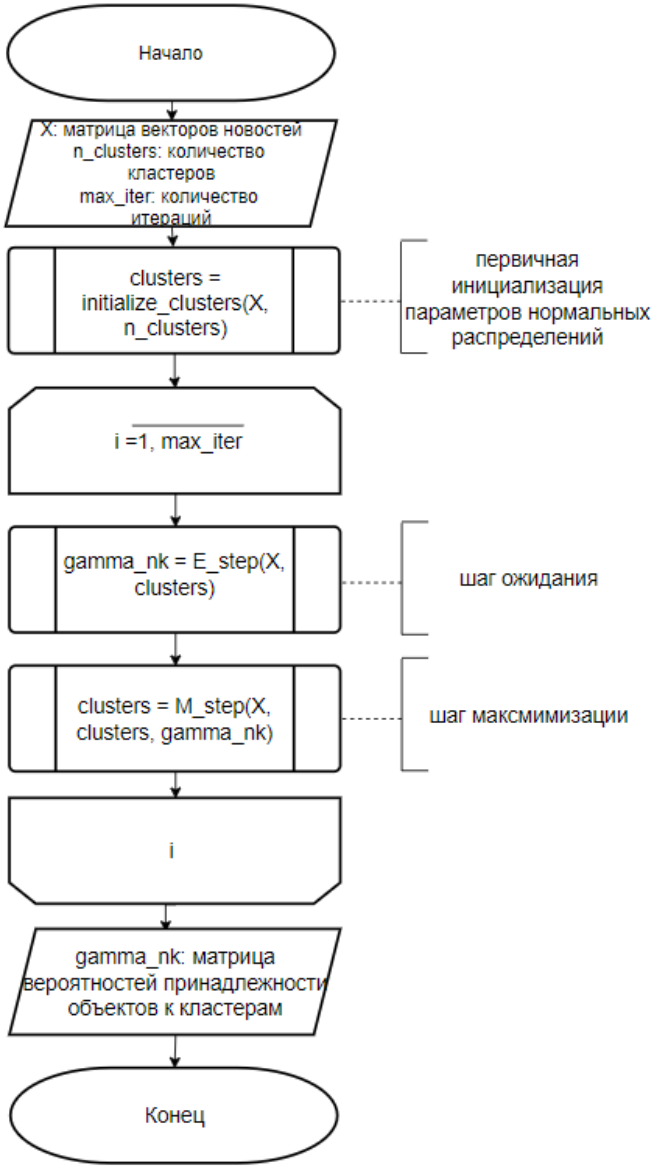
$$\gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\mu_k,\mathbf{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\mu_j,\mathbf{\Sigma}_j)}$$

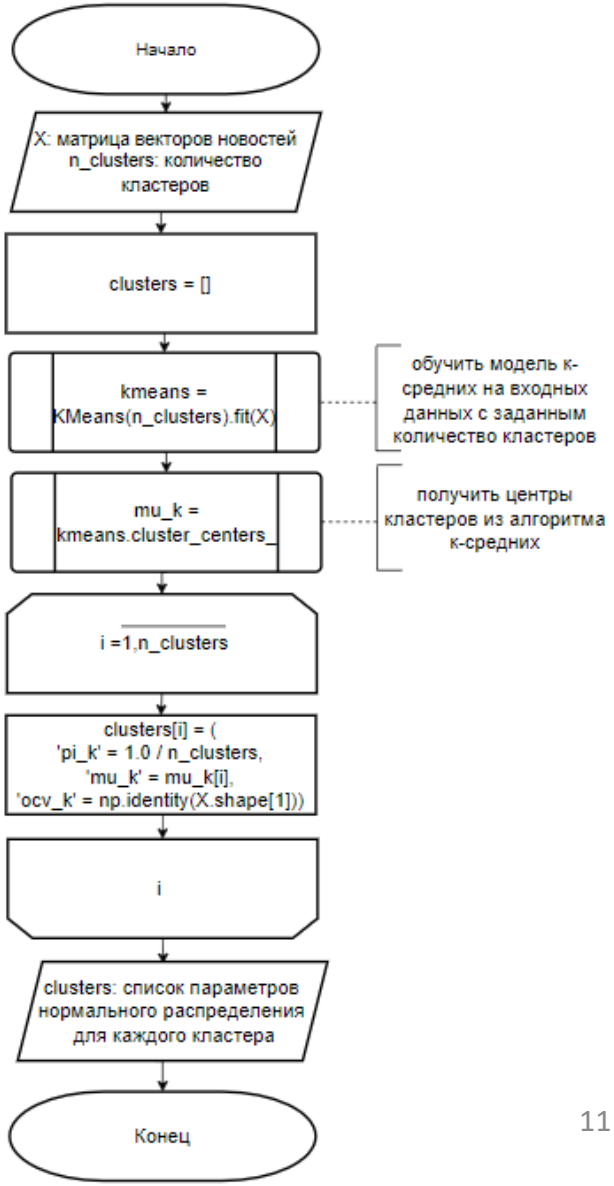
, где $\mathcal{N}$ — *функция многомерного нормального распределения*.

$$\pi_k^* = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N} \qquad \Sigma_k^* = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^T}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

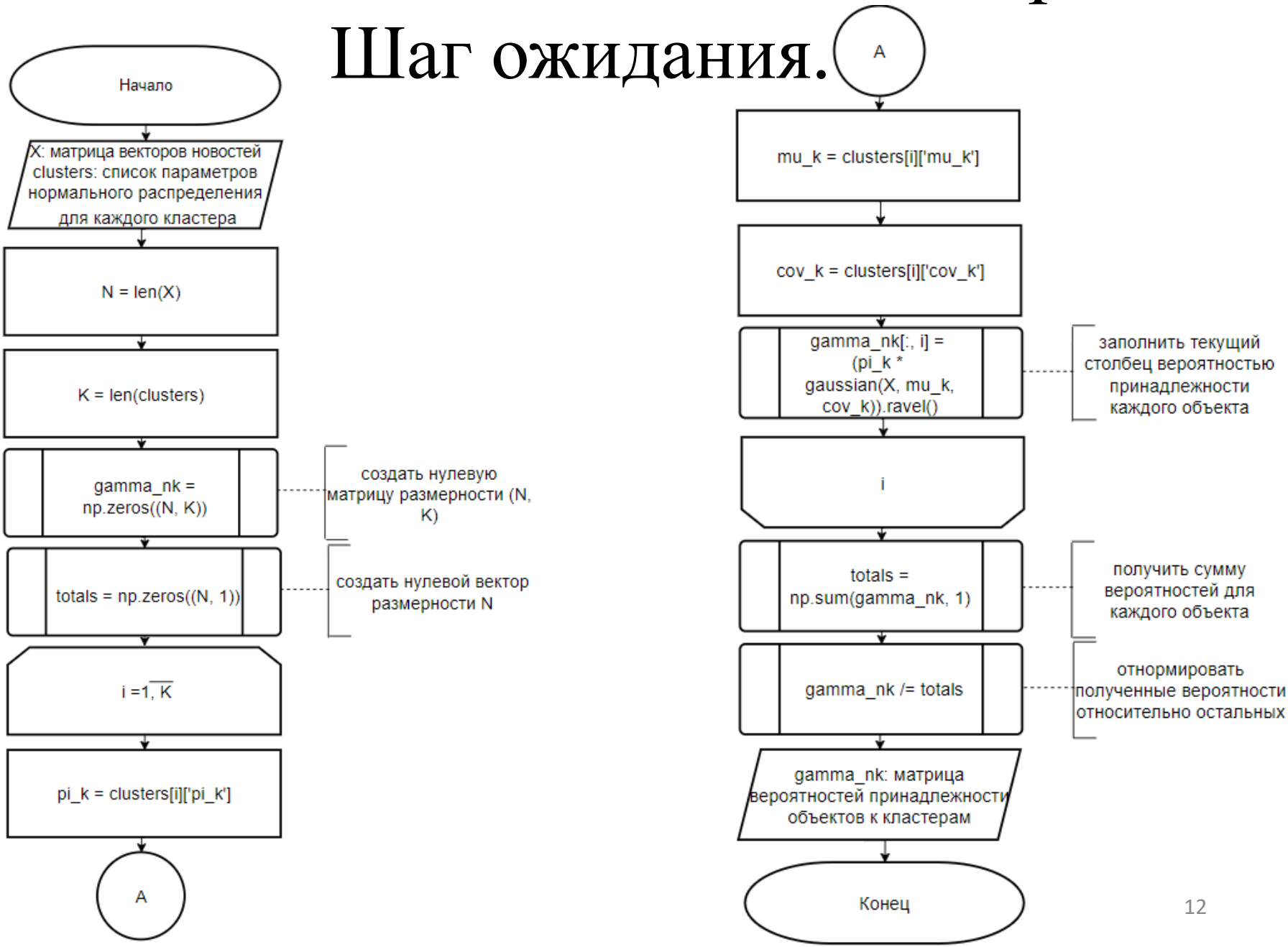
$$\mu_k^* = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

Разделение новостей на кластеры. Алгоритм Гауссовой смеси.

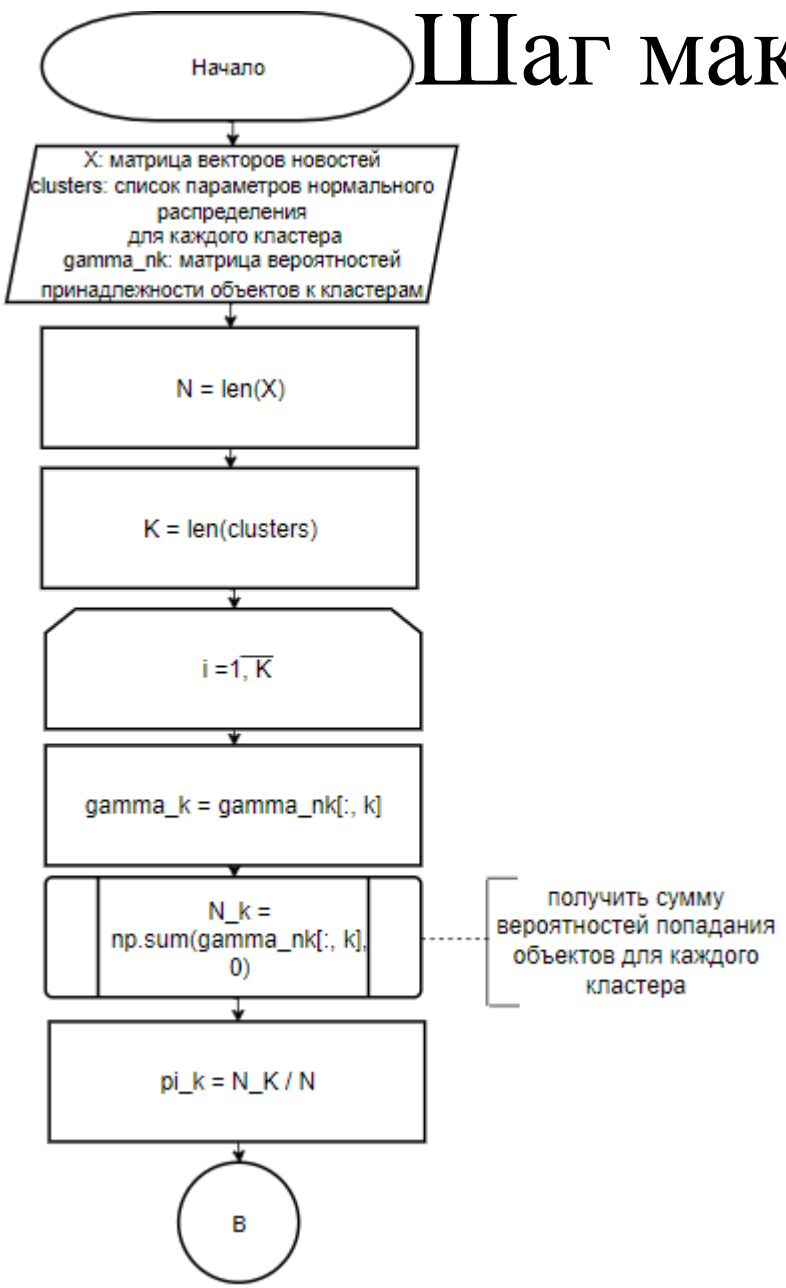


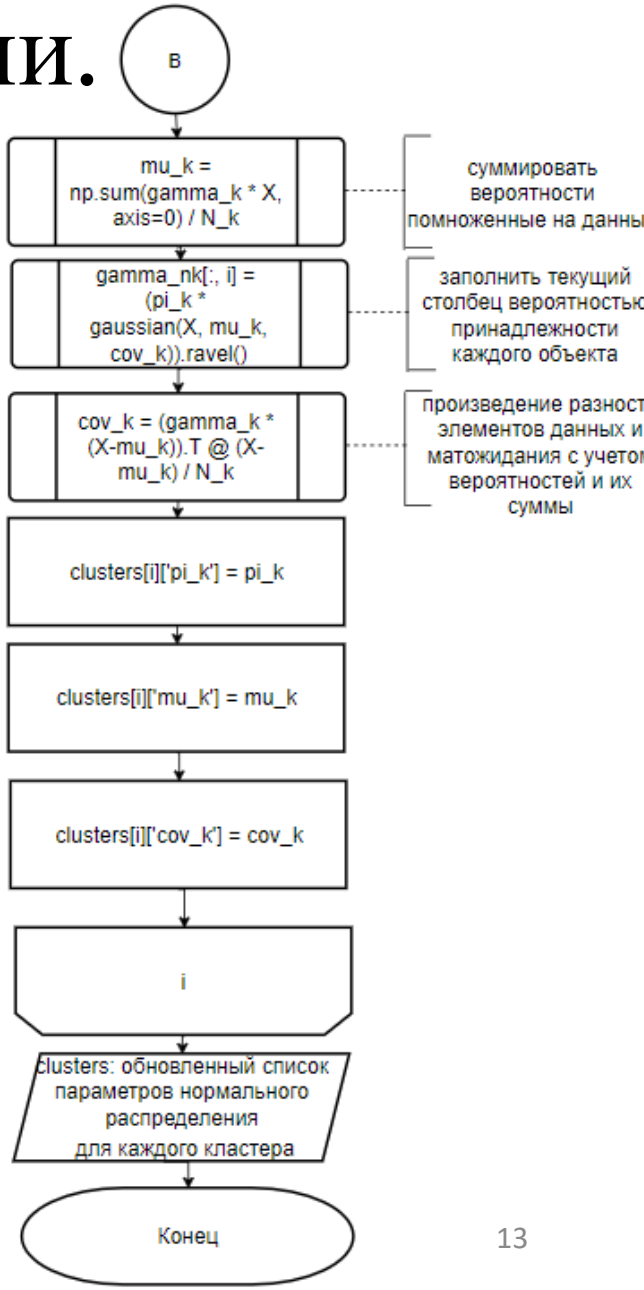


Разделение новостей на кластеры. Шаг ожидания.

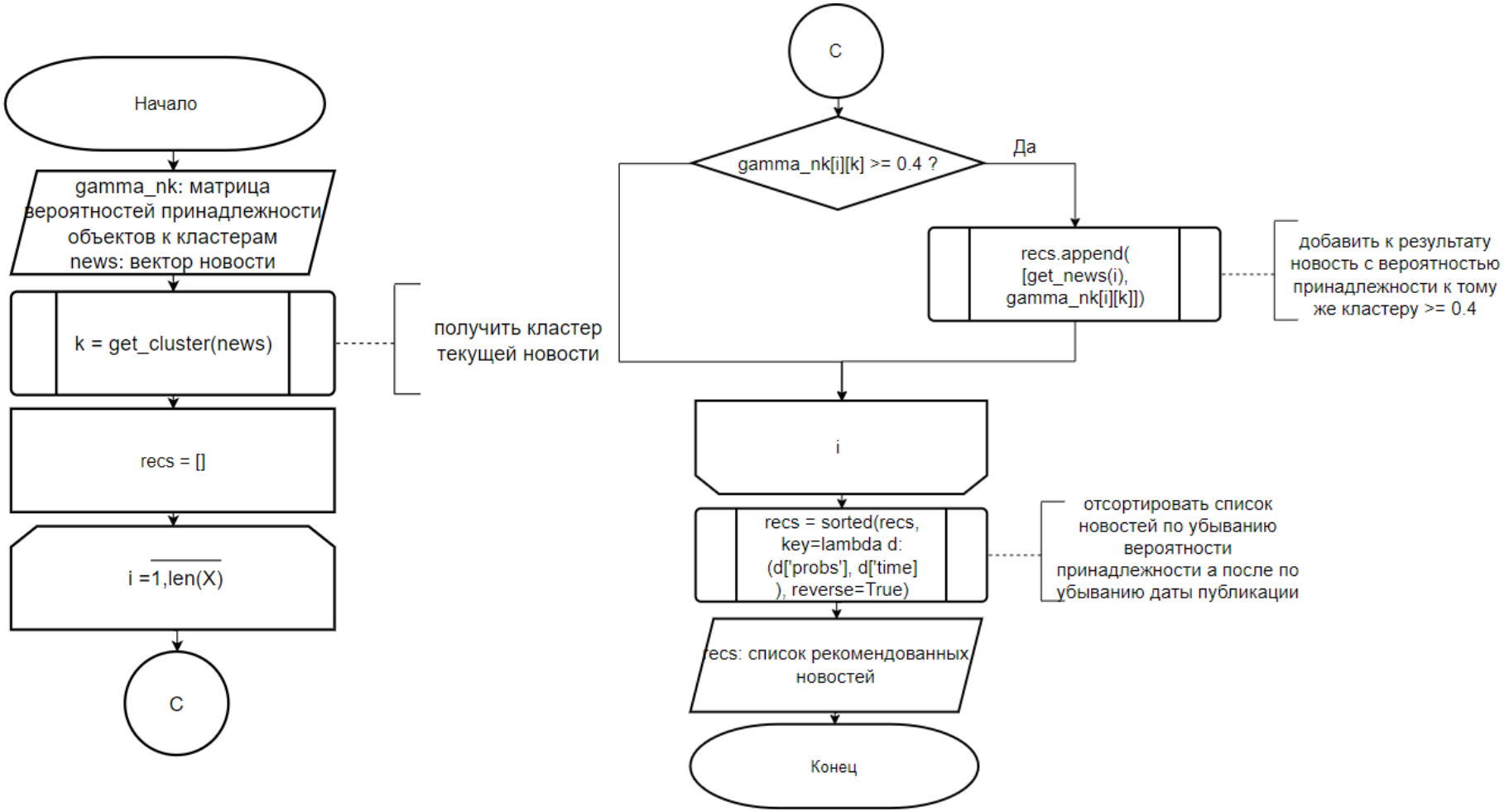


Разделение новостей на кластеры. Шаг максимизации.

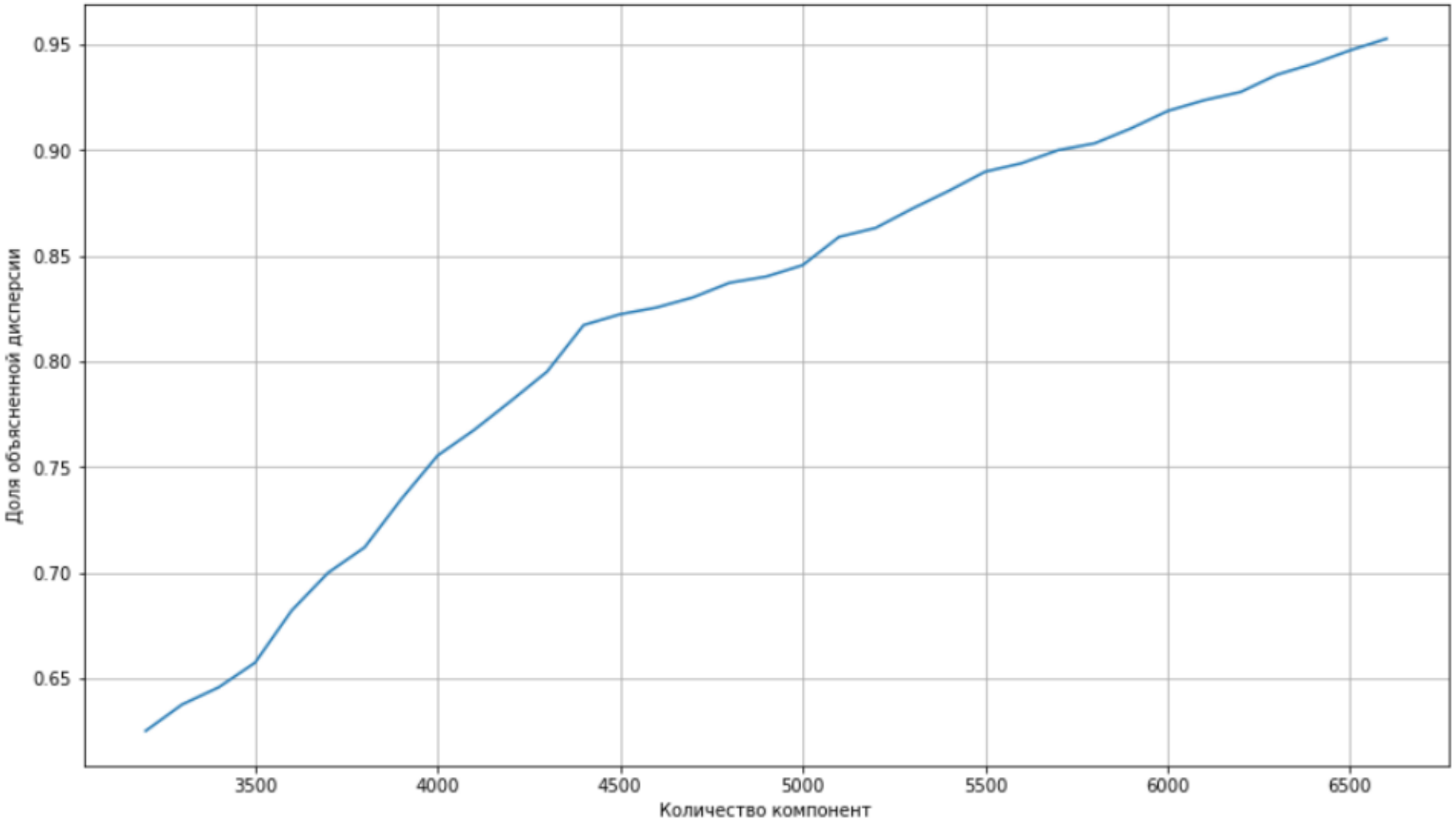




Рекомендация новостей

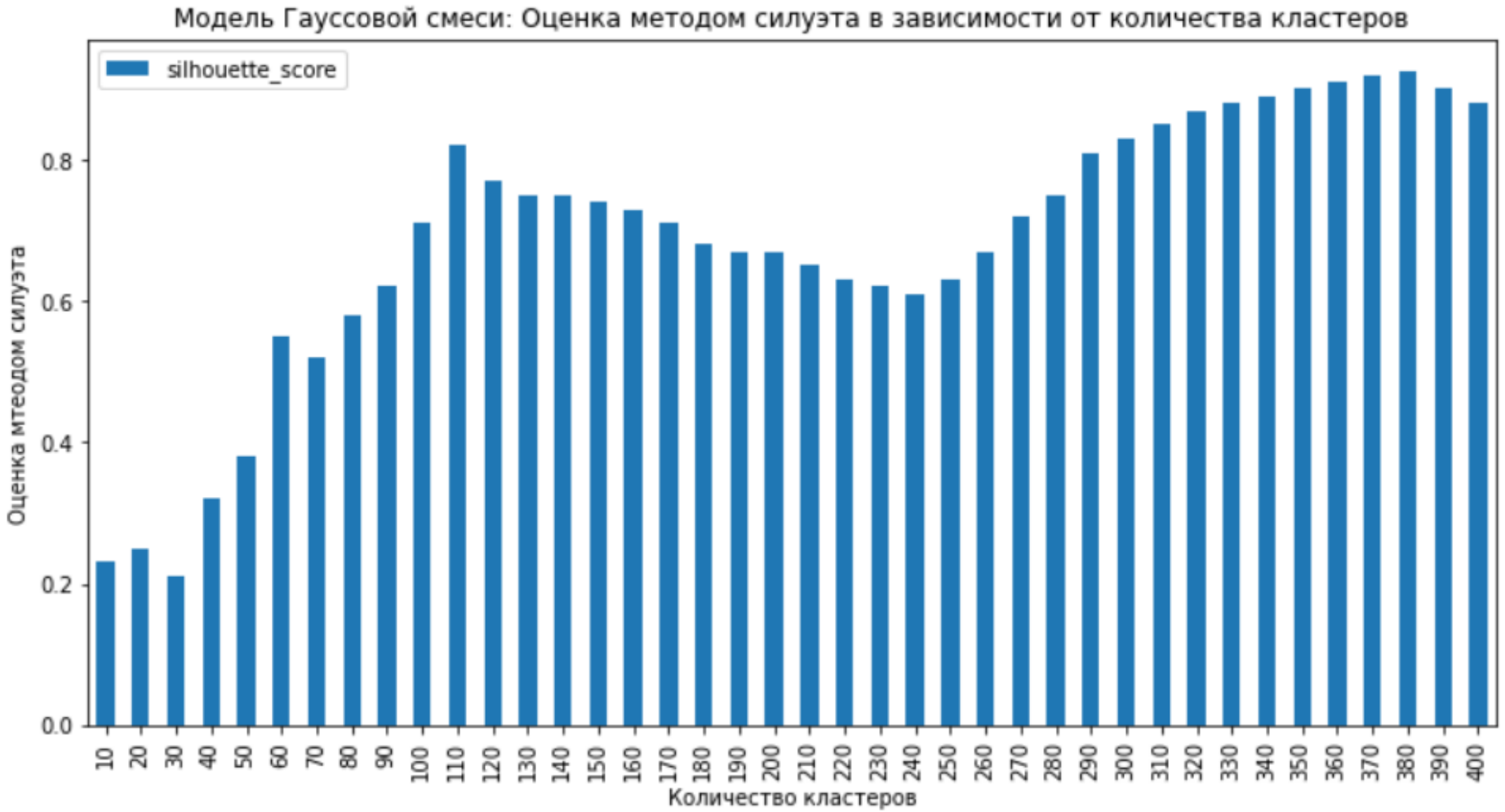


Исследования



Доля объясненной дисперсии в зависимости от количества компонент.

Исследования



Оценка методом Силуэта в зависимости от количества кластеров.

Заключение

Достигнута цель работы — разработана и программно реализована рекомендательная система новостей на основе нечеткой кластеризации.

Решены поставленные задачи:

- проведен анализ предметной области;
- проанализированы подходы к реализации рекомендательных систем ;
- разработана рекомендательная система на основе нечеткой кластеризации;
- разработано программное обеспечение для рекомендательной системы на основе нечеткой кластеризации;
- проведено исследование работоспособности реализованной рекомендательной системы и алгоритма нечеткой кластеризации.

Дальнейшее развитие

Разработанная система имеет перспективу дальнейшего развития и улучшения:

- использование альтернативного способа векторизации предобработанных новостей, улучшающих качество последующего обучения;
- адаптирование системы для работы с новостями на новостях, написанных не только на английском языке.