ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.62, 004.8, 004.93

Отчет об исследовательском проекте

на тему **«Spot the bot: семантические траектории текстов естественного языка»**

(промежуточный, этап 1)

Выполнил:
студент группы БПИ

Дата 15.02.2022

Принял:
руководитель проекта Василий Александрович Громов
Имя, Отчество, Фамилия
Заместитель руководителя департамента, старший научный сотрудник
Должность
Национальный исследовательский университет «Высшая школа экономики»
Место работы

Дата _____2022 _____ _____
Оценка (по 10-тибалльной шкале) Подпись

Москва 2022

РЕФЕРАТ

Отчёт 13 с., 2 кн., 5 рис., 2 табл., 14 источн., 2 прил.

ЧЕЛОВЕК, БОТ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО ЯЗЫКА, ХАОТИЧЕСКИЕ ВРЕМЕННЫЕ РЯДЫ, ГЕНЕРАЦИЯ ТЕКСТОВ, КЛАСТЕРИЗАЦИЯ УИШАРТА, SVD, WORD2VEC, ELMO, BERT.

Объектом исследования являются паттерны генерации текстов человеком, отличающие его от ботов.

Цель работы – разработка алгоритма, решающего задачу классификации человека или бота как автора сгенерированного текста.

В рамках данной работы должны быть выполнены следующие действия:

1) Реализация и оптимизация алгоритма кластеризации Д. Уишарта, модифицированного А. В. Лапко и С. В. Ченцовым.

2) Реализация алгоритмов вычисления внутренних метрик качества кластеризации.

3) Разработка словарей эмбеддингов для русского, английского и румынского языков с помощью сингулярного разложения, моделей Word2Vec, ELMo, BERT.

4) Исследование зависимостей мер качества кластеризации областей семантического пространства языка от способа получения эмбеддингов и объёма языкового корпуса, проектирование оптимального варианта пространства.

5) Поиск признаков траекторий из текстовых n-грамм, уникальных для текстов, сгенерированных человеком.

В качестве предполагаемого результата проекта должно быть представлено текстовое или кодовое описание бинарного классификатора, на вход получающего литературный текст и возвращающего одну из двух меток класса: «человек», «бот».

Итоговый алгоритм может быть внедрён в различные социальные сети с целью ускорения и оптимизации процесса поиска и блокировки спам-аккаунтов, что позволит компании-пользователю повысить уровень комфорта и безопасности сети, выгодно выделиться на фоне конкурентов.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем отчете о НИР применяются следующие термины с соответствующими определениями:

Бот – Программа, автоматически выполняющая какие-либо действия через интерфейсы, предназначенные для людей.

Временной ряд – Статистическая последовательность из значений параметров исследуемого процесса, синхронизированная с временной шкалой.

Внутренняя мера качества кластеризации – Метрика, вычисляющаяся исключительно по данным и результатам кластеризации, без доступа к внешней информации.

Датасет – Сгруппированный набор данных.

Документ – Файл, содержащий текст в том или ином виде.

Кластеризация – Процесс группировки объектов некоторого множества на основе похожести их признаков.

Компактность кластера – Метрика, основанная на идее того, что разделение тем лучше, чем ближе находятся к друг другу объекты внутри одного кластера. Таким образом, сама мера является внутрикластерным расстоянием, которое предполагается минимизировать.

Контекст – Набор слов или предложений, объединённый одной предметной областью и уточняющий значения входящих в него синтаксических структур.

«Мешок слов» или же «Bag of words» - Принцип обработки синтаксической сущности, подразумевающий отсутствие учёта упорядоченности входящих в неё слов.

Отделимость кластера – Метрика, вычисляемая в гипотезе о том, что качество кластеризации напрямую зависит от расстояния между объектами разных кластеров. Чем это расстояние больше, тем лучше разделение.

Семантическое пространство языка – Представление модели мира человека в терминах естественного языка; структура, ограничивающая всевозможные комбинаторные цепочки слов до осмысленных согласно с общим представлением о мире.

Сингулярное разложение - Декомпозиция вещественной матрицы с целью ее приведения к каноническому виду. Обладает множеством полезных свойств при решении задач, связанных с матричным представлением данных.

Субпредложение – Непрерывная часть языкового корпуса.

Субсэмплирование - Процесс изъятия наиболее частотных слов из текстового анализа с целью ускорения алгоритма обучения и увеличения качества оаботы модели.

Токен – Последовательность символов, соответствующая языковой лексеме, полученной в процессе анализа текста.

Траектория языка – Упорядоченная последовательность кластеров в языковой семантической сети.

Эмбеддинг – Результат некоторого отображения слова или любой другой языковой сущности в конечномерное векторное пространство.

CBoW – Continuous Bag of Words, модель, предсказывающая слово по окружающему его контексту.

LSTM – Рекуррентная нейронная сеть с механизмом долгосрочной кратковременной памяти.

Skip-gram – Модель, по конкретному слову в тексте предсказывающая окружающий его контекст.

SVD – Сингулярное разложение (аббревиатура).

ВВЕДЕНИЕ

С развитием направления обработки естественного языка к разного рода сомнительным личностям попадают в руки всё более мощные инструменты для синтеза качественных текстов, из-за чего за последние годы социальные сети заполнили спам-боты. Они мешают процессу взаимодействия пользователей, снижают общий уровень безопасности и, в целом, сильно ухудшают опыт работы с онлайн-мессенджерами. Крупные компании в попытках решить проблему организуют команды, занимающиеся поиском и блокировкой подозрительных аккаунтов. Однако зачастую человек оказывается неэффективным в задаче распознавания ботов по ряду причин: низкая скорость обработки текстов, размытые критерии классификации, большие денежные затраты на содержание штата работников. Сама собой возникает идея о том, что бороться с машинными мощностями следует ими

же: необходимо разработать алгоритм, способный за короткое время идентифицировать ботов и автоматически ограничивать им доступ.

Классифицировать ботов не имеет смысла, так как все они обладают сильно различающимися особенностями, зависящими от архитектуры той или иной модели. Вместо этого значительно проще абстрагироваться от конкретных видов ботов и анализировать тексты по наличию паттернов, присущих человеку: тогда задача сокращается до бинарной классификации «человек / не человек (бот)». Поведение человека достаточно исследовано и относительно предсказуемо, что позволяет построить хорошо интерпретируемую модель.

Алгоритм классификации предполагается разрабатывать на основе гипотезы о том, что люди в процессе составления текстов образуют траектории в семантическом пространстве языка, обладающие различимыми отличительными особенностями, которые могут быть выявлены в результате анализа следующих характеристик последовательностей текстовых n-грамм:

1) Частота посещения кластеров, избегаемых ботами.

2) Частота переходов из кластера в кластер, её постоянство.

3) Расстояния между соседними в траектории кластерами.

4) Наличие или отсутствие циклических участков в траектории.

5) Прочие особенности, выявленные в процессе экспериментов.

Так как в рамках НИР должны быть разработаны сущности, необходимые для исследования, настоящий отчёт фиксирует результаты подготовительного этапа работы: получение словарей эмбеддингов; реализация и оптимизация алгоритма кластеризации языкового пространства, а также внутренних метрик её оценивания.

1. ПОДГОТОВКА СЛОВАРЕЙ ЭМБЕДДИНГОВ

Этап подготовки словарей целесообразен только для моделей, работающих с предложениями по принципу «мешка слов», так как в данном случае отображение слова в векторное пространство будет с некоторой погрешностью единственным. Это касается методов SVD-разложения, Word2Vec.

В первую очередь из открытых источников[1-3] были загружены датасеты документов, представленные наборами литературных текстов, статей из газет, журналов, Википедии, так как такой выбор позволяет наиболее

широко охватить языковое пространство и не заострять внимание на какой-то определённой тематике. В дальнейшем с помощью загруженных данных составлялись три очищенных языковых корпуса: русский, английский и румынский. Подготовка корпусов необходима для того, чтобы алгоритмы моделей извлечения эмбеддингов не тратили собственные вычислительные ресурсы на изучение падежей, склонений, сокращений и т. п., а также не заостряли внимание на словах, не обладающих весомой смысловой нагрузкой, но при этом сильно зашумляющих текст. Процесс извлечения корпуса проводился для каждого языка в несколько этапов. У всех текстовых документов, входящих в корпус:

1) Удалялись спецсимволы, не являющиеся разделителями предложений или ключевыми для конкретного языка.

2) Буквы предложений приводились к нижнему регистру, за исключением первой заглавной.

3) Удалялись символы окончания строк и предложений, производилась развёртка в одну строку.

4) Слова приводились к начальной форме.

5) Заранее обозначенные стоп-слова заменялись на токены.

Затем обработанные тексты конкатенировались построчно в общий языковой корпус. Таким образом, в результате вышеописанных процедур на выходе получался текстовый файл, каждой строкой которого являлся один очищенный документ из ранее загруженного датасета.

Очищенные текстовые корпусы передавались на вход алгоритмов SVD и Word2Vec, в результате работы которых были получены словари эмбеддингов, использованные в дальнейшем для конструирования языковых семантических пространств. Более подробное описание моделей и алгоритмов приведено ниже.

2. МОДЕЛИ И АЛГОРИТМЫ ПОЛУЧЕНИЯ ЭМБЕДДИНГОВ

А. СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ

За основу метода был взят алгоритм, описанный в книге Jerome R. Bellegarda[4]. Сначала составляется матрица совпадений W размера $M \times N$ (число уникальных слов $\times$ количество документов), каждая строка и столбец которой соответствуют конкретным слову и документу соответственно. Значения ячеек вычисляются по формуле:

$$w_{i,j} = (1 - \varepsilon_i)\frac{k_{i,j}}{\lambda_j}, \tag{2.1}$$

где $k_{i,j}$ – число раз, которое слово r_i встретилось в документе c_j, λ_j – общее число слов в c_j, а ε_i – нормализованная энтропия слова r_i в полной коллекции слов $\mathcal{N}$. Обозначим за τ_i сумму $\sum_{j=1}^{N} k_{i,j}$. Тогда, в свою очередь, энтропия может быть вычислена как:

$$\varepsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\frac{k_{i,j}}{\tau_i}\log\frac{k_{i,j}}{\tau_i}, \tag{2.2}$$

Таким образом, значение ε_i, близкое к 1 означает, что слово r_i распределено по множеству всех документов в коллекции, в то время как близость к 0 означает присутствие только в определённом контексте. Это позволяет снизить влияние вводных слов, союзов, предлогов и т. п., относительно контекстуально-образующих единиц.

В теории, уже после этого этапа можно было бы сопоставить каждому слову r_i векторное представление w_i. Однако на практике часто оказывается, что размерности полученных векторов чрезвычайно велики, а сами векторы к тому же сильно разрежены. Эти проблемы решаются с помощью сингулярного разложения матрицы W (Рис. 1):

$$W \approx \widehat{W} = USV^T, \tag{2.3}$$

С его помощью единицу r_i можно представить в виде эмбеддинга:

$$v_i = u_i S(1{:}n), \tag{2.4}$$

где n – желаемая размерность векторного пространства. В данном случае используется одно из полезных свойств SVD разложения, вследствие которого

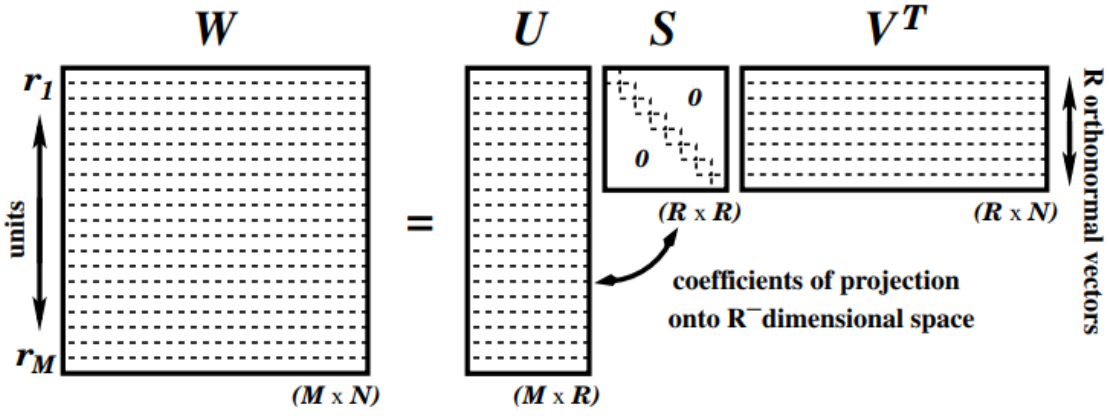

Рисунок 1

все измерения элементов векторного пространства L ($\forall i \in \overline{1,M},\ u_i S \in L$) ранжированы в порядке убывания своей информационной значимости.

При выполнении настоящей НИР в качестве инструмента получения эмбеддингов методом сингулярного разложения использовалась модель TfidfVectorizer[5] из бесплатной Python-библиотеки для машинного обучения Scikit-learn.

B. WORD2VEC

В данном пункте используется модель Word2Vec, основанная на работах Tomas Mikolov[6-7] и включает два механизма обработки текстов: CBoW и Skip-gram. Кратко описать принцип работы можно следующим образом. Переданный на вход алгоритма текст проходит этапы разбиения на субпредложения и субсэмплирования. По полученному набору сущностей проходит окно фиксированного размера и на каждой итерации прогоняет включённые в него слова через одну из вышеупомянутых моделей (а иногда и через обе сразу), принципиально различающихся по своей конструкции.

CBoW (Continuous Bag of Words) представляет собой перцептрон с одним скрытым слоем, на вход которого передаются закодированные посредством one-hot encoding слова. На выходе возвращается вектор

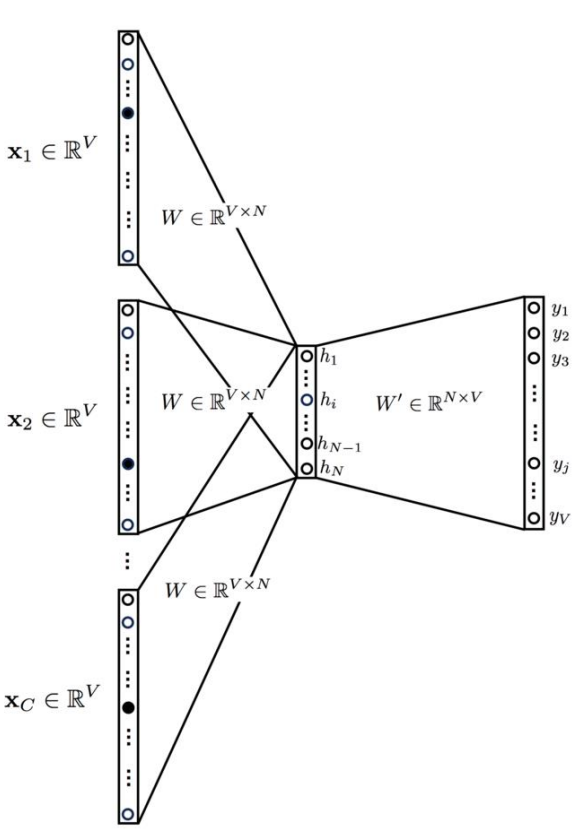

Рисунок 3

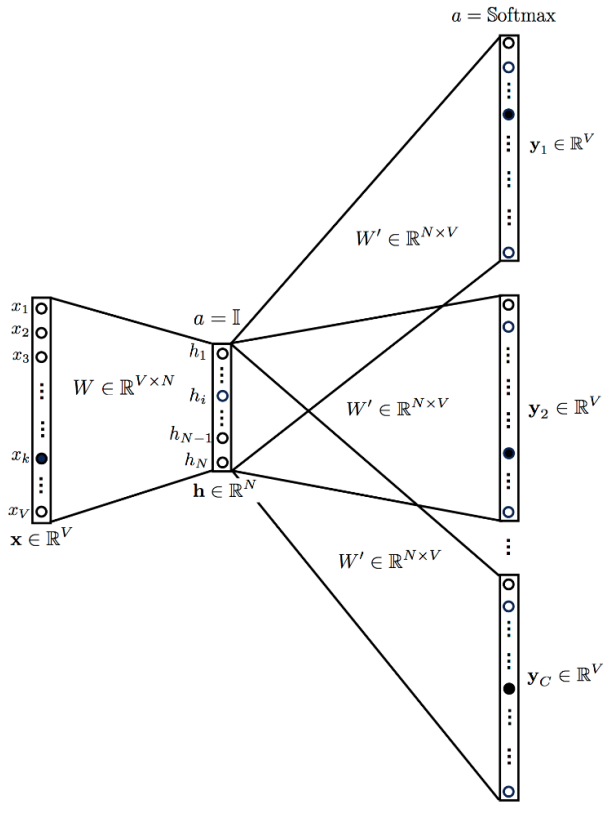

Рисунок 2

распределения вероятностей принадлежности некоторого скрытого в окне элемента к каждому слову в словаре. Таким образом происходит предсказание слова по окружающего его контексту. Полученная в результате обучения матрица W' (матрица весов на скрытом слое) и является словарём эмбеддингов (каждому слову в соответствие поставлен вектор-столбец).

Skip-gram отличается от CBoW тем, что, наоборот, стремится предсказать по слову контекст, в котором оно находится и является инвертированной копией описанного ранее перцептрона. Схемы сетей CBoW и Skip-Gram представлены на Рисунках 2 и 3 соответственно.

В рамках НИР использовалась модель Word2Vec[8] от фреймворка Gensim.

C. ELMO

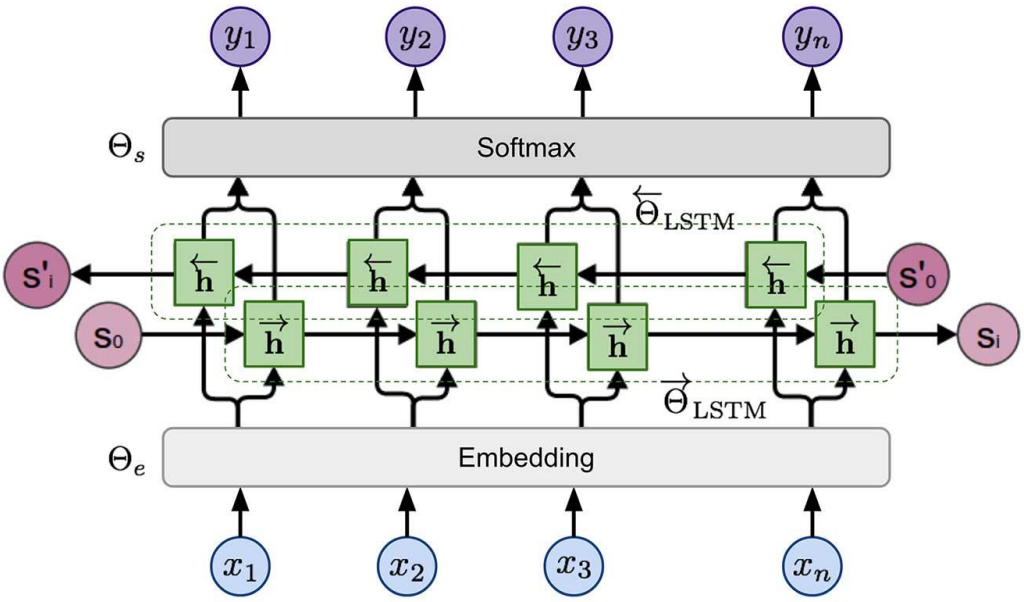

Рисунок 4

ELMo, так же, как и упомянутый ниже BERT, отличается от двух предыдущих моделей тем, что учитывает контекст слова в предложении, из-за чего пропадает всякий смысл в составлении словарей эмбеддингов. Архитектура сети представлена на Рисунке 4: слова предложения с помощью свёртки преобразуются в векторы с последующей передачей в слои двунаправленных LSTM. Таким образом задействуется механизм памяти для анализа контекста как до, так и после прогнозируемого слова. В процессе работы использовалась конкретная реализация[9] мультиязычной ELMo, предобученной преимущественно на текстах Википедии и описаниях веб-сайтов.

D. BERT

BERT (Рис. 5) является нейронной сетью на основе последовательности энкодеров модели OpenAI Transformer[10] и принципиально отличается от ELMo использованием исключительно механизма внимания без LSTM.

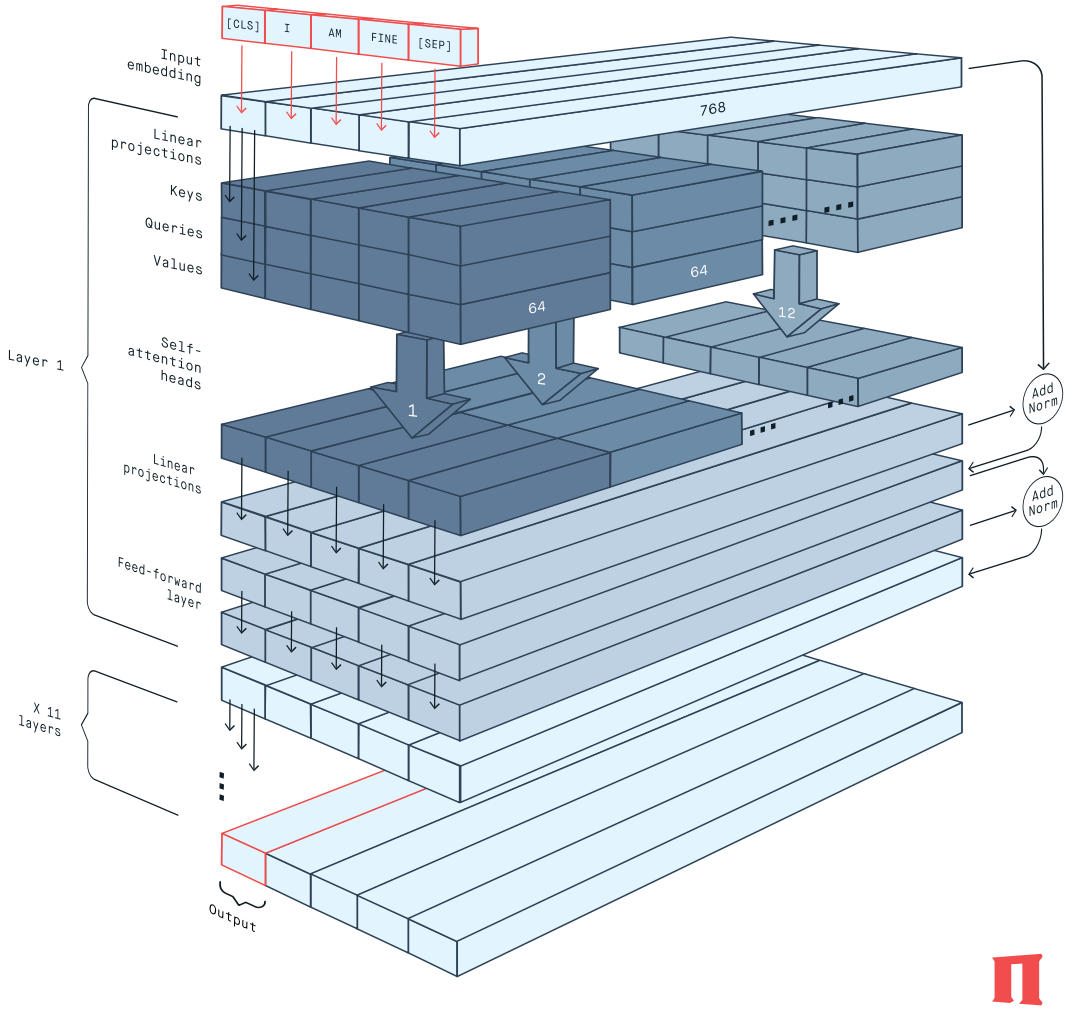

Рисунок 5

В рамках настоящей НИР использовались модели BERT интернет-сообщества специалистов по машинному обучению Hugging Face в различных вариациях в зависимости от языка.

3. АЛГОРИТМ КЛАСТЕРИЗАЦИИ УИШАРТА

Техника кластеризации, использованная в настоящей НИР, является алгоритмом Уишарта[11, 12] под модификациями Лапко и Ченцова[13]. Изначально планировалось использовать DBSCAN, но в силу меньшей требовательности к вычислительным мощностям и памяти при схожих положительных качествах: хороших результатах на сильно зашумлённых данных, способности к выделению субкластеров, разделению композиций невыпуклых многоугольников - был выбран именно алгоритм Уишарта.

В целях оптимизации времени и памяти алгоритма были добавлены механизмы подсчёта на графическом процессоре и разбиения выборки на пакеты.

4. ВЫБОР МЕТРИК КАЧЕСТВА КЛАСТЕРИЗАЦИИ

Так как работа с n-граммами подразумевает кластеризацию без учителя в силу неизвестности не только истинности разделения, но даже количества предполагаемых кластеров, для оценки качества кластеризации применимы только внутренние метрики. Все такие метрики используют понятия компактности (*cohesion*) и отделимости (*separation*). Изначально в качестве мер качества кластеризации планировалось использовать все внутренние меры, описанные в работе Hui Xiong, Zhongmou Li[14]. Однако в процессе экспериментов было принято решение о том, что стоит использовать лишь те метрики, которые способны к учёту наличия субкластеров. Такими оказались Silhouette index (S), Davies-Bouldin index (DB), Xie-Beni index (XB), SD and S-Dbw validity indexes, а также CVNN (cluster validity index based on nearest neighbours). Полный список метрик, а также оценка их возможностей доступны в приложениях A1 и A2 соответственно.

5. СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Romanian Wikisource Literature Corpus [Электронный ресурс] / Kaggle. Режим доступа: https://www.kaggle.com/crischir/romanian-wikisource-literature-corpus, свободный. (дата обращения: 05.12.2021).

2. Romanian Wikisource Literature Corpus [Электронный ресурс] / Kaggle. Режим доступа: https://www.kaggle.com/crischir/romanian-wikisource-literature-corpus, свободный. (дата обращения: 05.12.2021).

3. Romanian Wikisource Literature Corpus [Электронный ресурс] / Kaggle. Режим доступа: https://www.kaggle.com/crischir/romanian-wikisource-literature-corpus, свободный. (дата обращения: 05.12.2021).

4. Jerome R. Bellegarda – Latent Semantic Mapping: Principles and Applications, 2007 – 101 с.

5. TfidfVectorizer [Электронный ресурс] / Scikit-learn. Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, свободный. (дата обращения: 17.12.2021).

6. Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, and J. Dean // [Электронный ресурс]: Cornell University arXiv:1301.3781v3 [cs.CL] 2013 – Режим доступа: https://arxiv.org/pdf/1301.3781.pdf, свободный. (дата обращения: 23.12.2021).

7. Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean // [Электронный ресурс]: Cornell University arXiv:1310.4546v1 [cs.CL] 2013 – Режим доступа: https://arxiv.org/pdf/1310.4546.pdf, свободный. (дата обращения: 23.12.2021).

8. Word2Vec [Электронный ресурс] / Gensim. Режим доступа: https://radimrehurek.com/gensim/models/word2vec.html, свободный. (дата обращения: 16.01.2022).

9. ELMoForManyLangs [Электронный ресурс] / GitHub. Режим доступа: https://github.com/HIT-SCIR/ELMoForManyLangs, свободный. (дата обращения: 29.01.2022).

10. Vaswani, A. Attention is All you Need / I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett // [Электронный ресурс]: Cornell University arXiv:1706.03762v5 [cs.CL] 2017 – Режим доступа: https://arxiv.org/pdf/1706.03762.pdf, свободный. (дата обращения: 06.02.2022).

11. D. Wishart, A numerical classification methods for deriving natural classes // Nature – 1969 – C. 97-98.

12. H. H. Bock, Automatic Classification, Vandenhoeck and Rupert // Göttingen – 1974.

13. A.V. Lapko, S.V. Chentsov, Nonparametric Information Processing Systems // Nauka, Novosibirsk - 2000.

14. Hui Xiong, Zhongmou Li Clustering Validation Measures, in: Charu C. Aggarawal and Chandan K. Reddy (Eds.) Data Clustering: Algorithms and Applications // 2014 // Chapman and Hall / CRC C. 571-605.

6. ПРИЛОЖЕНИЯ

Приложение А1 – внутренние метрики качества кластеризации.

Метрика	Формула
RMSSTD	$\left[\frac{\sum_{i=1}^{NC}\sum_{x\in C_i}\|x-c_i\|^2}{P\sum_{i=1}^{NC}(n_i-1)}\right]^{\frac{1}{2}}$
R-squared (RS)	$\frac{\sum_{x\in D}\|x-c_i\|^2-\sum_{i=1}^{NC}\sum_{x\in C_i}\|x-c_i\|^2}{\sum_{x\in D}\|x-c_i\|^2}$
Modified Hubert Γ statistic (Γ)	$\frac{2}{n(n-1)}\sum_{x\in D}\sum_{y\in D}d(x,y)d_{x\in C_i,y\in C_j}(c_i,c_j)$
Calinski-Harabasz index (CH)	$\frac{\sum_{i=1}^{NC}n_i d^2(c_i,c)\,(n-NC)}{\sum_{i=1}^{NC}\sum_{x\in C_i}d^2\,(x,c_i)}$
I index (I)	$\left[\frac{1}{NC}\frac{\sum_{x\in D}d(x,c)}{\sum_{i=1}^{NC}\sum_{x\in C_i}d(x,c_i)}max_{i,j}d(c_i,c_j)\right]^p$
Dunn's Index (D)	$min_i\left[min_{i\neq j}\left(\frac{min_{x\in C_i,y\in C_j}d(x,y)}{max_j\left[max_{x,y\in C_k}d(x,y)\right]}\right)\right]$
Silhouette index (S)	$\frac{1}{NC}\sum_{i=1}^{NC}\left[\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x,i)-a(x,i)}{\max\,(b(x,i),a(x,i))}\right]$ $a(x,i)=\frac{1}{n_i-1}\sum_{y\in C_i,y\neq x}d(x,y)$ $b(x,i)=min_{j,j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j}d(x,y)\right]$
Davies-Bouldin index (DB)	$\frac{1}{NC}\sum_{i=1}^{NC}max_{j,j\neq i}\left(\frac{\frac{1}{n_i}\sum_{x\in C_i}d(x,c_i)+\frac{1}{n_j}\sum_{x\in C_j}d(x,c_j)}{d(c_i,c_j)}\right)$
Xie-Beni index (XB)	$\frac{1}{n}\frac{\sum_{i=1}^{NC}\sum_{x\in C_i}d^2(x,c_i)}{min_{i,i\neq j}d^2(c_i,c_j)}$
SD index (SD)	$Dis(NC_{max})Scat(NC)+Dis(NC)$ $Scat(NC)=\frac{1}{NC}\sum_{i=1}^{NC}\frac{\|\sigma(C_i)\|}{\|\sigma(D)\|}$ $Dis(NC)=\frac{max_{i,j}d(c_i,c_j)}{min_{i,j}d(c_i,c_j)}\sum_{i=1}^{NC}\left(\sum_{j=1}^{NC}d(c_i,c_j)\right)^{-1}$

S_Dbw index (S_Dbw)	$$Scat(NC) + Dens_{bw}(NC)$$ $$Dens_{bw}(NC) = = \frac{1}{NC(NC-1)} \sum_{i \neq j} \frac{\sum_{x \in C_i \cup C_j} f(x, u_{i,j})}{max\left(\sum_{x \in C_i} f(x, c_i), \sum_{x \in C_i} f\left(x, c_j\right)\right)}$$ $$f(x, y) = I(\|x - y\| < stdev)$$ $$u_{i,j} = 0.5(c_i + c_j)$$
CVNN index	$$Sep(NC, k) + Com(NC)$$ $$Com(NC) = \sum_{i=1}^{NC} \left[\frac{2}{n_i(n_i - 1)} \sum_{x,y \in C_i} d(x, y) \right]$$ $$Sep(NC, k) = max_i^{NC} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{q_j}{k} \right)$$

Таблица 1

Обозначения: D – набор данных; n – количество элементов в D; c – центр D; P – размерность D; NC – число кластеров; C_i – i-й кластер; n_i - количество объектов в кластере i; q_j – количество ближайших соседей элемента j кластера C_i, которые не принадлежат кластеру C_i; c_i – центр кластера C_i; $\sigma(A)$ – дисперсия вектора в множестве A; $d(x, y)$ – расстояние между x, y.

Приложение А2 – оценка возможностей внутренних метрик.

Метрика	Монотонность	Шум	Плотность	Субкластеры	Ассим. распределение	Произвольная форма
RMSSTD	×	-	-	-	-	-
RS	×	-	-	-	-	-
Γ	×	-	-	-	-	-
CH	✓	×	✓	✓	×	×
I	✓	✓	×	✓	✓	×
D	✓	×	✓	×	✓	×
S	✓	✓	✓	×	✓	×
DB	✓	✓	✓	×	✓	×
SD	✓	✓	✓	×	✓	×
S_Dbw	✓	✓	✓	✓	✓	×
XB	✓	✓	✓	×	✓	×

CVNN	✓	✓	✓	✓	✓	✓

Таблица 2

Легенда:

- – не проверялось;
× – проверка не пройдена;
✓ – проверка пройдена;