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

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 519.767.2

Отчет об исследовательском проекте на тему:

Исследование подобластей в семантическом пространстве естественного языка, связаных с описанием глубинных психологических процессов

Выполнил студент:

группы #БПМИ228, 2 курса Мазаев Илья Александрович

Принял руководитель проекта:

Громов Василий Александрович
Доктор физико-математических наук , Профессор
Факультет компьютерных наук НИУ ВШЭ

Москва 2024

Содержание

Аннотация

В данной работе проведён ряд исследований на текстовых данных для анализа семантического пространства естественного языка. С целью выделения наиболее важных признаков, которые могут быть использованы для обнаружения и выделения архетипов в тексте или речи человека. Для представления слов в семантическом пространстве рассматриваются их векторные представления - эмбеддинги, например, полученные с использованием сингулярного разложения (SVD) или других алгоритмов получения эмбеддингов. Для объективного анализа пространства языка применены алгоритмы кластеризации. Для оценки качества кластеризации используются внутренние метрики качества кластеризации. Также в ходе работы был проанализирован малоресурсный язык Аймара, изучены его структурные особенности.

Ключевые слова

Семантическое пространство, язык, архетипы, кластеризация, эмбеддинги, карта языка, язык Аймара

1 Введение

1.1 Описание предметной области

Исследование глубинных психологических процессов имеет давние корни, начиная с работ Карла Юнга и его концепции коллективного бессознательного. Юнг предполагал, что в глубинах психики существуют универсальные символические образы, или архетипы, которые влияют на наше поведение и мышление. В последние годы мы стали свидетелями революционных изменений в обработке естественного языка, появилась возможность проводить более объективный и масштабный анализ семантического пространства текстов, например, для выявления различных аномалий или закономерностей в языке. В данной работе применяется анализ естественного языка с использованием методов представления слов в виде сложных сетей и анализ различных характеристик полученых пространств. Анализ текстов с явным проявлением архетипов проводится на уникальном датасете - текстах, в которых психотерапевтами была промаркирована выраженность архетипов в предложениях. Также в ходе сбора информации для исследования был разобран малоресурсный язык Аймара, был найден носитель языка в том числе благодаря помощи которого, был собран корпус текстов(книги, диалоги, учебники).

1.2 Постановка задачи

Задача исследования заключается в разработке системы выделения подобластей в семантическом пространстве естественного языка, связаных с описанием глубинных психологических процессов(архетипов), применении современных методов анализа текстовых данных на русском языке, методов обработки естественного языка. Для этого необходимо провести предобработку(удаление знаков препинания, лемматизацию) и последующий анализ (сравнение) пространства текстов литературных (нейтральных) и текстов с явным проявлением архетипов. Для работы с текстовыми данными слова преобразуются в массивы чисел (эмбеддинги), затем формируется датасет семантических траекторий - из векторных представлений слов формируются n-граммы. Пространства семантических траекторий полученые для литературных текстов и текстов с проявлением архетипов на русском языке являются предметом исследования в целях выделения наиболее важных признаков и закономерностей, которые могут быть использованы для обнаружения и автоматического выделения архетипов в литературном тексте или речи человека. Математически указаная задача формулируется, как задача кластеризации отдельно всех n-грамм встречающихся в естественном языке и n-грамм

соответствующих проявлению архетипов в языке и сравнении указаных кластеризаций в семантическом пространстве. Для сравнения двух кластеризаций используется взаимоналожение и пересечение кластеров в семантическом пространстве.

2 Обзор литературы

2.1 Методы анализа данных

Основная задача исследования сводится к анализу собраного датасета - семантического пространства, представляющего собой многомерное пространство, в котором взаиморасположены векторы n-грамм. Множество методов анализа, описаных в [2] книге "Динамические процессы на сложных сетях применимы к данному исследованию. Для кластеризации данных использованы методы, описаные в книге [1] "Кластеризация данных". Для оценки качества кластеризации использованы методы, описаные в главе 23. Задача получения эмбеддингов слов решаемая на одном из этапов исследования раскрывается в работе [3], где описаны методы и модели, используемые для анализа семантических связей в текстовых данных, включая применение метода сингулярного разложения (SVD) для получения эмбеддингов. Так же применимы алгоритмы и инструкции для получения других видов эмбеддингов и работы с ними. Например, [12] инструкция по работе с эмбеддиннгами fastText или используемая для получения эмбеддингов CBOW библиотека Gensim [10] для языка программирования Python и документация к ней.

2.2 Глубинные психологические процессы

Основы вопроса исследования архетипов строятся на идеях Карла Густава Юнга, изложенных в его работе [5] "Архетипы и коллективное бессознательное". Юнг исследовал универсальные символические образы, или архетипы, которые лежат в основе коллективного бессознательного и влияют на поведение и мышление человека. В инструкции [14] описана методика подготовки (маркировки) текстов для формирования корпуса текстов с явным проявлением архетипов. Тексты, которые были отобраны психотерапевтами и промаркированы следуя этой инструкции и составляют уникальную часть корпуса текстов с проявлением архетипов.

3 Этапы исследования

3.1 Подготовка текста

Перед началом самого исследования была произведена работа по сбору и дополнению корпусов текстов на русском языке: общего корпуса (тексты литературных произведений) и корпуса текстов с проявлением архетипов. В рамках работы над сбором материалов с проявлением архетипов было разработано консольное приложение с использованием технологии асинхронного распознавания речи SaluteSpeech [7] для создания текстов из аудио материалов, предоставленных психотерапевтами для последующей маркировки ими архетипов в распознанном тексте и расширения корпуса текстов с проявлением архетипов.

Тексты с явным проявлением архетипов - представляет собой уникальный предоставленный психотерапевтами корпус текстов, в котором вручную были промаркированы архетипы. Одним из возможных применений результатов исследования является создание алгоритма для автоматической разметки архетипов в тексте, без участия человека.

После сбора текстов проведена предподготовка корпусов: очистка от знаков препинания и других специальных символов, приведение букв к нижнему регистру, удаление символов переноса строк и табуляции, развёртка текстов в одну строку, проведена лемматизация - приведение слов в тексте к начальной форме, замена некоторых слов(имён, местоимений) на токены. Для лемматизации русского яызка использована библиотека Natasha [9] для языка программирования Python. Обработанные таким образом тексты объединяются в один текстовый файл в котором построчно расположены исходные тексты.

3.2 Эмбеддинги

Подготовленный корпус далее был использован для получения словаря эмбеддингов. С ипользованием различных алгоритмов получения эмбеддингов в соответствие каждому слову ставился массив чисел(вектор). Эти данные используются для построения семантического пространства языка. На этом этапе были задействованы алгоритмы получения эмбеддингов: Word2Vec(CBOW, Skip-gram), SVD. В исследовании для построения пространства использовались эмбеддинги размерности 8, 20 и 100.

Для получения эмбеддингов типа CBOW(Continuous Bag of Words) используется модель Word2Vec из библиотеки Gensim [10] для Python. Модель включает два основных подхода: CBOW и Skip-gram. Суть CBOW (Continuous Bag of Words) заключается в прогнозировании слова на основе его контекста. Текстовый ввод рассматривается, как последова-

тельность слов, по которой проходит скользящее окно фиксированного размера, содержание которого представляется как контекст слова в центре этого окна. Слова, попавшие в окно используются, как входной слой однослойной нейросети, данные передаются на скрытый слой, где слова контекста агрегируются, а затем на выходном слое формируется вероятностное распределение предсказания целевого(центрального) слова. Результатом является матрица весов скрытого слоя, которая и представляет собой набор эмбеддингов слов, отображающих семантическую близость слов.

Метод получения эмбеддингов методом SVD заключается в построении TF-IDF матрицы и последующего её SVD разложения в целях снижения размерности. TF-IDF отражает важность слова w в тексте d относительно корпуса D через произведение частоты слова TF и обратной частоты текста IDF:

$$\text{TF-IDF}(w, d, D) = \text{TF}(w, d) \times \text{IDF}(w, D),$$

где

$$\text{TF}(w, d) = \frac{n_{w,d}}{\sum_{w' \in d} n_{w',d}}, \quad \text{IDF}(w, D) = \log \frac{|D|}{|\{d \in D : w \in d\}|}.$$

$n_{w,d}$ – число вхождений слова w в текст d, $\sum_{w' \in d} n_{w',d}$ – общее число слов в тексте.

Матрица TF-IDF X размера $n \times m$ (где n — уникальные слова, m — тексты) разреженная. SVD разложение приближает X к меньшей матрице X_r.

$$X \approx X_r = U \Sigma V^T,$$

с размерностями U (n × r) и V (m × r), где $r < \min(n, m)$. Для вычисления SVD разложения используется метод, предложеный Gene H. Golub и William Kahan [4].

Полученное пространство называется LSM (Latent Semantic Mapping [3]). Матрица X_r размера $n \times r$ содержит векторы слов, где каждая строка соответствует вектору слова(эмбеддингу) в пространстве размерности r.

3.3 Пространство языка

Этап построения и изучения пространства семантического пространства языка при рассмотрении в отдельности для корпусов: общего (с литературными произведениями), текстов с проявлением архетипов, и объединённого - корпус представляющий собой конкатенацию общего корпуса и корпуса с проявлением архетипов. Каждый из этих корпусов в отдель-

ности использовался для получения эмбеддингов слов, и построения пространства семантических траекторий языка из n-грамм - подпоследовательностей слов длины n, встреченных в тексте. Для этого строится множество n-грамм, затем каждой n-грамме в соответствие ставится вектор, сформированый из полученых ранее эмбеддингов слов, входящих в эту n-грамму. Именно эти пространства являются главным предметом исследования. При построении пространств не обязательно использовать эмбеддинги, и множество n-грамм построеные на одном корпусе. Во некоторых экспериментах используется, например множество биграмм встреченых в подмножестве общего корпуса, но эмбеддинги полученные по объединённому корпусу.

3.4 Визуализация

Метод t-SNE (t-distributed Stochastic Neighbor Embedding) представляет собой метод для визуализации многомерных данных путём сокращения их размерности с сохранением структуры соседства между точками данных. Этот метод был представлен Laurens van der Maaten и Geoffrey Hinton [6] в 2008 году.

Процесс t-SNE начинается с вычисления условных вероятностей парных подобий между точками в многомерном пространстве таким образом, что схожие объекты имеют высокую вероятность быть выбранными, а различные объекты - низкую. Для этого используется симметризированная версия условной вероятности, где вероятность встречаемости пары точек пропорциональна их сходству, которое обычно измеряется с помощью Гауссова распределения вокруг каждой точки.

Далее, в пространстве уменьшенной размерности создаётся новое распределение точек, для которого также вычисляются парные условные вероятности, но уже с использованием распределения Стьюдента. Это распределение выбирается из-за его более тяжелых хвостов по сравнению с нормальным распределением, что позволяет лучше разделять объекты, которые в многомерном пространстве были далеки друг от друга.

Основной принцип t-SNE — минимизировать расхождение между двумя распределениями, которое измеряется с помощью расхождения Кульбака-Лейблера(KL divergence). Одна из ключевых особенностей t-SNE заключается в том, что он способен эффективно обнаруживать локальные структуры в данных и выявлять глобальные группировки и формирование кластеров.

Для рассчёта визуализации в данной работе применялась библиотека sklearn [11] для Python.

Применение метода 3.1 t-SNE для визуализации 200-мерного пространства биграмм объединенного корпуса. Для построения биграмм в этом примере используются эмбеддинги SVD размерности 100 полученые по объединенному корпусу.

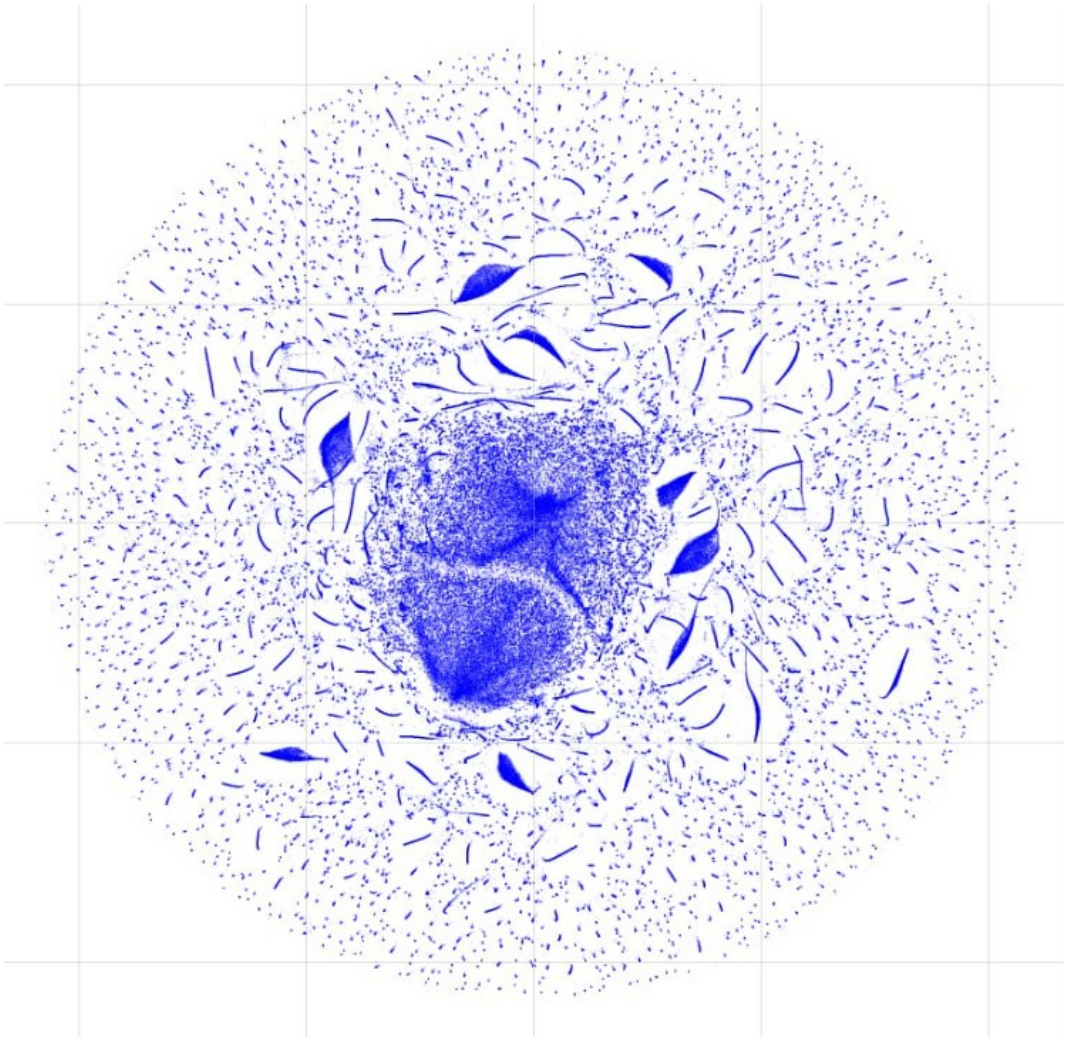

Рис. 3.1: t-SNE визуализация пространства биграмм объединенного корпуса, эмбеддинги SVD

На этапе построения пространства было выделено подмножество n-грамм, которые встречаются исключительно в текстах с проявлением архетипов. В следущем примере 3.2 визуализации 40-мерного пространства биграмм объединенного корпуса это подмножество точек отмечено красным. {Множество биграмм, отмеченых красным} = {множество биграмм в объединенном корпусе} \ {биграммы в корпусе общих текстов}.

Аналогичный пример с использованием эмбеддингов SVD 3.3 с выделением подпоследовательностей, уникальных для текстов с проявлением архетипов. Пространство размерности 200. В обоих примерах используется пространство биграмм в объединенном корпусе.

Заметим, что красные точки расположены близко друг к другу в общих кластерах, иногда составляют отдельные более плтоные кластеры меньшего размера, состоящие в основном из биграмм, уникальных для в текстов с проявлением архетипов.

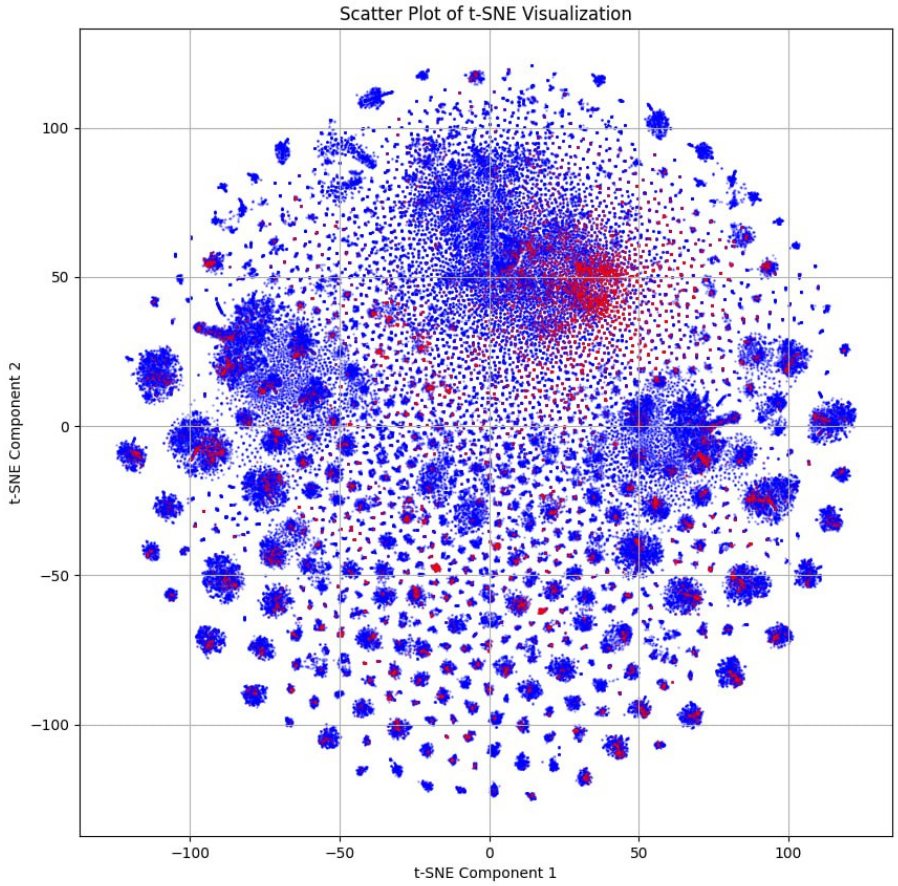

Рис. 3.2: t-SNE визуализация пространства размерности 200 биграмм объединенного корпуса, эмбеддинги CBOW по объединенному корпусу

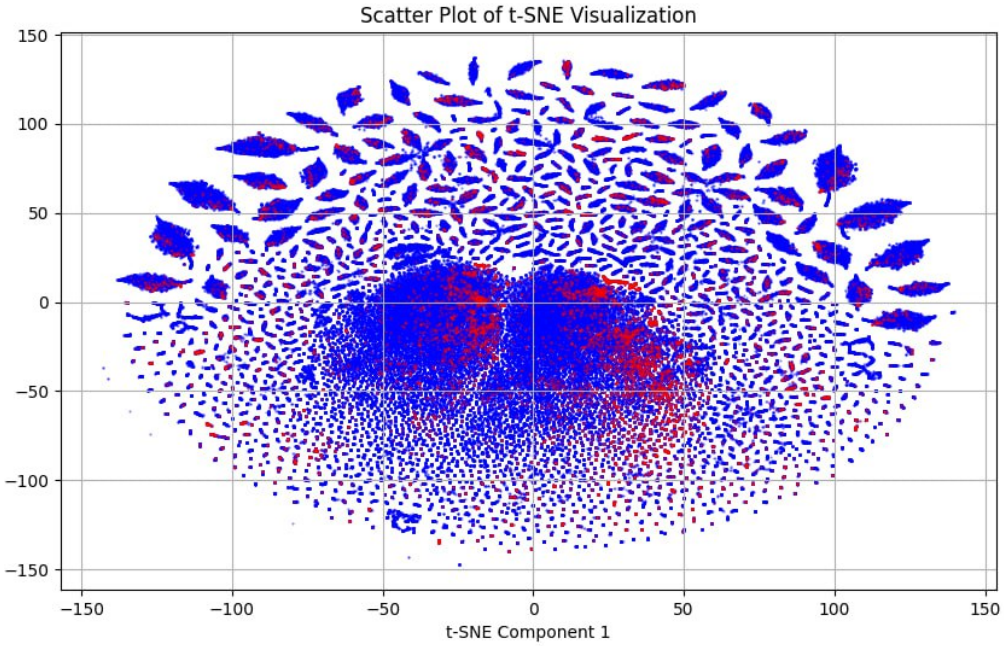

Рис. 3.3: t-SNE визуализация пространства размерности 200 биграмм объединенного корпуса, эмбеддинги SVD по объединенному корпусу

Ещё более явно видны плотные области красных точек при работе с пространством всех биграмм, которые встречаются в текстах с проявлением архетипов. Визуализации для эмбеддингов CBOW 3.4 и SVD 3.5.

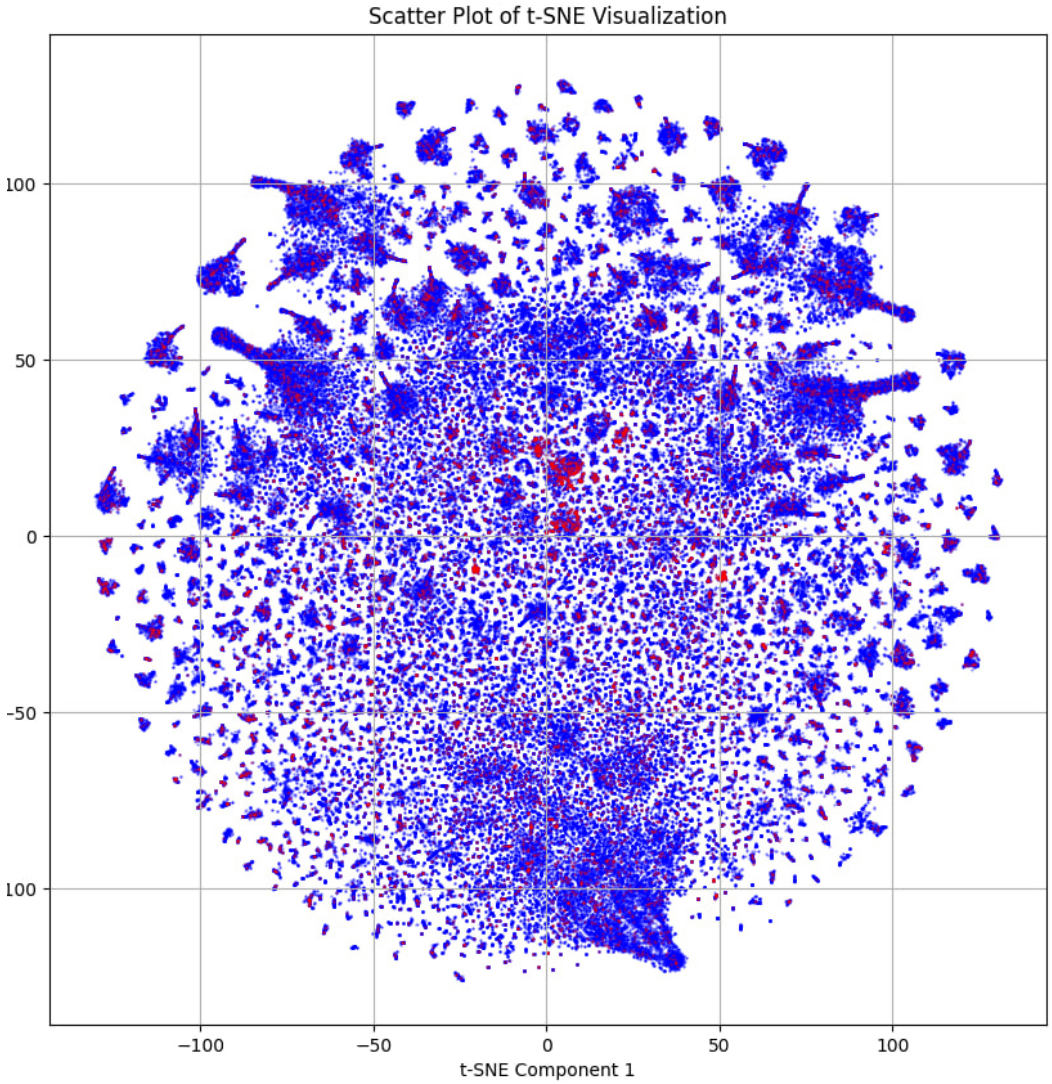

Рис. 3.4: t-SNE визуализация пространства размерности 40 биграмм архетипичного корпуса, эмбеддинги CBOW по объединенному корпусу

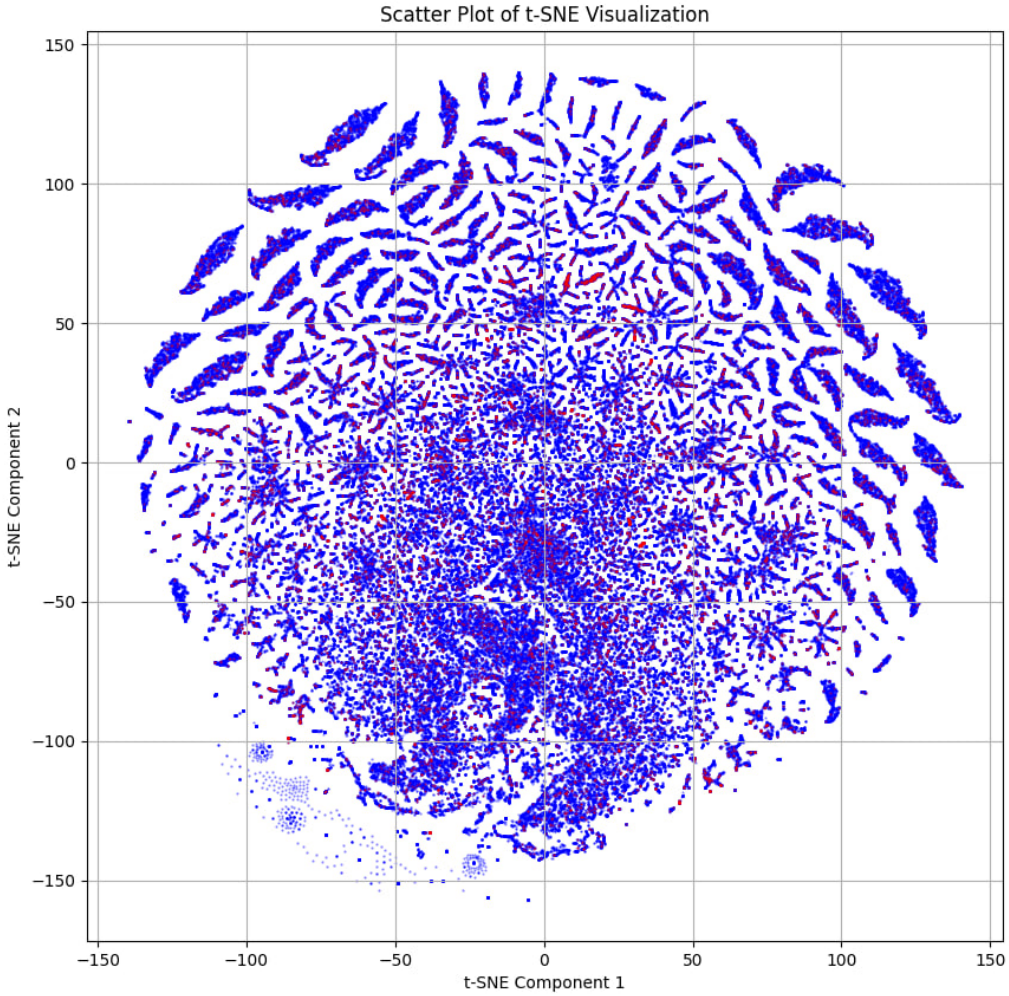

Рис. 3.5: t-SNE визуализация пространства размерности 16 биграмм архетипичного корпуса, эмбеддинги SVD по объединенному корпусу

В ходе работы с t-SNE была встречена некорректное поведени алгоритма при попытках визуализировать более 10**7 точек. Такое количество элементов пространства возникает, например, при визуализации биграмм всего общего корпуса литературных произведений Русского языка. Поэтому для всех удачных визуализаций n-граммы строились по выборке текстов методом выбора каждого k-го (k=25, 50) текста из корпуса и исключения остальных из рассмотрения в целях уменьшения общего количества n-грамм.

Пример неудачных визуализаций 3.6:

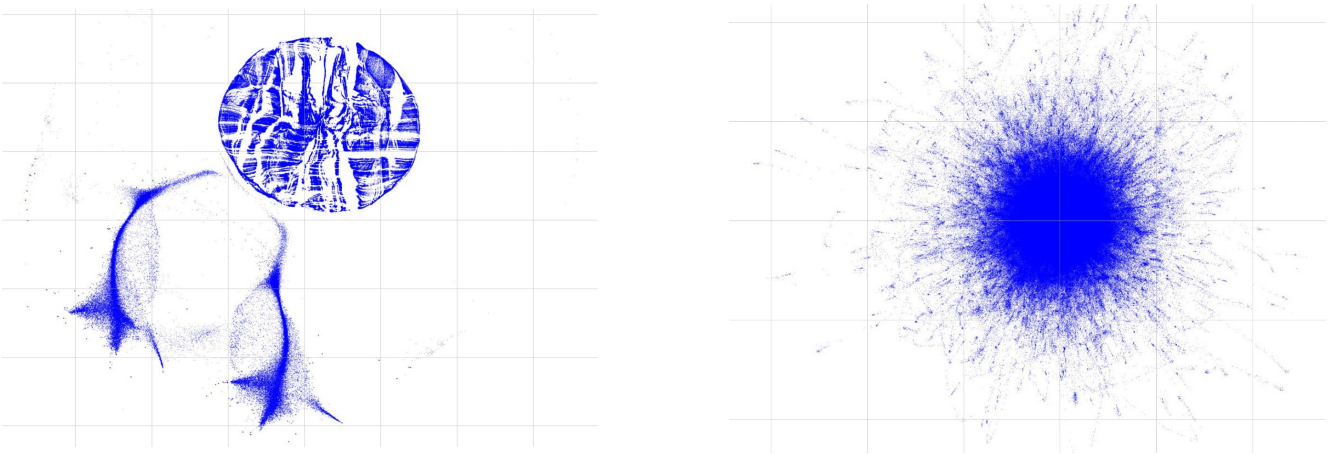

Рис. 3.6: Неудачные визуализации t-SNE для запредельного числа элементов пространства

3.5 Кластеризация

В целях выделения наиболее важных признаков и закономерностей, которые могут быть использованы для обнаружения подпоследовательностей слов в текстах проведён необходимый объективный анализ пространства семантических тректорий языка, поиск и выделение кластеров с наибольшим соотношением "красных точек" и других закономерностей. Для кластеризации использован алгоритм кластеризации Уишарта [8]. Алгоритм Уишарта наилучшим образом подходит для кластеризации сложноструктурированного семантического пространства языка за счёт его способности обнаруживать интенсивно сгруппированные кластеры без необходимости предварительного определения числа кластеров. Для эффективного использования ресурсов суперкомпьютера алгоритм был адаптирован для работы в многоядерном режиме.

Расчет t-SNE визуализации пространства и кластеризация производились на суперкомпьютере "cHARISMa"НИУ ВШЭ [13].

Для определения качества кластеризации и подбора обеспечивающих наиболее качественную кластеризацию гиперпараметров алгоритма Уишарта (минимального количества точек в кластере и статистической значимости) используются внутренние метрики качества кластеризации (см. таблица 1 в приложении). В исследовании применены Calinski-Harabasz index (CH) и Silhouette index (S). Calinski-Harabasz index (CH) измеряет степень дисперсии внутри каждого кластера по сравнению с дисперсией между кластерами, что позволяет оценить насколько четко выделены кластеры относительно друг друга. Silhouette index (S) оценивает схожеть элементов внутри кластеров, вычисляя среднее значение характеристики, которая показывает, насколько близко каждая точка находится к точкам своего кластера по

сравнению с точками из соседних кластеров, тем самым указывая на точность распределения точек по кластерам.

Наилучший результат кластеризации данных биграмм объединенного корпуса достиг значений 4576.571 по Calinski-Harabasz index и -0.1538 Shilouette index, что говорит об удовлетворительном, но пока недостаточном уровне точности кластеризации. Использованы эмбеддинги CBOW размерности 20, значения гиперпараметров: wishart neighbors = 1000 significance level = 50000. Был выделен кластер, содержащий 12207 "красных точек" из их общего количества, которое составляет 14180 - биграмм, встречающихся исключительно в текстах с проявлением архетипов. Гиперпатаметры для алгоритма Уишарта были подобраны эмпирическим методом, путём перебора значений гиперпараметров по сетке и улучшения результатов выбраных метрик кластеризации.

4 Заключение

В ходе исследования успешно применены методы обработки и подготовки корпусов языка для анализа. Разработано приложение для распознавания аудиофайлов и дополнения корпуса текстов с явным проявлением архетипов. Успешно применены алгоритмы построения эмбеддингов различных типов и любых размерностей. Эти данные служат фундаментом для дальнейшего исследования пространства естественного языка.

Дальнейшее применение метода t-SNE для визуализации семантического пространства позволило идентифицировать ключевые особенности областей пространства, связанных с описанием глубинных психологических процессов, несмотря на сложности с обработкой больших объемов данных.

Использование алгоритма Уишарта для кластеризации показало его потенциальную применимость для идентификации семантических траекторий в текстах, поиск кластеров, относящихся к проявлению архетипов в тексте. Однако, достигнутая точность кластеризации оказалась недостаточной для разработки алгоритма для автоматической разметки архетипов в тексте. Возможно использование большей длины n-грамм или других методов кластеризации, дополнительных внутренних метрик качества кластеризации для более точного подбора значений гиперпараметров алгоритмов кластеризации. В целях более точной кластеризации, выделе

Рассматривается возможность применения показавших свою эффективность методов исследования семантического пространства для продолжения исследования языка Аймара. Следующим этапом исследования может быть разработка лемматизатора для языка, его отладка с помощью носителя языка, дополнение корпуса литературы и исследования семантического пространства и особенностей языка.

5 Список литературы

[1] Charu C. Aggarwal и Chandan K. Reddy. “DATA CLUSTERING Algorithms and Applications”. В: Taylor Francis Group, LLC, 2014. Гл. 23.

[2] Alain Barrat, Marc Barthélemy и Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[3] Jerome R. Bellegarda. *Latent Semantic Mapping: Principles Applications*. Morgan Claypool, 2007.

[4] Gene H. Golub и William Kahan. *Calculating the Singular Values and Pseudo-Inverse of a Matrix*. 1965.

[5] Carl Gustav Jung. *Archetypes and the Collective Unconscious*. Princeton University Press, 1981.

[6] Laurens van der Maaten и Geoffrey Hinton. *Visualizing Data using t-SNE*. Journal of Machine Learning Research, 2008.

[7] *SaluteSpeech Асинхронное распознавание речи*. URL: https://developers.sber.ru/docs/ru/salutespeech/recognition/recognition-async (дата обр. 24.04.2024).

[8] David Wishart. “Numerical Classification Method for deriving Natural Classes”. В: *Nature* 221 (1969), с. 97—98.

[9] *Библиотека для лемматизации(токенизации) русского языка natasha*. URL: https://github.com/natasha/natasha (дата обр. 24.04.2024).

[10] *Документация библиотеки Gensim для gensim.models.Word2Vec*. URL: https://radimrehurek.com/gensim/models/word2vec.html (дата обр. 24.04.2024).

[11] *Документация библиотеки sklearn для sklearn.manifold.TSNE*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (дата обр. 25.04.2024).

[12] *Инструкция по работе с эмбеддингами fastText*. URL: https://fasttext.cc/docs/en/crawl-vectors.htmlhttps://fasttext.cc/docs/en/crawl-vectors.html (дата обр. 14.02.2024).

[13] *Суперкомпьютер cHARISMa НИУ ВШЭ*. URL: https://hpc.hse.ru/hardware/hpc-cluster (дата обр. 25.04.2024).

[14] Ирина Алексеевна Чеглова. *Инструкция по маркировке текстов сессий с использованием методики Ирины Алексеевны Чегловой*. URL: https://docs.google.com/document/d/1Jot_ivxfl0Qut4HYoTyk0fJN_iI9zd9Qy4ksKyG63bg/edit (дата обр. 13.02.2024).

6 Приложения

6.1 Внутренние метрики качества кластеризации

Источник : “DATA CLUSTERING Algorithms and Applications” глава 23 [1].

	Метрика	Определение
1	RMSSTD[1]	$\left\{\sum_i \sum_{x\in C_i} \|x-c_i\|^2 / [P\sum_i (n_i-1)]\right\}^{\frac{1}{2}}$
2	R-squared (RS)	$\left(\sum_{x\in D}\|x-c\|^2 - \sum_i \sum_{x\in C_i}\|x-c_i\|^2\right) / \sum_{x\in D}\|x-c\|^2$
3	Modified Hubert Γ statistic (Γ)	$\frac{2}{n(n-1)} \sum_{x\in D}\sum_{y\in D} d(x,y) d_{x\in C_i, y\in C_j}(c_i,c_j)$
4	Calinski-Harabasz index (CH)	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x\in C_i} d^2(x,c_i)/(n-NC)}$
5	I index (I)	$\left(\frac{1}{NC}\cdot \frac{\sum_{x\in D} d(x,c)}{\sum_i \sum_{x\in C_i} d(x,c_i)} \cdot \max_{i,j} d(c_i,c_j)\right)^p$
6	Dunn’s indices (D)	$\min_i \left\{\min_j \left(\frac{\min_{x\in C_i y\in C_j} d(x,y)}{\max_k \left\{\max_{x,y\in C_k} d(x,y)\right\}}\right)\right\}$
7	Silhouette index (S)	$\frac{1}{NC}\sum_i \left\{\frac{1}{n_i}\sum_{x\in C_i} \frac{b(x)-a(x)}{\max[b(x),a(x)}\right\}$ $a(x) = \frac{1}{n_i-1}\sum_{y\in C_i, y\neq x} d(x,y), b(x) = \min_{j,j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j} d(x,y)\right]$
8	Davies-Bouldin index (DB)	$\frac{1}{NC}\sum_i \max_{j,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)\right] / d(c_i,c_j)\right\}$
9	Xie-Beni index (XB)	$\left[\sum_i \sum_{x\in C_i} d^2(x,c_i)\right] / [n\cdot \min_{i,j\neq i} d^2(c_i,c_j)]$
10	SD validity index (SD)	$\mathrm{Dis}(NC_{\max})\,\mathrm{Scat}(NC) + \mathrm{Dis}(NC)$ $\mathrm{Scat}(NC) = \frac{1}{NC}\sum_i \|\sigma(C_i)\| / \|\sigma(D)\|$ $\mathrm{Dis}(NC) = \frac{\max_{i,j} d(c_i,c_j)}{\min_{i,j} d(c_i,c_j)} \sum_i \left(\sum_j d(c_i,c_j)\right)^{-1}$
11	$\mathrm{S}_D bwvalidityindex(S\_Dbw)$	$\mathrm{Scat}\ (NC) + \mathrm{Dens\_bw}(NC)$ $\mathrm{Dens\_bw}(NC) = \frac{1}{NC(NC-1)}\sum_i \left[\sum_{j,j\neq i} \frac{\sum_{x\in C_i\cup c_j} f(x,u_{ij})}{\max\left\{\sum_{x\in C_i} f(x,c_i), \Sigma_{r\in C_j} f(x,c_j)\right\}}\right]$
12	$CVNN$[2] index	$\mathrm{Sep}(NC,k)/\max_{NC}\mathrm{Sep}(NC,k) + \mathrm{Com}(NC)/\max_{NC}\mathrm{Com}(NC)$ $\mathrm{Com}(NC) = \sum_i \left[\frac{2}{n_i\cdot(n_i-1)}\sum_{x,y\in C_i} d(x,y)\right]$ $\mathrm{Sep}(NC,k) = \max_i \left(\frac{1}{n_i}\sum_{j=1,2,\ldots,n_i} \frac{q_j}{k}\right)$

Таблица 1

Примечание: D - набор данных; n - количество объектов в D; c - центр D; P - количество атрибутов D; NC - количество кластеров; C_i - i-ый кластер; n_i - количество объектов в C_i; c_i - центр C_i; k - количество ближайших соседей; q_j - количество ближайших соседей j-го объекта кластера C_i, которые не входят в кластер C_i; $\sigma(C_i)$ - вектор дисперсии кластера C_i; $d(x,y)$ - расстояние между x и y; $|X_i| = \left(X_i^T \cdot X_i\right)^{1/2}$. [1] RMSSTD: Root-mean-square standard deviation. [2] $CVNN$: Clustering Validation index based on Nearest Neighbors.

6.2 Эффективность различных внутренних метрик качества кластеризации

Источник : “DATA CLUSTERING Algorithms and Applications” глава 23 [1].

Метрика	Монотонность	Шум	Плотность	Субкластеры	Ассимпт. распр	Произвольн. форма
RMSSTD	×	-	-	-	-	-
RS	×	-	-	-	-	-
Γ	×	-	-	-	-	-
CH		×			×	×
I			×			×
D		×		×		×
S				×		×
DB^{**}				×		×
SD				×		×
S_Dbw						×
XB^{**}				×		×
$CVNN$						

Таблица 2

В таблице 2 обобщены свойства различных внутренних метрик качества кластеризации в различных задачах и при особенностях в структуре данных, что может быть полезно при выборе метрики. «–» означает, что свойство не проверено, а «×» означает, что ситуация не может быть обработана.