**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ**

**им. А.И. ГЕРЦЕНА»**



Институт информационных технологий и технологического образования

Реферат на тему: «Машинная обработка естественного языка»

Дисциплина: «Логика и методология науки»

студент группы 1А \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Момот Д.М.

профессор каф. ИС \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ д.т.н., проф. Фомин В.В.

Санкт-Петербург

2020

Оглавление

[1. NLP: общая характеристика 3](#_Toc53240055)

[2. NLP: приложения 5](#_Toc53240056)

[3. NLP: краткий обзор состояния и перспектив 7](#_Toc53240057)

[4. NLP: литература по теме 10](#_Toc53240058)

[Список использованной литературы 12](#_Toc53240059)

## NLP: общая характеристика

Обработка естественного языка (NLP) – это дисциплина на стыке искусственного интеллекта и математической лингвистики. Она изучает проблемы компьютерного анализа и синтеза текстов на естественном языке. При этом структуры естественного языка рассматриваются как объекты, которые могут быть смоделированы [1].

При моделировании могут использоваться как четкое моделирование (методами дискретной математики [8] и линейной алгебры [9]), так и нечеткое (методами нечеткой математики, в том числи нейросетями [2]) [1]. В наиболее простых случаях модель строится на основе одного поданного на вход текста [6][8], но чаще всего необходимо также учитывать контекст [7].

Структуры языка принято разделять на несколько взаимосвязанных иерархических уровней. Каждый уровень имеет свои правила сочетания составных единиц. На каждом уровне возникают свои задачи. Они описаны в таблице 1 (содержит данные из [3, 4, 5, 10]).

Таблица 1. Уровни естественного языка, их единицы и задачи.

|  |  |  |
| --- | --- | --- |
| Уровень | Составные части | Задачи |
| Фонологический | Фонемы[[1]](#footnote-1) / Графемы[[2]](#footnote-2) |  |
| Морфологический  (словоформа -> лемма, теги) | Морфемы  Словоформы | Лемматизация  Стемминг  Полный морфоанализ  Разрешение морфологической омонимии |
| Лексический | Лексемы (из лексикона) |  |
| Синтаксический (предложение + результат морфоанализа -> АСД) | Словосочетания  Предложения  Сверхфразовые единства[[3]](#footnote-3) | Токенизация  Векторизация  Парсинг (построение АСД)  Извлечение названий, терминов |
| Семантический (АСД -> семантическая структура) | Семы | Определение семантики слов и словосочетаний, разрешение неоднозначности  Установление семантических отношений между словами и словосочетаниями  Построение модели представления смысла  Локальный семантический анализ  Тематическое моделирование  Анализ тональности (эмоциональной окраски) |
| Дискурсивный | Структуры связного текста | Суммаризация, реферирование, аннотирование  Разрешение анафор[[4]](#footnote-4), кореференций[[5]](#footnote-5), извлечение отношений  Дедубликация |

Лемматизация – приведение слова к начальной форме.

Стемминг – получение основы слова.

Полный морфоанализ – извлечение леммы и морфологических характеристик (тегов).

Разрешение морфологической омонимии. Выполняется на основе применения лингвистических правил, например, анализа предлогов, либо с помощью машинного обучения.

Токенизация – деление на слова.

Векторизация – представление текста как вектора.

Парсинг – построение абстрактного синтаксического дерева предложения, АСД. АСД может принадлежать к одному из двух типов – дерево зависимостей/подчинения и дерево составляющих – для языков с разным синтаксисом.

Дедубликация – устранение повторяющихся документов, например, с помощью хэшей.

Тематическое моделирование – выделение темы, классификация текстов.

Для всех уровней характерны:

* Стандартная и нестандартная сочетаемость (боксер может быть и крепким, и сильным, а чай – только крепким)
* Неоднозначность
  + Полисемия – многозначность лексемы
  + Синонимия – слов, предлогов, приставок, …
  + Омонимия (в отличие от полисемии – нет смысловой связи)
    - Лексическая (совпадение слов)
    - Морфологическая (совпадение форм)
    - Лексико-морфологическая (совпадение форм слов)
    - Синтаксическая (неоднозначность синтаксической структуры)

## NLP: приложения

Существуют следующие прикладные задачи, решаемые с помощью NLP. В сносках приводятся ссылки на примеры публикаций по конкретным приложениям.

* Машинный перевод
  + На основе правил (rule-based)
  + На основе статистики (statistical-based)
  + Нейронный машинный перевод (neural machine translation)
* Информационный поиск [[6]](#footnote-6)
* Реферирование и аннотирование текстов [[7]](#footnote-7) [[8]](#footnote-8) [[9]](#footnote-9) [[10]](#footnote-10)
* Автоматизация генерации и редактирования текстов
* Формирование ответов на вопросы, естественно-языковой интерфейс [[11]](#footnote-11) [[12]](#footnote-12), организация диалога [[13]](#footnote-13)
* Распознавание и синтез звучащей речи, рукописного текста [[14]](#footnote-14)
* Text Mining
  + Извлечение информации из текста
    - Разрешение анафоры и кореференций (слов, словосочетаний, предложений, относящихся к одному объекту) [[15]](#footnote-15) [[16]](#footnote-16) [[17]](#footnote-17) [[18]](#footnote-18) [[19]](#footnote-19)
  + Классификация и кластеризация текстов [[20]](#footnote-20)
  + Извлечение имен собственных [[21]](#footnote-21) [[22]](#footnote-22) [[23]](#footnote-23) [[24]](#footnote-24) [[25]](#footnote-25), терминов и ключевых слов [[26]](#footnote-26) [[27]](#footnote-27) [[28]](#footnote-28) [[29]](#footnote-29) [[30]](#footnote-30) [[31]](#footnote-31) [[32]](#footnote-32) [[33]](#footnote-33) [[34]](#footnote-34) [[35]](#footnote-35)
  + Анализ мнений и оценка тональности текстов

## NLP: краткий обзор состояния и перспектив

Ниже приводится обзор современных (на октябрь 2019) достижений в этих областях по материалам альманаха МФТИ [11, 12].

Общие тренды и прогнозы.

* Будет развиваться end-to-end подход [см https://habr.com/ru/company/ru\_mts/blog/451008/], состоящий в переходе от системы белых или черных ящиков к одному черному ящику. Это позволяет повысить точность и скорость обучения модели, одновременно ухудшив понимание структуры происходящего внутри. Имеются в виду модели, занимающиеся NLP и генерацией текстов (уже сейчас есть GPT2 позволяет генерировать достаточно длинные статьи по заданной теме).
* Прогнозируется дальнейшее улучшение качества распознавания речи. В качестве пока что незадействованных ресурсов для этого предлагается определение пола и возраста говорящего, эмоциональной окраски речи и характера обстановки. Синтезированная речь будет неотличима от человеческой, можно будет синтезировать речь любого человека.
* Диалоговые сервисы будут больше учитывать контекст, а также смогут направлять разговор в нужное русло. Более глубокий учет контекста позволит также повысить качество извлечения информации из текстов.
* Прогнозируется стандартизация и повышение кроссплатформенности решений, возможность автоматического сохранения и экспорта контекста.
* Рост значения методов машинного обучения на небольших объемах сырых данных (transfer learning, knowledge transfer). Это означает повышение «спроса» на GAN (generative advertisal networks) для генерации недостающих данных.
* Появление запроса на энергоэффективные, возможно даже легковесные, решения. Например, применение разреженных матриц[[36]](#footnote-36), а также адаптация сложности моделей к объему обучающих данных.
* На рынке прогнозируется массовое внедрение голосовых интерфейсов и голосовых роботов-ассистентов. Коммерчески важной является также задача разрешения запросов к неструктурированным данным, которая будет решаться за счет совершения методов извлечения фактов и автореферирования. Причем среди крупных компаний наблюдается тренд решения задач NLP собственными силами, аутсорсинг не имеет распространения.

Распознавание и синтез звучащей речи.

Распознавание речи в данный момент является весьма актуальной темой ввиду понятного коммерческого применения (Cortana, Google Assistant, Siri, Алиса, …). При отсутствии шума и хорошего качества звука нейросети уже работают не хуже человека. Однако с распознаванием речи, записанной на удаленный микрофон, с шумом или при беседе нескольких лиц, нейросети сильно уступают человеку.

Синтез речи. Основным подходом является развертывание системы TTS (text to speech), включающей следующие компоненты, реализуемые с помощью нейросетей:

1. Транскриптор (текст -> последовательность фонем)
2. Преобразователь (последовательность фонем -> последовательность параметров сигнала)
3. Вокодер (параметры -> звуковая волна).

Нейронные системы синтеза речи позволили значительно увеличить плавность и натуральность звучания.

Машинный перевод. Проведено сравнение разных систем и сделан вывод, что:

* единого лидера нет — в зависимости от лингвистической пары и предметной области качество решений варьируется
* доступное качество со временем растет
* идет массовое внедрение доменно-адаптивного (основанного на корпусе текстов по определенной тематике) машинного перевода

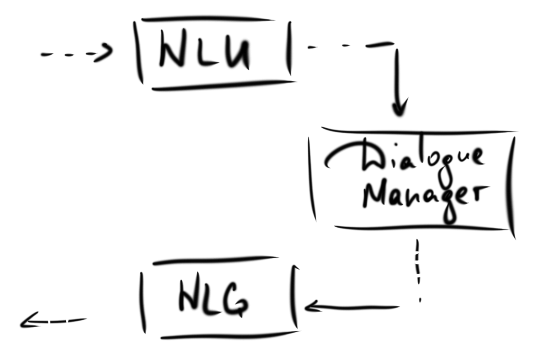
Нейронные системы синтеза речи позволили значительно увеличить плавность и натуральность звучания.

Генерация текста. В порядке появления:

* Нейросети LSTM
* Нейросети Generative Pretraining Transformer
* Нейросети GPT2

В настоящее время все модели генерируют тексты слово за словом, что не позволяет рассчитывать на интересный сюжет.

Диалоговые системы.

По способу построения:

* На основе правил
* На основе данных

Состоят из 3 частей:

* Natural language understanding – модуль понимания естественного языка
* Dialogue Manager – ядро диалоговой системы
* Natural language generator – генератор естественного языка

Метрики оценки:

* Показатели пользователей – количество юзеров, активные пользователи, вовлечение
* Метрики сообщений – всего сообщений, сообщений за разговор, процент обработанных
* Метрики бота – уровень удержания, время достижения цели общения, количество неудач

NLP, поиск и извлечение информации из текстов

Основным инструментом является представление слов как плотных векторов. Для этого используется технология Word2Vec, она представляет собой набор моделей для анализа семантики языка. Позднее стали применяться модели с использованием рекуррентных нейросетей, а именно LSTM и GRU (gated recurrent unit). Еще позднее добавились «механизм внимания» и большие предобученные модели.

## NLP: литература по теме

1 Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

<https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf>

https://istina.msu.ru/download/88770435/1gFEa3:QwqAnNiO\_3GPe\_3Erw4K2LXdtfM/

2 <https://cyberleninka.ru/article/n/sostoyanie-i-perspektivy-razvitiya-metodov-generatsii-i-opredeleniya-form-slov-i-obrabotki-kolichestvennyh-chislitelnyh-estestvennyh>

3 <https://medium.com/@mad_81298/https-medium-com-mad-81298-state-of-deep-learning-for-natural-language-processing-9b94dbba8d18>

4 <https://cyberleninka.ru/article/n/prozrachnoe-glubokoe-obuchenie-na-osnove-veroyatnostnyh-formalnyh-ponyatiy-v-zadache-obrabotki-estestvennogo-yazyka>

5 в медицине <https://www.sworld.com.ua/konfer30/661.pdf>

6 transfer learning

7 <https://www.hse.ru/data/2018/06/05/1150100121/program-2211378881-e9jDwINTe2.pdf>

8 диалоговые системы <https://www.elibrary.ru/item.asp?id=39549470>

9 учебник по NLP <https://coderlessons.com/tutorials/akademicheskii/obrabotka-estestvennogo-iazyka/obrabotka-estestvennogo-iazyka-kratkoe-rukovodstvo>

10 обзор ИИ <http://digital-economy.ru/images/easyblog_articles/530/DE-2018-03-05.pdf>

11 платный курс от яндекса <https://academy.yandex.ru/posts/sovremennye-metody-iskusstvennogo-intellekta>

12 книга Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit Steven Bird, Ewan Klein, and Edward Loper <http://www.nltk.org/book/>

13 книга SPEECH and LANGUAGE PROCESSING An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition by Daniel Jurafsky and James H. Martin <https://www.cs.colorado.edu/~martin/slp2.html>

14 обзор <http://vestnik.rsreu.ru/images/archive/2016/1-55/11_%D0%9F%D1%80%D1%83%D1%86%D0%BA%D0%BE%D0%B2%20%D0%90.%D0%92.pdf>

15 материалы большой конференции <http://www.dialog-21.ru/media/2767/dialogue_2008.pdf>

16 <http://khpi-iip.mipk.kharkiv.edu/library/ai/conspai/10.html>

17 <https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D1%81%D0%BB%D0%B8%D1%82%D0%BD%D0%BE%D0%B9_%D1%80%D0%B5%D1%87%D0%B8_%D0%BD%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%BC_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B5:_%E2%80%A6%D0%B4%D1%80%D1%83%D0%B3%D0%BE%D0%BC%D1%83_%D0%BA%D0%B0%D0%BA_%D0%BF%D0%BE%D0%BD%D1%8F%D1%82%D1%8C_%D1%82%D0%B5%D0%B1%D1%8F>

18 deep learn in NLP <http://top1top.ru/2019/02/25/%D0%BE%D0%B1%D0%B7%D0%BE%D1%80-%D0%B8%D1%81%D1%81%D0%BB%D0%B5%D0%B4%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B9-%D0%B2-%D0%BE%D0%B1%D0%BB%D0%B0%D1%81%D1%82%D0%B8-%D0%B3%D0%BB%D1%83%D0%B1%D0%BE%D0%BA%D0%BE/>

19 для чайников <https://habr.com/ru/company/oleg-bunin/blog/352614/>

## Список использованной литературы

[1] Заболеева-Зотова, А.В. Использование естественного языка при автоматизации слабо структурируемых процессов в проектировании технических, программных и информационных систем. [Текст]: дис. ... доктор техн. наук: 05.13.01: защищена 23.12.04 / Заболеева-Зотова Алла Викторовна. - Волгоград, 2004. - 378 с. - Введение: с.1.

[2] Гольдберг Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг; пер. с англ. А. Слинкин. – ДМК-Пресс, 2019. – 282 с.

[3] Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 269 с.

[4] Смазневич И. NLP: как стать специалистом по обработке естественного языка [Электронный ресурс] // Tproger — издание о разработке и обо всём, что с ней связано. URL: https://tproger.ru/blogs/nlp-professional-howto/ (дата обращения: 29.09.2020).

[5] Обработка естественного языка [Электронный ресурс] // Викиконспекты. URL: http://neerc.ifmo.ru/wiki/index.php?title=Обработка\_естественного\_языка (дата обращения: 29.09.2020).

[6] Adesina A., Azeez N., Abidoye P. Context-Aware Stemming Algorithm for Semantically Related Root Words. African Journal of Computing and ICTs, 2012, vol. 5.

[7] Mohler T. Context Analysis in NLP: Why It’s Valuable and How It’s Done [Электронный ресурс] // Lexalytics. Tell powerful stories from complex text data. URL: https://www.lexalytics.com/lexablog/context-analysis-nlp/ (дата обращения: 29.09.2020).

[8] Mou L., Zhou H., Li L. Discreteness in Neural Natural Language Processing [Электронный ресурс] // URL: https://lili-mou.github.io/resource/emnlp19-1.pdf (дата обращения: 29.09.2020).

[9] Brownlee J. 10 Examples of Linear Algebra in Machine Learning [Электронный ресурс] // Machine Learning Mastery: Making Developers Awesome at Machine Learning. URL: https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/ (дата обращения: 29.09.2020).

[10] Бессмертный, И. А.  Интеллектуальные системы: учебник и практикум для академического бакалавриата / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. — Москва: Издательство Юрайт, 2019. — 243 с. — ISBN 978-5-534-01042-8. // Студми. Учебные материалы для студентов. — URL: https://studme.org/139940/informatika/intellektualnye\_sistemy (дата обращения: 08.10.2020).

[11] Альманах «Искусственный интеллект» / Центр компетенция НТИ на базе МФТИ. – 2019. – Вып. 2  : Обработка естественного языка, распознавание и синтез речи. – 99 с – URL: https://singapore-academy.org/libcdo/100.pdf (дата обращения: 10.10.2020).

[12] Alexander Polomodov. Обзор Альманаха по ИИ #2 про NLP. [Электронный ресурс] / Alexander Polomodov. – URL: https://medium.com/@alexanderpolomodov/обзор-альманаха-по-ии-2-про-nlp-db36f4c6a8f3 (дата обращения: 10.10.2020).

1. Фонема – звук. [↑](#footnote-ref-1)
2. Графема – буква. [↑](#footnote-ref-2)
3. Сверхфразовое единство – примерно равно абзацу. [↑](#footnote-ref-3)
4. Анафора – повторение начальных созвучий, частей слов, словосочетаний, синтаксических конструкций. [↑](#footnote-ref-4)
5. Кореференция – обозначение одной сущности (референта) в разных частях текста с помощью различных обозначений. Пример: *Иван Алексеевич Бунин* — выдающийся русский *писатель и поэт*. *Он* является первым *лауреатом Нобелевской премии* по литературе из России. [↑](#footnote-ref-5)
6. https://www.researchgate.net/publication/331232339\_Metody\_informacionnogo\_poiska\_v\_komputernyh\_setah\_ssverhnasysennymi\_informacionnymi\_resursami [↑](#footnote-ref-6)
7. https://cyberleninka.ru/article/n/avtoreferirovanie-na-osnove-assotsiativnyh-poley-dominant [↑](#footnote-ref-7)
8. https://habr.com/ru/post/271771/ [↑](#footnote-ref-8)
9. https://ddecisions.ai/testsummary [↑](#footnote-ref-9)
10. http://masters.donntu.org/2013/fknt/mashchenko/library/translate.htm [↑](#footnote-ref-10)
11. http://www.swsys.ru/index.php?page=article&id=4170 [↑](#footnote-ref-11)
12. https://cyberleninka.ru/article/n/estestvenno-yazykovye-interfeysy-intellektualnyh-voprosno-otvetnyh-sistem [↑](#footnote-ref-12)
13. https://www.dissercat.com/content/ierarkhicheskaya-obrabotka-potokov-tekstovykh-soobshchenii-na-baze-naivnogo-baiesovskogo-kla [↑](#footnote-ref-13)
14. http://www.recognition.mccme.ru/pub/RecognitionLab.html/tasks.html [↑](#footnote-ref-14)
15. https://www.hse.ru/data/2017/07/27/1173643651/Ionov.pdf [↑](#footnote-ref-15)
16. https://translatedby.com/you/anaphora-resolution-the-state-of-the-art-introduction/into-ru/ [↑](#footnote-ref-16)
17. http://www.ccas.ru/avtorefe/034avtor.pdf [↑](#footnote-ref-17)
18. http://itas2016.iitp.ru/pdf/1570285121.pdf [↑](#footnote-ref-18)
19. https://dspace.spbu.ru/bitstream/11701/5052/1/Bodrova\_masterThesis\_Rus\_Final.pdf [↑](#footnote-ref-19)
20. http://www.spiiras.nw.ru/dissovet/wp-content/uploads/2017/08/dissertacija\_karpovich.pdf [↑](#footnote-ref-20)
21. https://elar.urfu.ru/bitstream/10995/79844/1/vopon\_2018\_1\_014.pdf [↑](#footnote-ref-21)
22. http://www.dialog-21.ru/media/2393/kreydlinl.pdf [↑](#footnote-ref-22)
23. https://cyberleninka.ru/article/n/problema-omonimii-imen-sobstvennyh-i-imen-naritsatelnyh-v-rechevyh-naimenovaniyah [↑](#footnote-ref-23)
24. https://cyberleninka.ru/article/n/problema-semantiki-imeni-sobstvennogo [↑](#footnote-ref-24)
25. https://habr.com/ru/post/349864/ [↑](#footnote-ref-25)
26. https://cyberleninka.ru/article/n/avtomatizirovannoe-vydelenie-terminologii-iz-tekstov-dogovorov-na-angliyskom-yazyke-pri-pomoschi-sredstv-korpusnoy-lingvistiki [↑](#footnote-ref-26)
27. https://cyberleninka.ru/article/n/vydelenie-terminologicheskogo-polya-neobhodimoe-uslovie-izucheniya-spetsialnoy-leksiki-angliyskogo-grazhdanskogo-sudoproizvodstva [↑](#footnote-ref-27)
28. http://www.dialog-21.ru/digest/2001/articles/bolshakova/ [↑](#footnote-ref-28)
29. http://www.dialog-21.ru/digests/dialog2008/materials/html/11.htm [↑](#footnote-ref-29)
30. Д.Д. Голомазов Выделение терминов из коллекции текстов с заданным тематическим делением https://istina.msu.ru/download/9027061/1daCqR:RMDzAhnvS\_SpDE7gmtAJ31vnXiE/ [↑](#footnote-ref-30)
31. https://www.science-education.ru/ru/article/view?id=13485 [↑](#footnote-ref-31)
32. АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ ТЕРМИНОВ ИЗ ТЕКСТОВ ПРЕДМЕТНЫХ ОБЛАСТЕЙ И УСТАНОВЛЕНИЕ СВЯЗЕЙ МЕЖДУ НИМИ Новикова Д.С. http://www.conf.sci.pfu.edu.ru/index.php/ittmm/2012/paper/download/245/409 [↑](#footnote-ref-32)
33. Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации В.П. Захаров, М.В. Хохлова http://ojs.itmo.ru/index.php/IMS/article/download/268/264 [↑](#footnote-ref-33)
34. https://www.elibrary.ru/item.asp?id=35568592 [↑](#footnote-ref-34)
35. https://pureportal.spbu.ru/ru/publications/%D0%B2%D1%8B%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5-%D1%82%D0%B5%D1%80%D0%BC%D0%B8%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D1%85-%D1%81%D0%BB%D0%BE%D0%B2%D0%BE%D1%81%D0%BE%D1%87%D0%B5%D1%82%D0%B0%D0%BD%D0%B8%D0%B9-%D0%B8%D0%B7-%D1%81%D0%BF%D0%B5%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D1%85-%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2 [↑](#footnote-ref-35)
36. См, например, https://machinelearningmastery.com/sparse-matrices-for-machine-learning/ и https://habr.com/ru/company/it-grad/blog/344320/ [↑](#footnote-ref-36)