МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ

КАЗАХСТАН

Казахский национальный исследовательский
технический университет имени К.И.Сатпаева

Институт информационных и телекоммуникационных технологий

УДК 004.942 На правах рукописи

Махамбетәлі Темірлан Қалыбекұлы

# МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

На соискание академической степени магистра

| | |
|---|---|
| Название диссертации | Анализ трендов изменения наукометрических показателей |
| Направление подготовки | 6М070500 – «Математическое и компьютерное моделирование» |

Научный руководитель
Профессор, д-р техн. наук
_____ Р.И.Мухамедиев
"08"июля 2020г.

Рецензент, к.т.н.
_____ Р.Р.Мусабаев
"08" июля 2020г.

**ДОПУЩЕН К ЗАЩИТЕ**
Директор НОЦ МиК
_____ Н.С.Даирбеков
" 08" июля 2020г

Нормоконтроль
Ведущий инженер
_____ М.С.Амирбекова
"08" июля 2020г

Алматы 2020

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный исследовательский технический университет имени К.И. Сатпаева

Институт кибернетики и информационных технологий

Научно-образовательный центр математики и кибнернетики

6М070500 – Математическое и компьютерное моделирование

**УТВЕРЖДАЮ**
Директор НОЦ МиК
_____ Н.С.Даирбеков

“08” июля 2020г.

# ЗАДАНИЕ
## на выполнение магистерской диссертации

Магистранту *Махамбетәлі Темірлан Қалыбекұлы*
Тема: *Анализ трендов изменения наукометрических показателей*
Утверждена *приказом Ректора*
*Университета № 1001 –М от " 1 6 " марта 20 20г.*
Срок сдачи законченной диссертации *"8" июля 2020г.*
Исходные данные к магистерской диссертации: *Анализ трендов изменения наукометрических показателей*
Перечень подлежащих разработке в магистерской диссертации вопросов:
а) *Сбор и анализ информации по исходным данным*
*б) Анализ существующих индикаторов*
*в) Изучение методов оценки публикаций*
*г) Разработка предсказательных моделей на python*
*д) Заключение*
*е) Список использованной литературы*
Перечень графического материала (с точным указанием обязательных чертежей):
*Презентация диссертации на 42 слайдах*

Рекомендуемая основная литература:


1. Garfield E. Citation indexes for science //Science. – 1955. – Т.122. – №3159. - p.108–11
2. Muhamedyev R. et al. New bibliometric indicators for prospectivity estimation of research fields //Annals of Library and Information Studies (ALIS). – 2018. – Т. 65. – №. 1. – p. 62-69.
3. Barakhnin V., Duisenbayeva A., Kozhemyakina O., Yergaliyev Y., Muhamedyev R. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area// Journal of Physics: Conference Series.- 2018 - №1117. 012001.

**ГРАФИК**

подготовки магистерской диссертации

| Наименование разделов, перечень разрабатываемых вопросов | Сроки представления научному руководителю | Примечание |
|---|---|---|
| Сбор и изучение материала по наукометрии и обзор литературы | 21.03.2020 | |
| Обзор литературы. Обзор существующих методов и изучение данных | 20.04.2020 | |
| Предварительная обработка данных, построение регрессии | 10.05.2020 | |
| Анализ полученных результатов и выводы. | 29.05.2020 | |

**Подписи**

консультантов и нормоконтролера на законченную магистерскую диссертацию с указанием относящихся к ним разделов диссертации

| Наименования разделов | Консультанты, И.О.Ф. (уч. степень, звание) | Дата подписания | Подпись |
|---|---|---|---|
| Сбор и изучение материала по наукометрии | Р.И.Мухамедиев, профессор, д-р техн.наук | 8.07.2020 | |
| Обзор существующих методов и изучение данных | Р.И.Мухамедиев, профессор, д-р техн.наук | 8.07.2020 | |
| Применение регрессионной модели и анализ результатов | Р.И.Мухамедиев, профессор, д-р техн.наук | 8.07.2020 | |
| Нормоконтроль | М.С.Амирбекова, ведущий инженер | 8.07.2020 | |

Научный руководитель _____ Р.И.Мухамедиев

*(подпись)* *(Ф.И.О.)*

Задание принял к исполнению обучающийся _____ Т.Қ. Махамбетәлі

*(подпись)* *(Ф.И.О.)*

Дата “08” июля 2020 г.

## Abstract

In this work, we consider scientometric indicators of such a rapidly developing field of research as automatic text processing (natural language processing). Differential indicators of speed and acceleration were used to assess the dynamics of the development of NLP domains. The assessment was based on data from a direct bibliographic database Science and eLibrary.ru. Calculations were performed for the following NLP subdomains: grammar checking, information extraction, text categorization, dialogue systems, speech recognition, machine translation, information search, answers to questions, opinion analysis, intelligent advisers and others. Areas with high growth rates (sentiment analysis, statistical methods, deep learning) and areas that lost the preexisting dynamics of publication activity growth (automatic summarization, speech recognition, information retrieval) were identified. The proposed indicators allow to visually express changes in the dynamics of scientometric indicators, which may be useful in assessing the prospects of research areas.

**Аннотация**

В данной работе рассматриваются наукометрические показатели такой быстро развивающейся области исследований, как автоматическая обработка текста (обработка естественного языка). Дифференциальные показатели скорости и ускорения были использованы для оценки динамики развития доменов NLP. Оценка была основана на данных из прямой библиографической базы данных Science и eLibrary.ru. Расчеты проводились для следующих разделов NLP: проверка грамматики, извлечение информации, категоризация текста, системы диалогов, распознавание речи, машинный перевод, поиск информации, ответы на вопросы, анализ мнений, интеллектуальные консультанты и другие. Были определены области с высокими темпами роста (анализ тональности, статистические методы, глубокое обучение) и области, которые потеряли существующую динамику роста публикационной активности (автоматическое обобщение, распознавание речи, поиск информации). Предлагаемые показатели позволяют наглядно выразить изменения в динамике наукометрических показателей, что может быть полезно при оценке перспективности исследований.

**Аңдатпа**

Бұл мақалада сөздерді автоматты өңдеу (тілді өңдеу) сияқты қарқынды дамып келе жатқан зерттеу саласының ғылымиометриялық көрсеткіштері қарастырылады. Жылдамдық пен үдеудің дифференциалды көрсеткіштері NLP бөлім даму динамикасын бағалау үшін пайдаланылды. Бағалау тікелей библиографиялық мәліметтер базасының мәліметтеріне Science и eLibrary.ru негізделген. Есептер NLP-нің келесі қосалқы бөлімдерінде жүргізілді: грамматикалық тексеру, ақпаратты шығару, мәтінді санаттау, диалог жүйелері, сөйлеуді тану, машиналық аударма, ақпаратты іздеу, сұрақтарға жауаптар, пікірлерді талдау, ақылды кеңесшілер және басқалар. Өсудің жоғары қарқыны бар (көңіл анализ, Статистикалық тәсілдер, тереңдетіп оқыту) және жарияланым белсенділігінің өсу динамикасын жоғалтқан (автоматты жалпылау, сөйлеуді тану, ақпаратты іздеу) бағыттары анықталды. Ұсынылған индикаторлар зерттеу перспективаларын бағалауда пайдалы болуы мүмкін ғылымиометриялық көрсеткіштер динамикасындағы өзгерістерді көзбен көрсетуге мүмкіндік береді.

## СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

Обработка естественного языка является одной из тех наук, развитие которых отличается особой динамикой, технология быстро развивается благодаря растущему интересу к методам коммуникаций между людьми и компьютером, также сыграла роль доступность больших данных, увеличение мощности вычислений, разработка новых и модификация существующих алгоритмов. Данная область характеризуется широким спектром задач и приложений. Особый интерес представляет для исследователя понимание того ,как каждая из областей науки развивается, что может быть полезно для поиска предпочтительного направления.

Количество публикаций, количество цитирований (индекс цитирования), количество соавторов, индекс Хирша[1] и другие библиометрические показатели можно использовать для оценки развития научной области. Выявление перспективных областей, в которых эти показатели имеют большие значения, позволяет при прочих равных условиях более четко представить ситуацию в анализируемой области научных исследований. Тем не менее, количество публикаций растет во многих областях обработки естественного языка. Поэтому простого заявления о растущем интересе основываясь количестве публикаций недостаточно. Для выявления закономерностей изменения публикационной активности были введены дифференциальные индикаторы D1 («скорость») и D2 («ускорение») [2]. Дифференциальные индикаторы позволяют оценить динамику изменения использования выбранных ключевых терминов авторами научных публикаций. Таким образом, можно более четко отразить рост или снижение интереса исследователей к использованию ключевых слов, характеризующих область исследований.

NLP как область исследований решает проблему разработки методов автоматического анализа и представления естественного человеческого языка. Исследования в этом направлении ведутся с 50-х годов прошлого века. Практические задачи NLP включают в себя:

- Автоматический перевод

- Выработка ответов на пользовательские запросы
- Извлечение информации
- Поиск информации
- Анализ тональности
- И другие области, связанные с обработкой устной или письменной речи

В последние годы был достигнут значительный прогресс в области автоматического перевода (машинного перевода), обобщения (автоматического реферирования), поиска информации, систем ответа на вопросы, анализа тональности (анализа настроений), извлечения информации, проверки авторства. Успех NLP является следствием развития методов машинного обучения, многократного увеличения вычислительной мощности, доступности большого количества лингвистических данных и развития понимания структуры естественного языка . в социальном контексте[2].

Необходимость решения практических проблем NLP послужила катализатором для разработки нескольких новых методов, среди которых можно упомянуть машинное обучение и его подразделы: нейронные сети и глубокое обучение (deep learning –DL), векторное представление слов и текстов, автоматический морфологический анализ  и так далее. NLP и DL как область исследований быстро меняются.

Интересно и важно оценить динамику этих изменений, оценивая публикационную активность в разных разделах NLP и DL. Идея оценки публикационной активности может быть приписана работам Э. Гарфилда[3], который ввел понятие индекса научного цитирования (SCI). Позднее библиометрические показатели (количество публикаций, индекс цитирования, число соавторов и т. Д.)

Однако, основной целью этих показателей является оценка личного вклада ученого. Для оценки динамики изменений в областях исследований можно использовать показатели цитирования и количество публикаций.

Оценка изменений этих показателей с течением времени с ежегодным приростом не представляет трудностей, но для сравнительного анализа важны числовые показатели роста. С этой целью на основе дифференциальных индикаторов, введенных в [2] для оценки роста областей исследований, были рассчитаны динамические показатели публикационной активности в области NLP и различных приложений DL.

## 1. Актуальность и цели работы

Обработка естественного языка (NLP) является быстро развивающейся областью исследования. При решении задач NLP наряду с традиционными методами, основанными на статистической модели языка, используются методы машинного обучения (ML). В работе рассматриваются библиометрические показатели NLP и DL. Оценены динамические показатели[2] и определены области исследований с самыми высокими темпами роста. Индикаторы были рассчитаны для следующих приложений NLP: проверка грамматики, извлечение информации, классификация текста, диалоговые системы, распознавание речи, машинный перевод, поиск информации, ответы на вопросы, анализ мнений, интеллектуальные советники и т. д. Зафиксированы самые высокие значения динамических показателей по разделам: нейронные сети, извлечение информации, анализ тональностей, обучение с учителем. В свою очередь, глубокое обучение используется для решения широкого спектра задач. Его особенностью являются повышенные требования к объему обрабатываемых данных. В статье дается оценка динамики роста библиометрических показателей для некоторых приложений глубокого обучения. Одним из наиболее динамично развиваемых исследований является область применения глубокого обучения для решения проблем здравоохранения.

**Целью работы** является представление объективных библиометрических индикаторов оценивающих динамику развития научной сферы.

**Задачи исследования:**

1) Сгруппировать выборку статей по научным подразделам

2) Построить регрессионную зависимость связывающую год и количество публикаций (цитирований)на основе данных по количеству цитирований и публикаций
3) Рассчитать дифференциальные индикаторы D1 и D2
4) Построить графики изменения D1 и D2 во времени

**Научные статьи и публикации**. По теме диссертации опубликована 1 статья в IT & M2020 The international Scientific Conference, Апрель 2020

**Объем и структура работы.** Магистерская диссертация состоит из 8 разделов, изложена на 42 страницах основного текста. Работа содержит 11 формул, 23 рисунка, 3 таблицы и скрипт кода на программе Python.

## 2.1. Библиометрические индикаторы

Оценка научных исследований всегда была трудной задачей. Процесс рецензирования, который был основой научной оценки в течение почти столетия, требует времени, опыта и немалых ресурсов для правильной работы. Но некоторые тенденции в научных исследованиях сделали этот процесс еще более сложным. Огромное количество научных публикаций, выпускаемых в год, росло экспоненциальными темпами в течение более пятидесяти лет и не показало никаких признаков замедления в ближайшее время. Эти публикации также становятся все более техническими и специализированными, что затрудняет поиск квалифицированных рецензентов. Оценка научных исследований в этом контексте становится не только все более сложной, но и все более важной для обеспечения того, чтобы подходящие исследователи получали поощрения и финансирование для продолжения своей работы.

В этой среде ряд наблюдательных советов, учреждений и даже стран обращаются к библиометрии для облегчения процесса обзора. Библиометрия - это количественный анализ публикаций. Он по существу собирает данные из публикаций и анализирует эти данные различными способами, чтобы ответить на вопросы об исследованиях, которые представляют эти публикации. Таким образом, область библиометрии охватывает широкий спектр подходов и

методов, она стала наиболее известной своими попытками измерить влияние научных исследований посредством использования различных библиометрических показателей, таких как, например, импакт-фактор[12] и иднекс Хирша[1]. Эти показатели воспринимаются как более объективные, чем рецензирование, потому что они могут быть рассчитаны с гораздо меньшими затратами времени и усилий, чем рецензирование, и потому что есть некоторые свидетельства того, что эти показатели согласуются с мнением коллег, рецензенты и политики все чаще используют индикаторы как дополнение, а в некоторых случаях вместо экспертной оценки.

Хотя использование библиометрических показателей может служить ценным дополнением к процессу экспертной оценки, эти показатели слишком часто вынимаются из контекста и применяются без полного понимания библиометрического исследования, на котором они основаны. В результате они часто используются, чтобы измерить вещи, которые они не должны были измерять или сделать сравнения, которые они фактически не способны сделать. В этой статье также дается краткое введение в основные идеи, лежащие в основе этих индикаторов.

## 2.2. Что можно измерить библиометрическими индикаторами

Все библиометрические показатели основаны на идее, что мы можем измерить влияние статьи, подсчитав количество других статей, которые ее цитировали. Цитаты, согласно теории, действуют как вотум доверия или знак влияния от одного документа к другому. Тот факт, что одна статья ссылается на другую, свидетельствует о том, что цитируемая статья оказала какое-то влияние цитирующую ее статью. Таким образом, подсчет количества ссылок, полученных статьей, позволяет нам оценить влияние, которое статья оказала на науку в целом.

Считая ссылки на набор статей - одним автором, учреждением или даже целой страной, мы можем измерить влияние, которое набор публикаций оказал на научные исследования. Больше цитирований - больше влияния.

Проблема этой идеи заключается в том, что признание влияния является лишь одной из многих причин, по которым авторы цитируют другие статьи [13]. Авторы ссылаются на другие статьи  по разным причинам: ссылаются на конкретную методологию, указывают на примеры другой работы, проделанной по той же теме, чтобы подчеркнуть точку зрения, которую они делают в статье, отдать должное своим наставникам или экспертам в поле, или даже обсудить примеры ошибочных методов или вводящих в заблуждение результатов. Современные библиометрические показатели не могут объяснить это разнообразие, обоснованное, с одной стороны, субъективностью автора при выборе статей оказавших влияние на его публикацию; они считают все цитаты одинаковыми, независимо от фактической причины цитирования. В результате нельзя с уверенностью сказать, что цитируемая статья действительно очень влиятельна. Что мы, вероятно, можем сказать, следуя примеру Э. Гарфилда,[14] одного из основателей библиометрии, так это то, что цитируемые статьи очень полезны для авторов при написании других работ. Однако для чего эти статьи полезны, неясно.

Это означает, что число цитирований измеряет очень конкретное определение воздействия. Подсчет цитирования только измеряет влияние или полезность статей для авторов других работ; они не измеряют влияние этих документов на что-либо еще. Из числа цитирований статьи невозможно определить, сообщали ли они о прорыве в биомедицинском понимании, прогрессе в клинической практике, который значительно улучшил результаты лечения пациентов, особенно полезном методе анализа данных или своевременном обзоре существующей литературы.

Количество цитирований, полученных статьей, не может измерить, улучшило ли исследование, опубликованное в этой статье, здоровье людей. Можно только измерить, была ли статья полезна другим авторам для написания их собственных статьей. Конечно, это форма воздействия, но не обязательно та, которую, по мнению рецензентов, они измеряют.

### 2.3. Наукометрические методы анализа

Определение наукометрии впервые было предложено в [15] как «количественное исследование развития науки». Его можно рассматривать как методику, которая включает измерение воздействия на исследования, понимание процесса цитирования, составление карты структуры знаний и эволюции в области, основанной на крупномасштабном наборе научных данных [9]. Благодаря обработке огромных библиометрических данных, наукометрические методы помогают исследователям находить систематические литературные открытия, связывая литературные концепции, которые могут быть упущены при ручном обзоре [10].

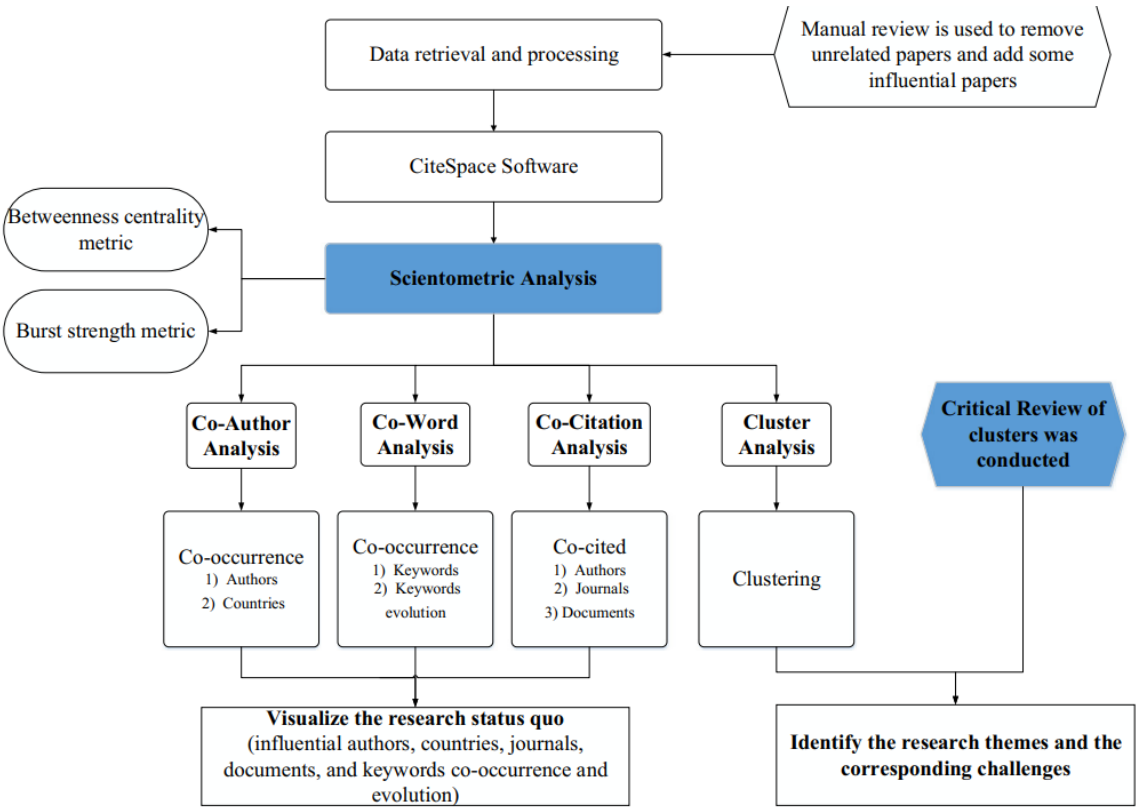



*Рисунок 1. Наукометрический анализ*

## 2.4. Ограничения и ошибки при измерении

В дополнение к путанице в отношении того, что измеряют цитаты, существует также путаница в том, как на самом деле работают библиометрические показатели, что приводит к тому, что оценщики допускают ошибки при их использовании. Одна из самых распространенная ошибок - использование импакт-фактора журнала для измерения влияния статьи, опубликованной в этом журнале. Оказывается, что влияние любого журнала в первую очередь определяется цитатами, полученными небольшой частью (10–30%) статей[7], опубликованных в этом журнале. То есть, несколько статей в этом журнале получают чрезвычайно большое количество ссылок, в то время как подавляющее большинство статей получает мало или совсем не цитируется. Фактор влияния может быть действительной мерой воздействия цитирования журнала; однако это не является действительной мерой влияния цитирования статьи.

Вторая распространенная ошибка, которую делают оценщики, заключается в том, что они не принимают во внимание время. Цитаты не только накапливаются, но и накапливаются с течением времени. Исследования показали, что для публикации достаточно цитат, чтобы библиографические индикаторы были достоверными, по крайней мере, через два-три года после публикации. Это означает, что самым последним статьям, включенным в любую институциональную оценку с использованием библиометрических показателей, должно быть не менее двух лет [16]. С другой стороны, цитаты продолжают накапливаться после этого начального периода времени, что означает, что более старые статьи, как правило, цитируются более высоко, чем более молодые, потому что у них было больше времени для накопления цитат. В результате любой список наиболее цитируемых статей, опубликованных в дисциплине или журнале за определенный период времени, будет смещен в сторону более старых статей [17]. Это также означает, что библиометрические показатели для отдельных авторов всегда будут смещены в сторону более старых авторов. Это связано с тем, что у молодых авторов не так много публикаций, и потому, что на их публикации не было так много времени для цитирования.

## 2.5. Возможности

Итак, со всеми ограничениями и ошибками, связанными с библиометрическими показателями, их всё же можно использовать потому, что у золотого стандарта, рецензирования, есть свои проблемы. В дополнение ко времени и затратам, которые требуются для выполнения, процесс рецензирования редко дает согласованные или воспроизводимые результаты. Рекомендации одной экспертной комиссии могут напрямую противоречить рекомендациям другой, даже если они рассматривают одну и ту же заявку. Рецензенты также подвержены сознательным и бессознательным формам

предвзятости, которые могут серьезно повлиять на их суждение и окончательные рекомендации. Наконец, при оценке отдельных лиц или учреждений с сотнями или тысячами публикаций рецензенты не могут оценить все из них, поэтому они могут читать только несколько статей и вынуждены игнорировать остальные. Эти предубеждения означают, что, хотя рецензирование остается золотым стандартом оценки научных исследований, оно не лишено недостатков.

Библиометрические показатели, если они используются ответственно, могут в некоторой степени нивелировать эти недостатки. Так уж сложилось, что сильные стороны библиометрических показателей точно соответствуют слабым сторонам рецензирования. Библиометрические показатели могут быть рассчитаны для всего набора публикаций и представляют собой коллективное суждение широкого сегмента научного исследовательского сообщества, а не мнения отдельных лиц, выбранных для группы обзора. Библиометрические показатели также могут быть более прозрачными и воспроизводимыми, чем рецензирование. Они также могут помочь в процессе рецензирования, указав на неточности в публикациях или записях цитирования, на которых рецензенты, возможно, пожелают сосредоточиться во время рецензирования. В результате, сообщество библиометрических исследователей рекомендует использовать библиометрические показатели в качестве дополнения, а не замены для информированного экспертного обзора при оценке научных исследований. Каждый метод уравновешивает слабости другого.

Сочетание библиометрических показателей и экспертной оценки приводит к более справедливым, сбалансированным и точным оценкам научных исследований. Поскольку в этих оценках поставлено на карту не что иное, как будущее научных исследований, жизненно важно, чтобы их понимали правильно.

## 3. Способы оценки

### 3.1. Количественные индикаторы

Количественные показатели предназначены для измерения продуктивности исследователя или группы. Самый простой способ - подсчитать количество статей, опубликованных конкретным автором или исследовательской группой за определенный период времени [18]. Хотя этот подсчет является очень простым показателем, который может быть легко рассчитан самими авторами, при сравнении его с авторами или исследовательскими группами следует быть очень осторожным. Хотя количество публикаций действительно отражает продуктивность автора или исследовательской группы, оно не влияет на качество статей [19]. Очевидно, что влияние сообщения о случае не эквивалентно влиянию клинического исследования. При сравнении групп следует учитывать, что на количество публикаций также влияет размер группы [20].

### 3.2. Индикаторы эффективности

Чтобы преодолеть некоторые из этих ограничений, более избирательным подходом будет подсчет количества публикаций в журналах самого высокого качества в соответствии, например, с их фактором воздействия [11]. Этот подход, однако, не учитывает влияние размера группы. Следовательно, с этим индексом вышеупомянутое ограничение сохраняется при сравнении групп на основе количества публикаций, имеющих самый высокий рейтинг. Хотя этот подход может выглядеть как показатель эффективности, он был разработан для устранения недостатков вышеупомянутого количественного показателя. Помимо производительности, существуют дополнительные критерии, которые

полезно учитывать. Показатели эффективности помогают определить уровень качества работы автора или исследовательской группы и могут использоваться для оценки воздействия исследований на научное сообщество. То, как часто другие люди цитируют статью, автора или журнал, является показателем эффективности: чем больше число ссылок, тем выше уровень эффективности. Затем число цитирований можно разделить на количество лет в течение определенного периода времени, чтобы получить среднее число цитирований за год. Для исследователей это количество ссылок может быть разделено на количество статей для получения среднего количества ссылок.

### 3.3. Импакт-фактор журнала

Фактор воздействия журнала (ИФ) является мерой ссылок на журнал и предназначен для оценки важности журнала в данной области. ИФ был впервые предложен в 1955 году Э. Гарфилдом, пионером в исследованиях цитирования, и был концептуально разработан в начале 1960-х годов самим и И.Х. Шер [21]. Этот индикатор, который в настоящее время называется «импакт фактор журнала», является, вероятно, наиболее широко используемым индикатором важности журнала в различных научных областях. ИФ доступны в отчетах цитирования журнала SCI (Science Citation Index) и в Web of Knowledge для более чем 8000 избранных научных журналов. Импакт-фактор журнала устанавливается каждый год на основе предыдущего 2-летнего периода. Он определяется следующим образом: каждый рассчитывает, сколько раз статьи, опубликованные в течение данного двухлетнего периода, цитировались в журналах в течение года после указанного периода. Это число делится на общее

количество «цитируемых предметов», которые были опубликованы в этом журнале за то же 2- летний период. ИФ имеет несколько ограничений:

Во-первых, хотя более высокий ИФ может предполагать большее влияние журнала, он не отражает качество каждой конкретной статьи, опубликованной этим журналом. Следовательно, неясно, является ли высокий ИФ результатом умеренной степени цитирования всех опубликованных статей или высокой степенью цитирования только некоторых статей [22].

Во-вторых, междисциплинарные журналы, как правило, имеют более высокий ИФ, чем специализированные журналы. Одним из объяснений этого является то, что междисциплинарные журналы имеют очень большую читательскую аудиторию из разных областей и, следовательно, цитируются чаще. Отдавая предпочтение только журналам с высоким ИФ, авторы могут игнорировать специализированные журналы, которые на самом деле могут быть более подходящими для охвата их целевой аудитории, несмотря на то, что ИФ ниже, чем у междисциплинарных журналов.

В-третьих, существуют различия между областями исследований, в том числе в интенсивности исследований. Журнал с самым высоким рейтингом в каждой специализированной области может иметь разный ИФ от специальности к специальности. Эти различия могут быть объяснены различиями в популярности, привычках цитирования и динамике цитирования. Популярность относится к числу авторов, статей и, следовательно, цитат, связанных с областью исследования. Популярность сильно варьируется в зависимости от сферы - чем больше популярность в данной области, тем больше число исследователей, вовлеченных в эту область, тем больше число опубликованных статей и тем выше ИФ соответствующих журналов. Например, статьи в журналах по клеточной биологии цитируются в среднем в пять раз чаще, чем статьи в журналах по кристаллографии[23]. Таким образом, не разумно сравнивать

журналы разных областей только на основе их соответствующих ИФ. Среднее значение также сильно варьируется в зависимости от сферы.

### 3.4. Индекс непосредственности

Индекс непосредственности измеряет текущую значимость статьи, опубликованной журналом, путем расчета среднего числа цитирований статей, опубликованных в течение определенного года конкретным журналом в течение того же года [24]. Данный индекс рассчитывается на базе количества, когда статьи, опубликованные в данном журнале, цитируются другими, а затем делением этого числа на количество статей, опубликованных в этом журнале в том же году.

### 3.5. Полу-период цитирования

Указанный в журнале полу-период цитирования- это количество лет, начиная с текущего года, которые составляют 50% от общего количества ссылок, полученных журналом в текущем году [25]. Полу-период цитирования относится к промежутку времени между публикацией цитируемого исследования и публикацией статей, цитирующих это исследование [26]. Хотя, полу-период цитирования не отражает научную ценность конкретного журнала, он может дать информацию о своей редакционной политике или области исследований. Полу-период цитирования может отражать редакционную политику, которая подчеркивает текущую осведомленность или быстро развивающаяся область исследований, в то время как полный период цитирования может отражать либо акцент на архивной литературе или медленно развивающуюся область [25].

### 3.6. Собственный фактор

Как упоминалось ранее, ИФ отражает качество журнала через количество цитируемых им статей: чем выше ИФ конкретного журнала, тем выше его подразумеваемое качество и его предполагаемое влияние на научное сообщество. При расчете ИФ качество цитирующих журналов не учитывается. Тем не менее, одно цитирование в престижном журнале может быть более ценным, чем многократное цитирование в журналах низкого качества [27]. Так называемый, собственный фактор - учитывает качество цитирующих журналов, взвешивая их цитирование через их влияние на научное сообщество [28]. Собственный фактор предполагает, что научная литература образует обширную сеть статей, связанных друг с другом своими цитатами, и использует структуру этой сети для измерения относительного влияния журналов [29]. Чтобы оценить уровень значимости занимаемое определенным журналом в этой сети, Собственный фактор оценивает время, проведенное исследователем за этим журналом [30].

### 3.7. Индекс Хирша

Э.Хирш предложил индекс[1], который можно использовать для измерения научного воздействия отдельных исследователей: h-индекс - чем выше индекс h, тем показательнее деятельность исследователя. Индекс конкретного исследователя равен h, если h из его N статей имеют по меньшей мере h цитат в каждой, а в других N-h статьях цитирований меньше чем h. Индекс Хирша обладает рядом положительных свойств среди которых, например :

- индекс можно рассчитать вручную с помощью бесплатной базы данных, такой как Google Scholar. Данные, необходимые для его расчета, также доступны в Web of Science на основе подписки, где они автоматически рассчитываются для определенного исследователя.

- H-индекс нечувствителен к статьям, которые редко или никогда не были цитированы, а также к чрезвычайно часто цитируемым статьям (например, обзорам). Поскольку индекс Хирша объединяет как количество статей отдельных исследователей, так и их количество цитирований, он отдает предпочтение постоянным исследователям, которые публикуют непрерывный поток статей с длительным воздействием и уровнем выше среднего, а не авторам одной хоть и очень влиятельной статьи.

Тем не менее, индекс Хирша не лишен недостатков. Поскольку он основан на показателях цитирования, он остается очень чувствительным к характеристикам области исследования (популярность сферы, привычки цитирования и динамика цитирования) [31]. Таким образом, индекс Хирша не следует использовать для сравнения исследователей в различных областях науки. Кроме того, поскольку индивидуальные значения индекса увеличиваются со временем, этот индекс ставит исследователей с короткой карьерой в невыгодное положение, независимо от качества их производства в течение этого карьерного периода. Следовательно, исследователей следует сравнивать только в пределах одной и той же области исследований и на аналогичном периоде.

# 4. Методология

## 4.1. Библиометрические методы оценки публикационной активности

Библиометрия - это набор математических и статистических методов, используемых для анализа и измерения количества и качества книг, статей и других форм публикаций[30]. Существует три типа библиометрических показателей:

- количественные показатели, которые измеряют продуктивность конкретного исследователя;
- качественные показатели, которые измеряют качество (или «производительность») результатов исследования;
- структурные показатели, которые измеряют связи между публикациями, авторами и областями исследований.

В данной работе были использованы количественные показатели публикаций и цитирований, изучены и рассчитаны дифференциальные индикаторы впервые представленные в [2].

Библиометрические показатели особенно важны для исследователей и организаций, так как эти измерения часто используются при принятии решений о фондировании, назначениях и продвижении по службе исследователей. По мере увеличения научных открытий и их публикаций следует и рост в количестве цитирований другими исследователями, в таком случае библиометрические показатели становятся все более важными как инструмент объективной оценки. Идея оценки публикационной активности в области науки с применением индекса цитирования (SCI) был предложен Е.Гарфилдом [3]. Позже, библиометрические показатели как количество публикаций, индекс цитирования, количество соавторов и т. д. стали учитываться при оценке качества научной статьи и научной деятельности автора. Однако, несмотря на широкое использование этого показателя, его основным назначением является оценка личного вклада ученого. В данной же работе основная цель заключается

не в том чтобы оценить исследователя, а в том чтобы оценить область его исследований и насколько перспективной она является. Сам индекс Хирша мог бы войти в расчет представленных индикаторов, но в некоторой нормированной форме, по ранее описанной причине, где говорится о контрасте показателя в зависимости от самой области исследования.

Чтобы оценить развитие той или иной области, и выявить ее перспективность, важно обращать внимание не на само количество публикаций и цитирований, а на то как быстро эти показатели меняются, ведь более старые области будут всегда иметь большее количество публикаций и цитирований нежели только формирующиеся. Поэтому индикаторы увеличения количественных показателей используются для сравнительного анализа, чтобы отделить области с наиболее возрастающим интересом. С этой целью дифференциальные метрики для оценки развития областей исследований были введены в [2, 31]. Динамика прироста в каждом разделе была оценена на основании количества публикаций и цитат в области NLP.

Для этого определены следующие дифференциальные показатели научной области [2], определяемые как

$$D1_i = f_1(n_i, \frac{dn_i}{dt}, \frac{dc_i}{dt}) \quad (1)$$

$$D2_i = f_2(n_i, \frac{d^2n_i}{dt^2}, \frac{d^2c_i}{dt^2}) \quad (2)$$

Другими словами, индикаторы $D1_i$ это функция, зависимая от количества публикаций $n_i$, скорости изменения количества публикаций $\frac{dn_i}{dt}$ и количества цитирований $\frac{dc_i}{dt}$ в определенной изучаемой области.

Индикатор $D2_i$ это функция зависимая от количества публикаций $n_i$, ускорения в изменении количества публикаций $\frac{d^2n_i}{dt}$ и цитирований $\frac{d^2c_i}{dt}$ в определенной изучаемой области.

Функции $f_1$ и $f_2$ вычисляют тем или иным способом совокупный вклад цитирований и публикаций. В некоторых случаях агрегация может быть произведена с помощью взвешенного суммирования. Тогда, для определенной наукометрической базы $j$, индикаторы научной перспективности конкретной области в момент $t_k$, представленные в [32], могут быть записаны в следующем виде:

$$D1_i^j(t_k) = \alpha * n_i^j(t_k) + \beta * \frac{dn_i^j(t_k)}{dt} + \gamma * \frac{dc_i^j(t_k)}{dt} \quad (3)$$

$$D2_i^j(t_k) = \alpha' * n_i^j(t_k) + \beta' * \frac{d^2n_i^j(t_k)}{dt^2} + \gamma * \frac{dc_i^j(t_k)}{dt^2} \quad (4)$$

где $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$ - это определенные эмпирические коэффициенты которые взвешивают вклад количества публикаций, скорость и ускорение изменения в количестве публикаций $n_i$ и в количестве цитирований $c_i$ соответственно. Так как в работе анализируются относительные изменения показателей, сравниваемые с предыдущим периодом, и учитывая тот факт, что само количество публикаций и цитирований в некоторых областях намного больше чем в других, то часть напрямую включающая эти количества т.е. $\alpha * n_i^j(t_k)$ может быть опущена путем присваивания ей нулевого веса $\alpha = 0$ [32].

$$D1_i^j(t_k) = \beta * \frac{dn_i^j(t_k)}{dt} + \gamma * \frac{dc_i^j(t_k)}{dt} \quad (5)$$

$$D2_i^j(t_k) = \beta' * \frac{d\left(\frac{dn_i^j(t_k)}{dt}\right)}{dt} + \gamma' \frac{d\left(\frac{dc_i^j(t_k)}{dt}\right)}{dt} \quad (6)$$

Далее вычисления производятся по вышеуказанным формулам.

## 4.2. Данные для исследования

Для изучения динамики публикаций в области обработки естественного языка (OEJA) на русском языке были предоставлены данные [33] из ведущей Российской научной электронной библиотеки eLIBRARY.ru. В рамках работы [32] были определены следующие области публикационной активности для исследований. Сначала область NLP была рассмотрена с точки зрения решаемых задач, основанные на [34-40], были агрегированы по областям («NLP задачи» или «Задачи»): проверка грамматики, извлечение информации, категоризация текста, диалоговые системы, распознавание речи, машинный перевод, поиск информации, ответы на вопросы, анализ мнений и анализ настроений, интеллектуальные советники и автоматическое суммирование. Во-вторых, область NLP характеризуется быстрым ростом технологий и методов, которые способствуют решению вышеуказанных проблем. Были связаны следующие задачи с количеством методов (группа «Научные методы NLP» или «Методы»): машинное обучение, нейронные сети, глубокое обучение, нечеткая логика, логика первого порядка, представление знаний, эволюционные вычисления и генетическое программирование, системы на основе правил, обучение без учителя, кластеризация, обучение с учителем, статистические методы, байесовские сети, семантические сети , определение ключевых слов, лексическая близость, онтология, слияние информации, таксономия .

## 4.3. Регрессия, функция стоимости, градиентный спуск

Линейная регрессия для имплементации модели, функции стоимости и градиентного спуска. В обучении с учителем нам дают набор данных на вход и целевую переменную, например, исторические данные домов, проданных в городе. Для данного обучающего набора наша цель - изучить функцию h: X → Y, чтобы h (x) был «хорошим» предиктором для соответствующего значения y.

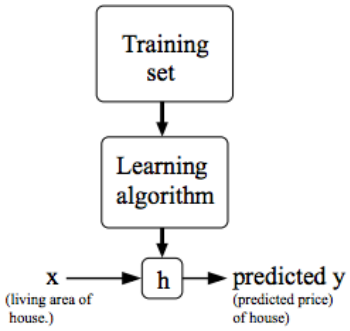



*Рисунок 2.Регрессия*

Когда целевая переменная, которую мы пытаемся предсказать, является непрерывной, как, например, в нашем примере, мы называем проблему обучения проблемой регрессии.

Итак, учитывая эти исторические данные, наша цель состоит в том, чтобы найти функцию, которая, с высокой точностью может предсказать количество публикаций. В случае линейной регрессии это будет функция прямой линии. Мы могли бы использовать другие квадратичные функции, которые могли бы помочь вписаться в наш набор данных, но для простоты мы начнем с линейной функции здесь. Нашей целью будет найти значение двух параметров theta0 и theta1, чтобы функция обеспечивала точный прогноз.

Как только мы определим функцию стоимости для нашей модели, нам нужно будет выяснить, как выбрать значения theta0 и theta1 так, чтобы ошибка для нашей модели была минимальной. Здесь мы начинаем с некоторого значения параметров, а затем постоянно продолжаем изменять параметры таким образом, чтобы в результате мы получили минимальную ошибку для нашего обучающего набора данных, используя нашу функцию модели.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \quad (7)$$

то есть выбрать $\theta$ так чтобы гипотетические значения были максимально близки к истинным. Для этого нужно задать начальные значения для коэффициентов $\theta$, затем продолжать менять значения коэффициентов уменьшая функцию стоимости до достижения минимума.

То, как мы продолжаем изменять значение параметров, называется градиентным спуском. Визуально мы видим, как мы можем начать с некоторого значения наших параметров, а затем продолжать изменять значения так, чтобы достичь минимума.

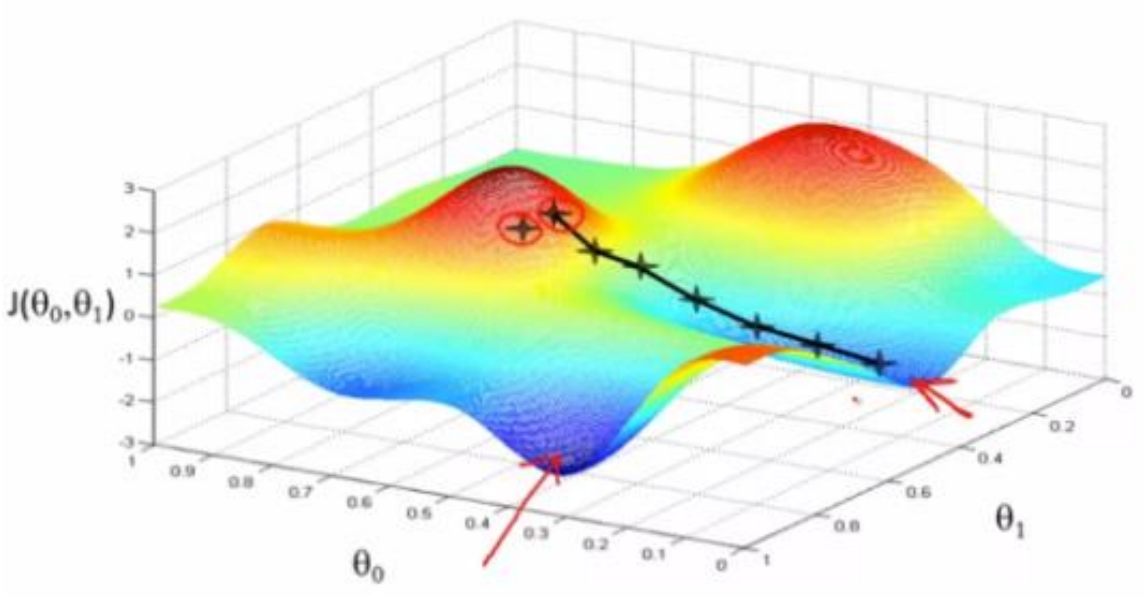



*Рисунок 3. Градиентный спуск*

На данный момент мы должны помнить, что, используя алгоритм градиентного спуска, мы можем продолжать изменять значение всех наших параметров так, чтобы мы могли достичь локального минимума. Альфа - это скорость обучения, которая определяет, насколько большой прыжок делается, чтобы приблизиться к минимуму. Мы умножаем скорость обучения на производную функции стоимости и затем вычитаем это значение из параметра.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (8)$$

$$temp_0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \quad (9)$$

$$temp_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

$$\theta_0 \coloneqq temp_0 \tag{10}$$
$$\theta_1 \coloneqq temp_1$$

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} * \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{11}$$

$$j = 0 \, , \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$j = 1 \, , \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x^{(i)}$$

- Если точка минимуму пройдена, то $\theta_j = \theta_j - (+slope)$, что понизит значение $\theta$.

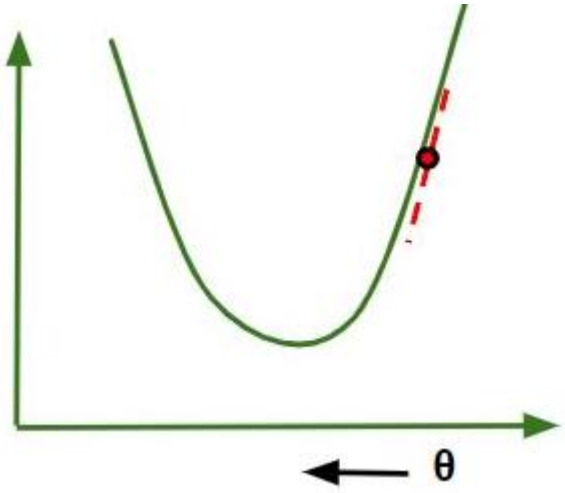



*Рисунок 4. Градиентный спуск*

- Если точка минимума еще не достигнута, то $\theta_j = \theta_j - (-slope)$, что увеличит значение $\theta$.

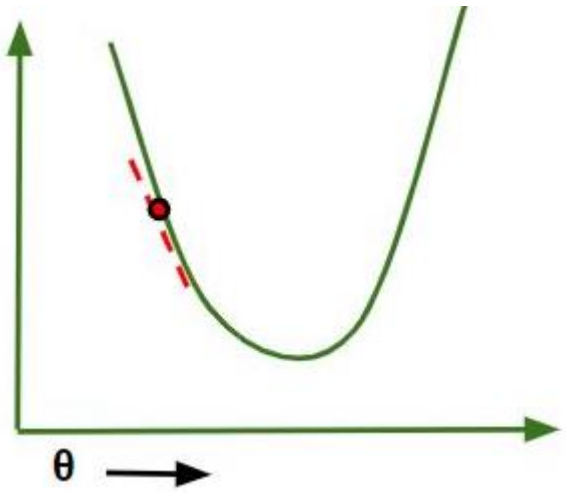



*Рисунок 5.Градиентный спуск*

- Если выбрать коэффициент обучения слишком большим, то градиентный спуск может пропустить точку минимума

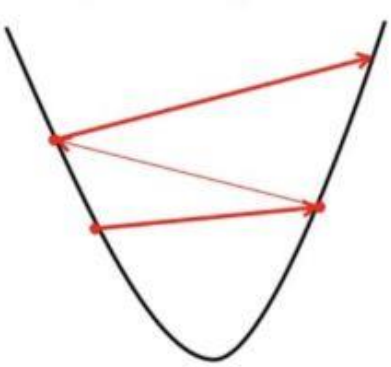


*Рисунок 6. Градиентный спуск*

- Если выбрать в качестве коэффициента обучения слишком маленькое число, то может быть потрачено намного больше времени для достижения минимума

*Рисунок 7.Градиентный спуск*

Градиентный спуск является одним из самых простых и широко используемых алгоритмов в машинном обучении, главным образом потому, что его можно применять к любой функции для его оптимизации. Есть несколько дополнительных концепций в применении градиентного спуска такие как:

- Выпуклость - в некоторых задачах линейной регрессии бывает только один минимум. Если поверхность ошибки выпуклая, то независимо от того, с чего мы начали, мы в конечном итоге достигнем абсолютного минимума. В общем, это не обязательно так. Возможно возникновение проблемы с локальными минимумами, из-за которой поиск по градиенту может застрять. Существует несколько подходов для преодоления этого (например, стохастический градиентный спуск).
- Производительность - существуют подходы к такому поиску параметров, которые могут сократить количество необходимых итераций. Например, линейный поиск сокращает число итераций для достижения разумного решения в несколько сотен раз.
- Сходимость - говорит о том, как определить, когда поиск находит решение. Обычно это делается путем поиска небольших изменений в ошибках от итерации к итерации (например, когда градиент близок к нулю).

## 5. Техническое инструменты

### 5.1. Python

Python был реализован в 1991 году Гвидо ван Россумом. С момента своего создания python был открытым исходным кодом. Программный Фонд Python управляет стандартизацией и проектированием языка и его библиотек. Процесс предложения по улучшению Python руководил его разработкой [Kevlin Henney, 2017].

### 5.2. Интерактивная Среда Разработки

Anaconda - это бесплатный и открытый дистрибутив языков программирования Python и R для научных вычислений (наука о данных, приложения для машинного обучения, крупномасштабная обработка данных, прогнозная аналитика и т. Д.), Целью которых является упрощение управления пакетами и развертывание. Дистрибутив включает в себя информационные пакеты, подходящие для Windows, Linux и macOS. Он разработан и поддерживается компанией Anaconda, Inc., которая была основана Питером Вангом и Трэвисом Олифантом в 2012 году. Как продукт Anaconda, Inc., он также известен как Anaconda Distribution или Anaconda Individual Edition, в то время как другие продукты компании - Anaconda Team Edition и Anaconda Enterprise Edition, которые не являются бесплатными.

Версии пакетов в Anaconda управляются системой управления пакетами conda. Этот менеджер пакетов был выделен как отдельный пакет с открытым исходным кодом, так как он оказался полезным сам по себе и для других целей, кроме Python. Существует также небольшая загрузочная версия Anaconda под названием Miniconda, которая включает в себя только conda, Python, пакеты, от которых они зависят, и небольшое количество других пакетов.

## 5.3. Выбранные библиотеки

NumPy - это хорошо известный универсальный пакет для обработки массивов. Обширный набор математических функций высокой сложности делает NumPy мощным средством обработки больших многомерных массивов и матриц. NumPy очень полезен для обработки линейной алгебры, преобразований Фурье и случайных чисел. Другие библиотеки, такие как TensorFlow, используют NumPy в бэкэнде для манипулирования тензорами.

С NumPy вы можете определять произвольные типы данных и легко интегрироваться с большинством баз данных. NumPy также может служить эффективным многомерным контейнером для любых общих данных любого типа данных. Ключевые особенности NumPy включают мощный N-мерный объект массива, функции трансляции и готовые инструменты для интеграции кода C / C ++ и Fortran.2.3.2 TensorFlow

В компьютерном программировании pandas - это библиотека программного обеспечения, написанная для языка программирования Python для обработки и анализа данных. В частности, он предлагает структуры данных и операции для работы с числовыми таблицами и временными рядами. Это бесплатное программное обеспечение, выпущенное под лицензией BSD с тремя пунктами. Название происходит от термина «групповые данные», эконометрического термина для наборов данных, которые включают наблюдения за несколькими периодами времени для одних и тех же людей.
Возможности библиотеки предоставляют:

- Объект DataFrame для манипулирования данными со встроенной индексацией.
- Инструменты для чтения и записи данных между структурами данных в памяти и различными форматами файлов.

- Выравнивание данных и интегрированная обработка отсутствующих данных.
- Изменение формы и поворот наборов данных.
- Срезы на основе меток, необычное индексирование и подмножество больших наборов данных.
- Вставка и удаление столбца структуры данных.
- Группирование по движку, позволяющее выполнять операции разделения-применения-объединения над наборами данных
- Слияние и объединение данных.
- Индексация по иерархической оси для работы с многомерными данными в низкоразмерной структуре данных.
- Функциональность временных рядов: генерация диапазона дат и преобразование частоты, статистика движущегося окна, линейные регрессии движущегося окна, сдвиг даты и отставание.

# 6. Результаты

## 6.1. Интерпретация индикаторов

В таблицах 4 и 5 показано ежегодное количество публикаций и цитирований, соответственно, относительно основных задач, решаемых в области NLP согласно eLIBRARY.ru .

| category | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fuzzy logic | 2.0 | 3.0 | 6.0 | 7.0 | 4.0 | 9.0 | 7.0 | 13.0 | 12.0 | 5.0 | 6.0 | 19.0 | 14.0 |
| Grammar checking | 3.0 | 2.0 | 1.0 | 3.0 | 15.0 | 1.0 | 4.0 | 3.0 | 5.0 | 6.0 | 3.0 | 8.0 | 10.0 |
| Knowledge representation | 30.0 | 40.0 | 49.0 | 73.0 | 48.0 | 93.0 | 72.0 | 96.0 | 111.0 | 119.0 | 175.0 | 221.0 | 267.0 |
| NLP & Automatic summariz | 13.0 | 10.0 | 23.0 | 25.0 | 22.0 | 43.0 | 20.0 | 36.0 | 54.0 | 62.0 | 58.0 | 69.0 | 57.0 |
| NLP & Deep Learning | 8.0 | 11.0 | 13.0 | 17.0 | 15.0 | 26.0 | 28.0 | 27.0 | 40.0 | 60.0 | 102.0 | 158.0 | 236.0 |
| NLP & Dialog Systems | 14.0 | 16.0 | 16.0 | 19.0 | 15.0 | 35.0 | 30.0 | 38.0 | 39.0 | 41.0 | 37.0 | 51.0 | 58.0 |
| NLP & Information Extraction | 76.0 | 91.0 | 98.0 | 133.0 | 126.0 | 212.0 | 156.0 | 205.0 | 273.0 | 264.0 | 334.0 | 373.0 | 388.0 |
| NLP & Information Retrieval | 91.0 | 94.0 | 131.0 | 140.0 | 145.0 | 200.0 | 152.0 | 198.0 | 260.0 | 282.0 | 311.0 | 344.0 | 356.0 |
| NLP & Machine Learning | 70.0 | 87.0 | 106.0 | 131.0 | 124.0 | 207.0 | 174.0 | 225.0 | 283.0 | 305.0 | 386.0 | 505.0 | 521.0 |
| NLP & Machine Translation | 22.0 | 51.0 | 39.0 | 68.0 | 48.0 | 109.0 | 65.0 | 101.0 | 128.0 | 133.0 | 141.0 | 218.0 | 162.0 |
| NLP & Opinion Mining | 4.0 | 3.0 | 9.0 | 13.0 | 10.0 | 33.0 | 26.0 | 44.0 | 64.0 | 67.0 | 82.0 | 111.0 | 137.0 |
| NLP & Question Answering | 28.0 | 28.0 | 53.0 | 43.0 | 30.0 | 53.0 | 38.0 | 54.0 | 55.0 | 66.0 | 80.0 | 115.0 | 104.0 |
| NLP & Sentiment analysis | 2.0 | 2.0 | 5.0 | 9.0 | 11.0 | 34.0 | 28.0 | 48.0 | 75.0 | 96.0 | 125.0 | 188.0 | 225.0 |
| NLP & Speech Recognition | 27.0 | 30.0 | 39.0 | 57.0 | 52.0 | 91.0 | 72.0 | 95.0 | 96.0 | 117.0 | 140.0 | 175.0 | 165.0 |
| NLP & Statistical methods | 72.0 | 65.0 | 99.0 | 97.0 | 93.0 | 166.0 | 125.0 | 151.0 | 201.0 | 193.0 | 226.0 | 268.0 | 267.0 |
| NLP & Text Categorization | 29.0 | 17.0 | 33.0 | 25.0 | 34.0 | 47.0 | 39.0 | 65.0 | 74.0 | 68.0 | 89.0 | 83.0 | 104.0 |
| Neural networks | 17.0 | 27.0 | 23.0 | 30.0 | 37.0 | 58.0 | 44.0 | 60.0 | 77.0 | 88.0 | 146.0 | 231.0 | 323.0 |
| Ontology | 72.0 | 65.0 | 99.0 | 97.0 | 93.0 | 166.0 | 125.0 | 151.0 | 201.0 | 193.0 | 226.0 | 268.0 | 267.0 |
| Rule based system | 31.0 | 51.0 | 52.0 | 75.0 | 74.0 | 117.0 | 80.0 | 108.0 | 118.0 | 161.0 | 185.0 | 212.0 | 221.0 |
| Supervised learning | 11.0 | 11.0 | 16.0 | 22.0 | 38.0 | 67.0 | 53.0 | 81.0 | 99.0 | 96.0 | 135.0 | 185.0 | 185.0 |

*Таблица 4. Количество публикаций*

В настоящее время количество публикаций и цитирований показывает устойчивый рост практически во всех рассматриваемых разделах. Использование дифференциальных индикаторов D1 и D2 позволяет нам более четко представить динамику этого роста. Использование выражений 5,6 иллюстрируется на рисунках [9-14, 17-19, 21,24]. Стоит отметить, что из-за особенностей поисковых запросов данные, представленные в таблице, вероятно, отражают только некоторые из доступных публикаций в русскоязычном сегменте.

| category | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fuzzy logic | 9.0 | 16.0 | 12.0 | 11.0 | 12.0 | 6.0 | 15.0 | 21.0 | 16.0 | 28.0 | 26.0 | 40.0 | 42.0 |
| Grammar checking | 21.0 | 15.0 | 20.0 | 27.0 | 32.0 | 27.0 | 32.0 | 36.0 | 45.0 | 46.0 | 41.0 | 62.0 | 46.0 |
| Knowledge representation | 34.0 | 54.0 | 32.0 | 39.0 | 49.0 | 38.0 | 52.0 | 57.0 | 50.0 | 51.0 | 53.0 | 62.0 | 66.0 |
| NLP & Automatic summariz | 2.0 | 7.0 | 10.0 | 2.0 | 5.0 | 4.0 | 8.0 | 11.0 | 9.0 | 8.0 | 13.0 | 10.0 | 12.0 |
| NLP & Deep Learning | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 2.0 | 1.0 | 0.0 | 2.0 | 10.0 | 22.0 | 72.0 | 114.0 |
| NLP & Dialog Systems | 19.0 | 19.0 | 9.0 | 5.0 | 9.0 | 10.0 | 6.0 | 7.0 | 7.0 | 4.0 | 8.0 | 14.0 | 7.0 |
| NLP & Information Extraction | 21.0 | 42.0 | 31.0 | 34.0 | 46.0 | 48.0 | 47.0 | 64.0 | 66.0 | 69.0 | 61.0 | 83.0 | 108.0 |
| NLP & Information Retrieval | 49.0 | 90.0 | 73.0 | 76.0 | 92.0 | 67.0 | 99.0 | 118.0 | 138.0 | 167.0 | 205.0 | 236.0 | 228.0 |
| NLP & Machine Learning | 49.0 | 69.0 | 67.0 | 71.0 | 105.0 | 77.0 | 108.0 | 125.0 | 169.0 | 232.0 | 277.0 | 385.0 | 451.0 |
| NLP & Machine Translation | 14.0 | 40.0 | 15.0 | 20.0 | 17.0 | 22.0 | 20.0 | 34.0 | 46.0 | 48.0 | 48.0 | 61.0 | 66.0 |
| NLP & Opinion Mining | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 3.0 | 9.0 | 9.0 | 12.0 | 37.0 | 36.0 | 57.0 | 50.0 |
| NLP & Question Answering | 15.0 | 32.0 | 26.0 | 22.0 | 20.0 | 18.0 | 22.0 | 37.0 | 44.0 | 41.0 | 45.0 | 49.0 | 58.0 |
| NLP & Sentiment analysis | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 | 3.0 | 9.0 | 25.0 | 18.0 | 49.0 | 67.0 | 96.0 | 102.0 |
| NLP & Speech Recognition | 38.0 | 52.0 | 32.0 | 31.0 | 38.0 | 35.0 | 23.0 | 43.0 | 42.0 | 52.0 | 74.0 | 85.0 | 116.0 |
| NLP & Statistical methods | 21.0 | 46.0 | 11.0 | 26.0 | 29.0 | 21.0 | 29.0 | 38.0 | 36.0 | 55.0 | 58.0 | 74.0 | 83.0 |
| NLP & Text Categorization | 5.0 | 10.0 | 5.0 | 15.0 | 18.0 | 11.0 | 17.0 | 19.0 | 18.0 | 33.0 | 31.0 | 30.0 | 34.0 |
| Neural networks | 41.0 | 52.0 | 32.0 | 48.0 | 54.0 | 34.0 | 53.0 | 74.0 | 67.0 | 112.0 | 122.0 | 182.0 | 237.0 |
| Ontology | 43.0 | 70.0 | 51.0 | 51.0 | 69.0 | 61.0 | 72.0 | 110.0 | 123.0 | 136.0 | 164.0 | 171.0 | 195.0 |
| Rule based system | 6.0 | 15.0 | 3.0 | 10.0 | 15.0 | 8.0 | 8.0 | 16.0 | 17.0 | 22.0 | 22.0 | 26.0 | 31.0 |
| Supervised learning | 10.0 | 14.0 | 11.0 | 18.0 | 18.0 | 14.0 | 31.0 | 42.0 | 40.0 | 66.0 | 71.0 | 117.0 | 142.0 |

*Таблица 5. Количество цитирований*

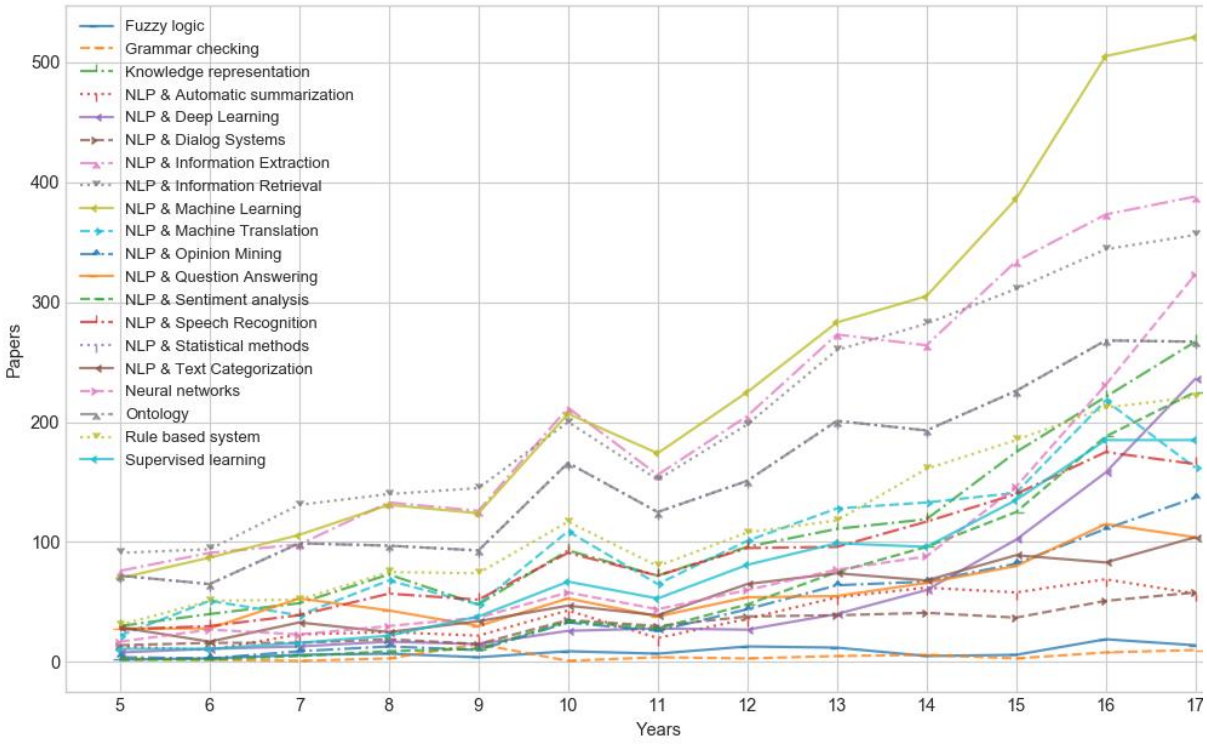



*Рисунок 8. Количество публикаций в год*

Построены графики количества публикаций и цитирований в разрезе года выпуска и раздела. Между данными можно сгенерировать дополнительные точки, затем интерполировать и построить регрессию для более гладкой картинки.

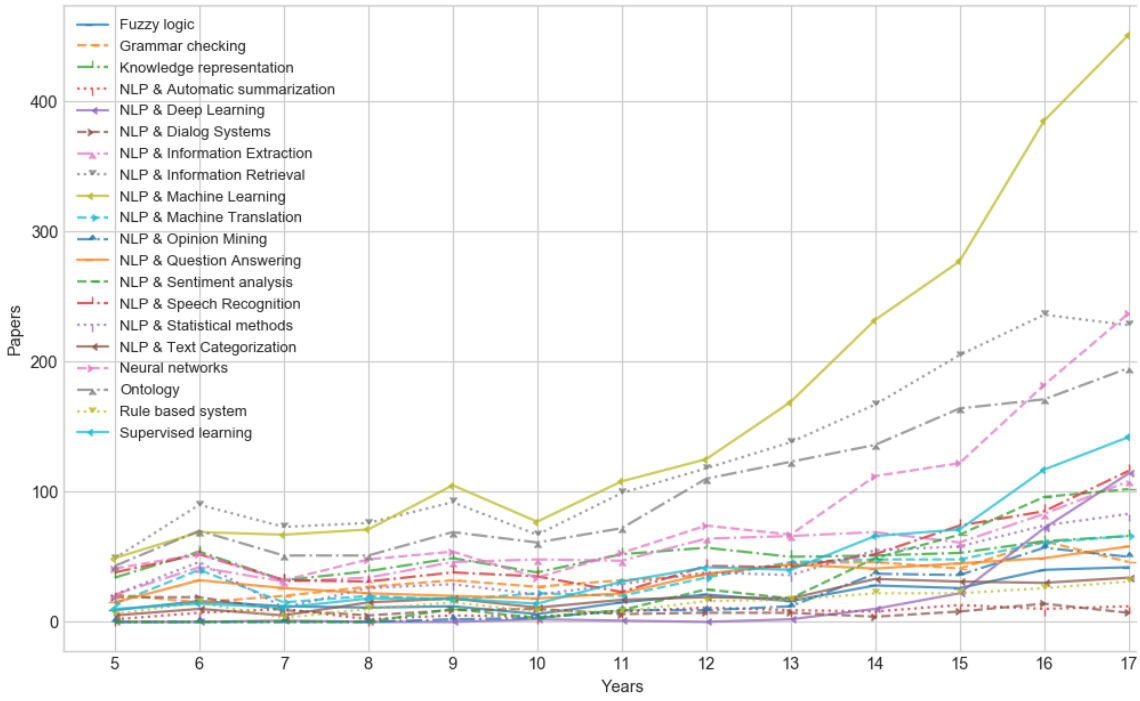



*Рисунок 9. Количество цитирований в год*

Увеличение индикатор D1 указывает на рост в целом а D2 увеличение темпа, то есть скорости изменения количества публикаций и цитат в области исследований. Негативная динамика, в свою очередь, показывает замедление роста по сравнению с предыдущими периодами. Для некоторых разделов замечены колебания, где за начальным увеличением следует замедление, а затем происходит повторное ускорение.

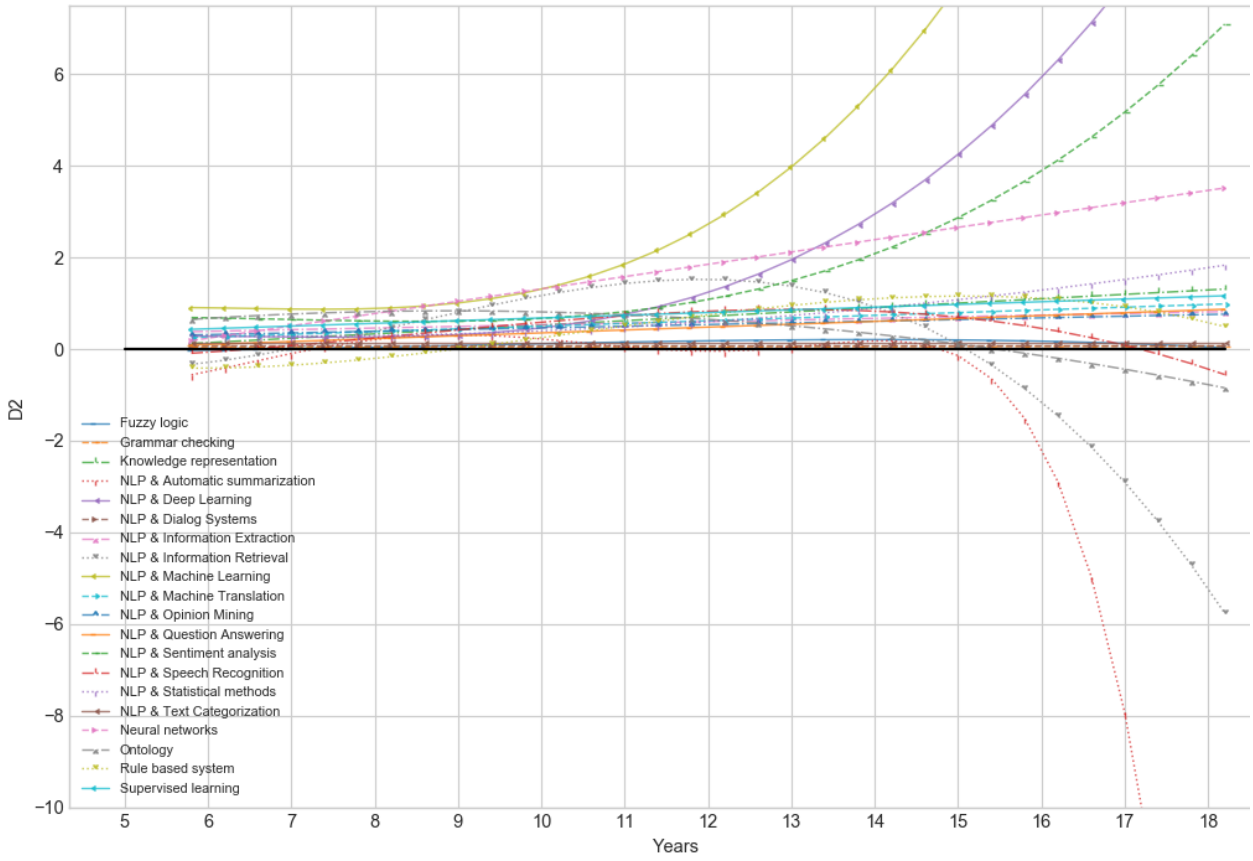



*Рисунок 10. Скорость изменения количества публикаций*

Можно предположить, что эта динамика характеризует интенсивность развития в области исследований, а также принятие новой концепции исследователями и ее применение в исследованиях. Несколько из представленных разделов (глубокое обучение, анализ тональностей, нейронные сети, машинное обучение) характеризуются постоянным увеличением цитирования и публикаций (показатель D2 имеет только положительные значения за весь рассматриваемый период).

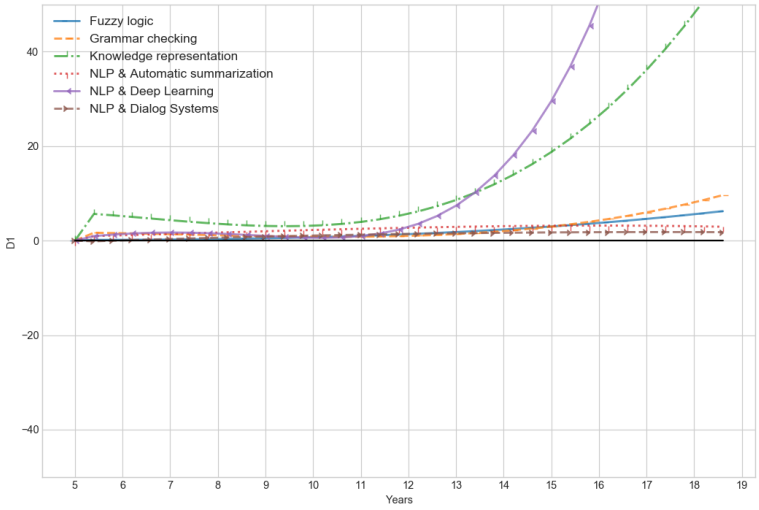



*Рисунок 11 Скорость изменения количества публикаций*

Разделы (анализ тональностей, глубокое обучение), как упоминалось выше, характеризуются резким ростом индикаторов, тогда как для разделов, связанных с поиском информации, категоризацией текста и автоматическими сводками по тексту, онтология показывают снижение индикаторов. В целом, анализ исследования NLP показывает стабильное увеличение динамики публикационной активности в областях: статистических методов, извлечения информации, машинного перевода, нечеткой логики.

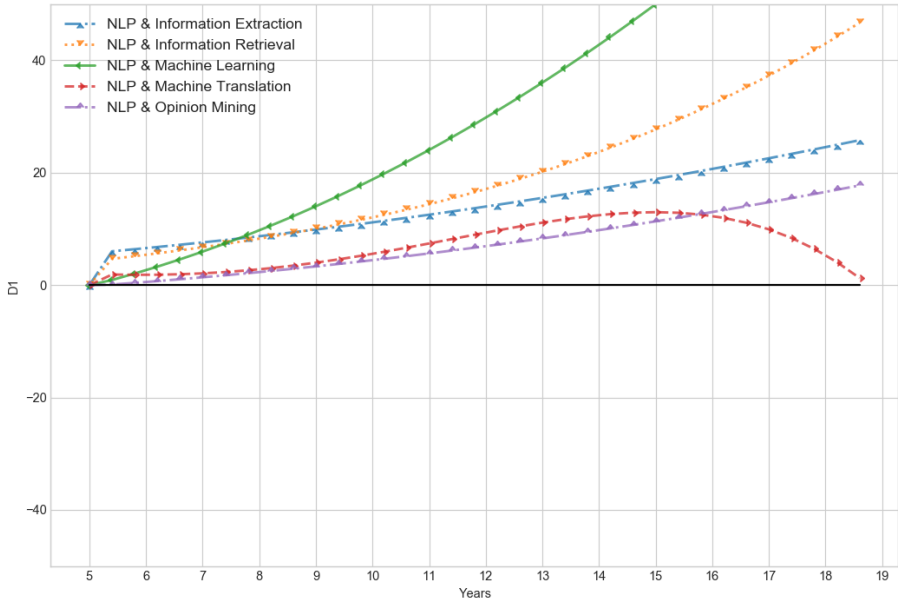



*Рисунок 12. Скорость изменения количества публикаций*

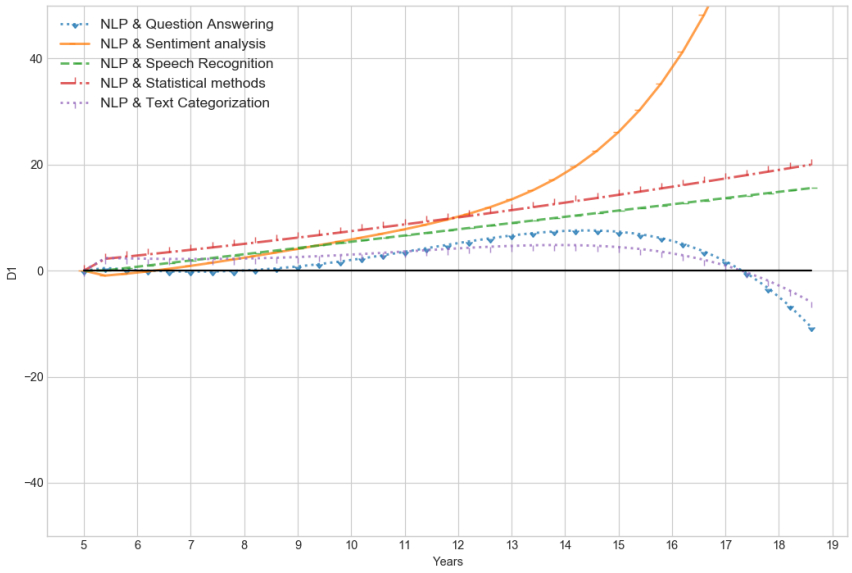



*Рисунок 13. Скорость изменения количества публикаций*

По графику ниже можно заметить увеличение скорости количества публикаций в таких разделах как нейронные сети, и обучение с учителем. Напротив, системы основанные на правилах и онтология показывая рост до 2015 года, в последующих поколениях рост начинает замедляться, что свидетельствует об угасающем интересе.

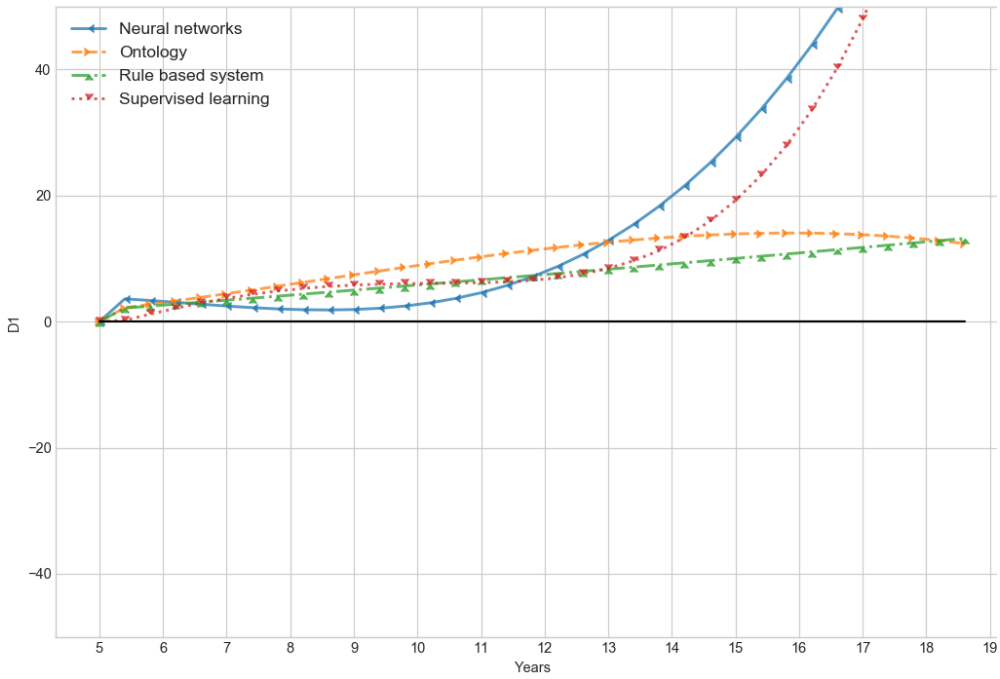



*Рисунок 14. Скорость изменения количества публикаций*

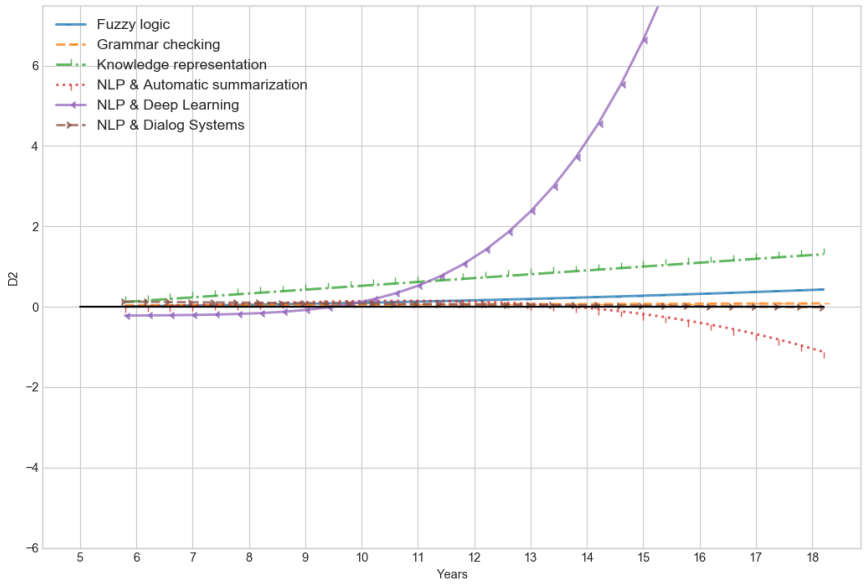



*Рисунок 15. Скорость изменения количества публикаций*

Порядок регрессии был подобран так, чтобы минимизировать ошибку :

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

где $\hat{y}$ это гипотетическое значение полученное из регрессии а $n$ количество наблюдений. Для каждого порядка считается корень суммы квадратов разности ошибок, затем из них выбирается минимальный.

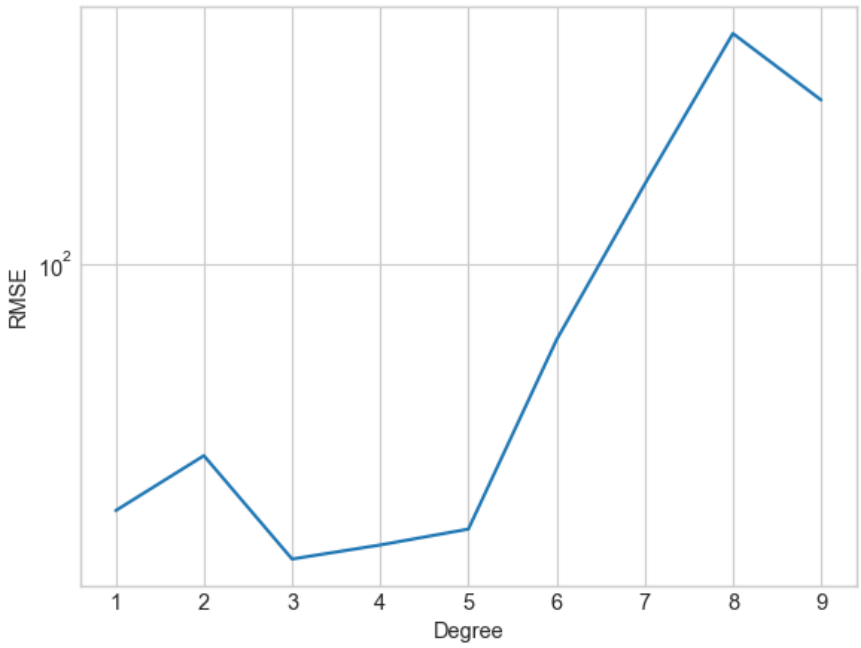



*Рисунок 16. График ошибки по порядку регрессии*

По графику заметно, что минимум ошибки достигается при регрессии 3 порядка, на которой мы остановимся. Этот шаг проделывается для каждого подраздела научной сферы.

```
Best degree 2 with RMSE 6.1946981783694355 for Fuzzy logic
Best degree 4 with RMSE 6.18723832619838 for Fuzzy logic
Best degree 4 with RMSE 2.713087076227063 for Grammar checking
Best degree 4 with RMSE 4.098521038107589 for Grammar checking
Best degree 4 with RMSE 18.873342672029022 for Knowledge representation
Best degree 2 with RMSE 9.116526099504837 for Knowledge representation
Best degree 2 with RMSE 12.40658419500485 for NLP & Automatic summarization
Best degree 4 with RMSE 3.2072126888607033 for NLP & Automatic summarization
Best degree 6 with RMSE 3.8428542746016756 for NLP & Deep Learning
Best degree 5 with RMSE 3.0623185108015805 for NLP & Deep Learning
Best degree 3 with RMSE 0.9302806722263526 for NLP & Dialog Systems
Best degree 2 with RMSE 2.9719031708205974 for NLP & Dialog Systems
```

*Рисунок 17. Расчет ошибки*

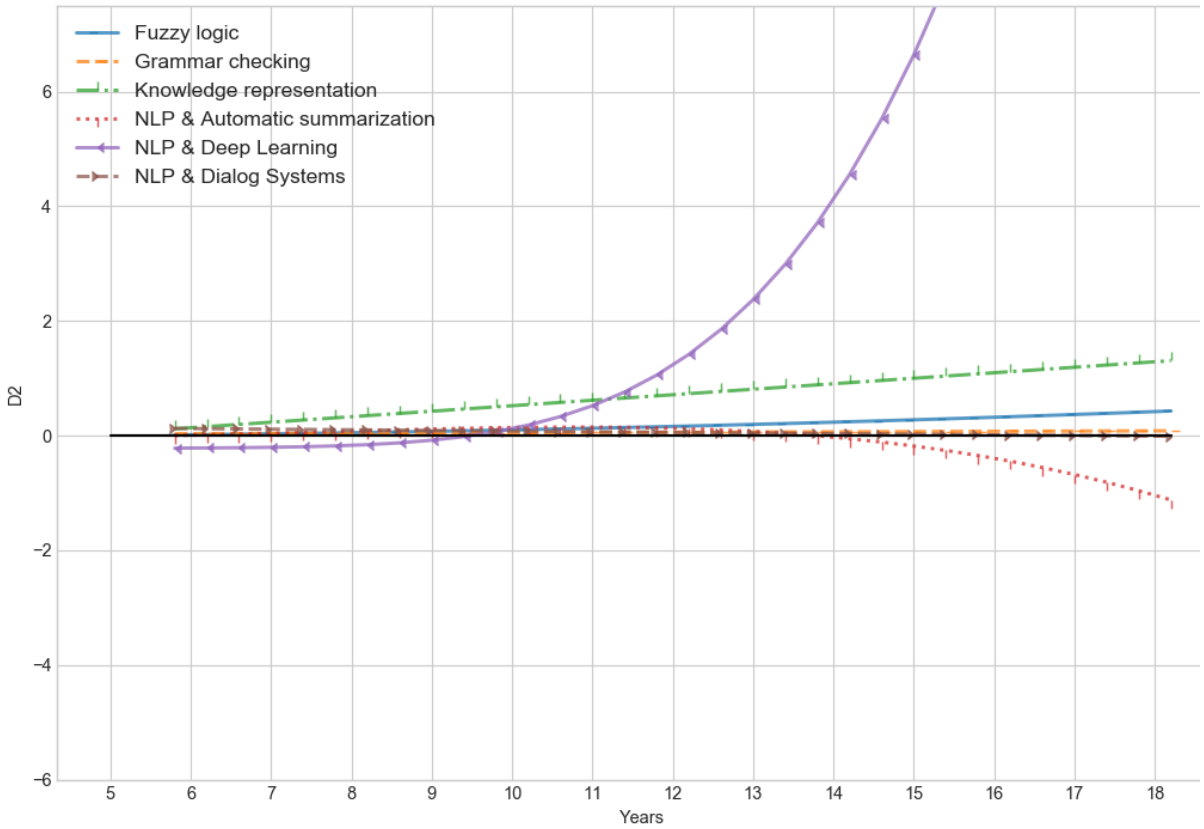



*Рисунок 18. D2 индикатор*

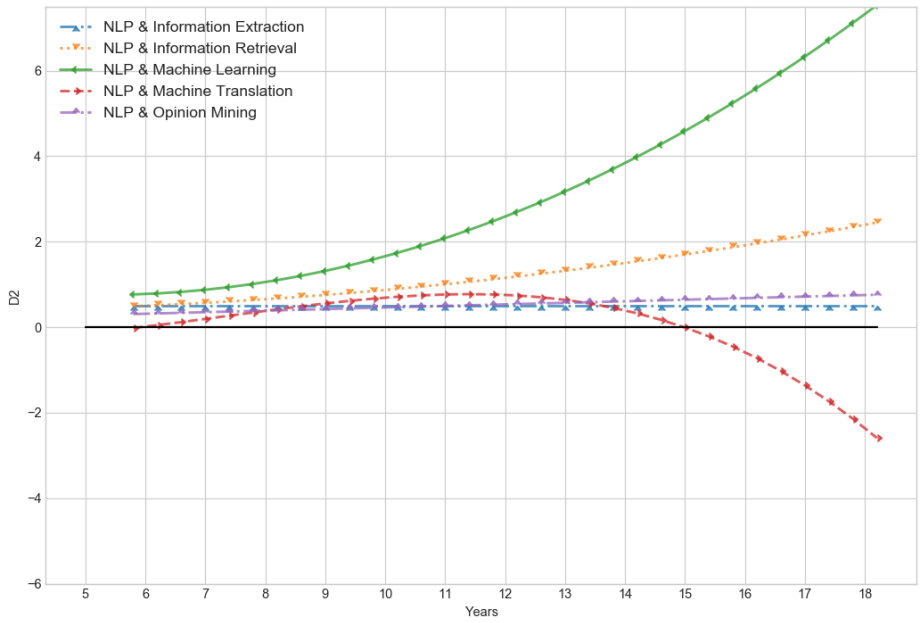



*Рисунок 19. D2 индикатор*

По графику можно заметить, что среди данных разделов стремительный рост в ускорении имеется у машинного обучения, что говорит о высокой перспективности данного раздела. Напротив, машинный перевод показывает снижение индикатора D2, что говорит об угасающем интересе, по крайней мере до 2015 года рост имел положительный характер. Разделы связанные с извлечением информации и анализом мнений показывают стабильное ускорение, но не явно позволяет судить о возможной перспективности. По разделам статистических методов, существенных изменений в ускорении не наблюдаются.

```
Best degree 2 with RMSE 31.665300416052215 for NLP & Information Extraction
Best degree 2 with RMSE 7.754701867873499 for NLP & Information Extraction
Best degree 2 with RMSE 9.28961807745307 for NLP & Information Retrieval
Best degree 4 with RMSE 12.586580883207603 for NLP & Information Retrieval
Best degree 2 with RMSE 15.866721736471895 for NLP & Machine Learning
Best degree 4 with RMSE 15.841236933452546 for NLP & Machine Learning
Best degree 3 with RMSE 20.72835891839442 for NLP & Machine Translation
Best degree 5 with RMSE 9.91459788332195 for NLP & Machine Translation
Best degree 2 with RMSE 6.708208493612333 for NLP & Opinion Mining
Best degree 3 with RMSE 1.6237127554803155 for NLP & Opinion Mining
```

*Рисунок 20. Ошибка*

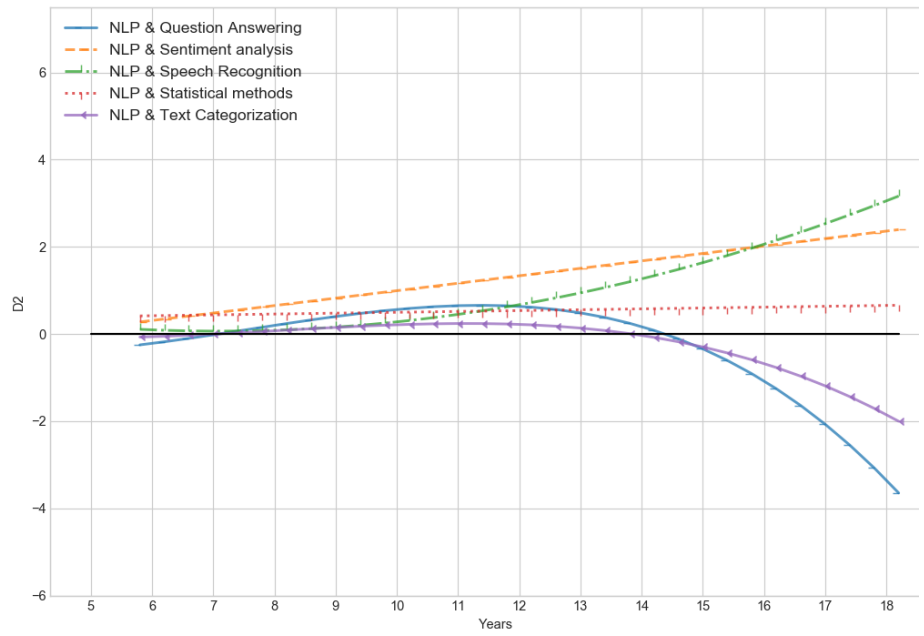



*Рисунок 21. D2 индикатор*

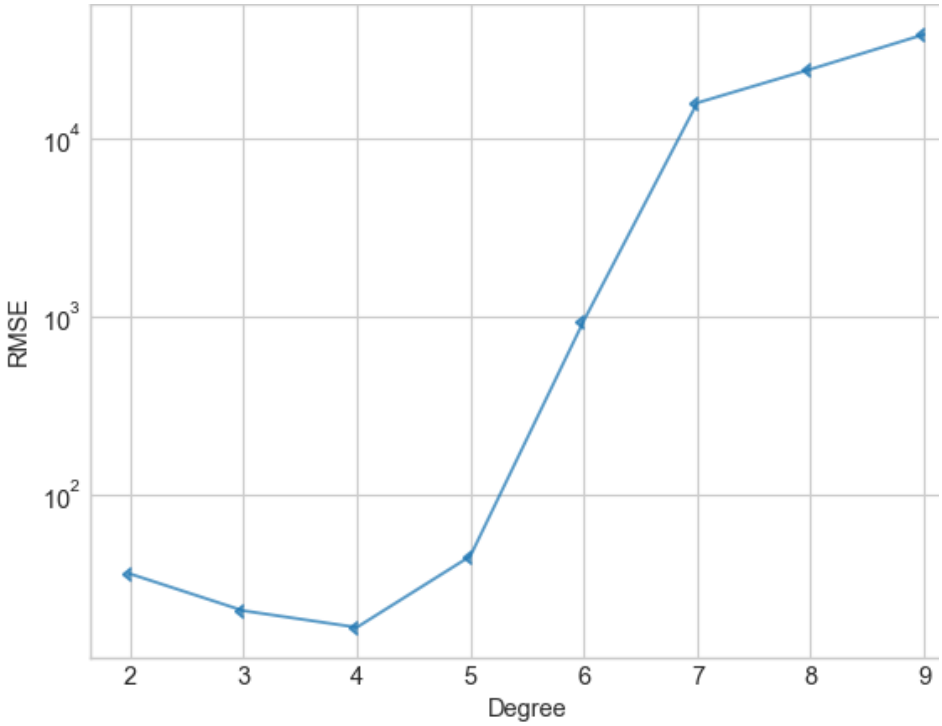



*Рисунок 22. Ошибка*

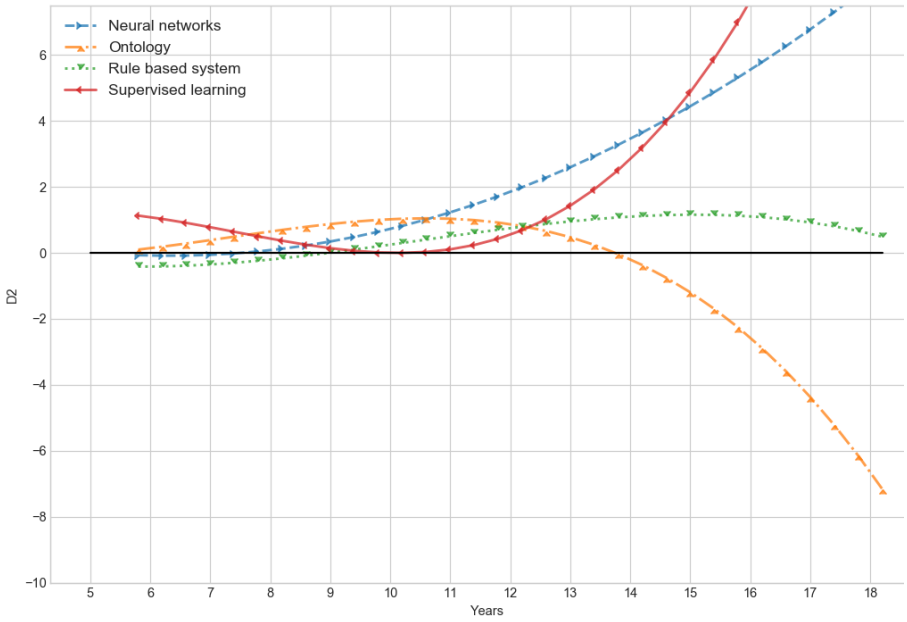



*Рисунок 23. D2 индикатор*

## 6.2. Заключение

Для получения адекватного прогноза перспективности следует корректно подобрать порядок регрессии и весовые коэффициенты. Как уже было описано, порядок подбирается такой, что ошибка RMSE при этом минимальная. Весовые коэффициенты $\beta, \gamma$ были эмпирическим путем подобраны в [32] и приравнены к 0,95. Таким образом можно проанализировать и объяснить рост публикационной активности на основе дифференциальных показателей[2]: изменения скорости D1 и ускорения D2. Положительное значение D2 отражает факт увеличения, а D1 характеризует темпы роста публикационной активности в области исследований. С другой стороны, отрицательное значение D2 указывает на замедление публикационной активности по сравнению с предыдущими периодами. По этим показателям мы можем предположить растущий интерес к таким областям NLP, как анализ тональности, нейронные сети, распознавание речи, представления знаний , стабильность интереса к нечеткой логике, статистическим методам, и снижающийся интерес к категоризации текста, ответы на вопросы, автоматические сводки текста и т.д. Такие дифференциальные показатели позволяют выявлять тенденции интереса в различных областях исследований с более объективной точки зрения. В этих показателях не столько важно численное значение индикаторов сколько их знак , и наклон касательной определяющие направление и темп развития области. Данные методы могут быть экстраполированы на другие отрасли научных исследований связанных с например с медициной, геологией и тд.

Основываясь на данных дифференциальных индикаторах выявлены наиболее быстро развивающиеся области NLP:

- Нейронные сети
- Распознавание речи
- Анализ тональности и другие;

Области со снижающейся публикационной активностью:

- Категоризация текста
- Автоматические реферирование
- Онтология и другие;

Области со стабильной динамикой развития

- Статистические методы
- Извлечение информации
- Нечеткая логика и другие.

Точность данного метода зависит от базы данных и ограничена ее объемами. Несколько исследований показали, что покрытие статей в Web of Science и Scopus значительно отличаются в зависимости от научной области, например, если в естественных науках покрытие хорошее, то в социальных науках и в искусстве намного ниже[43].

# 7. Список использованной литературы


1. Hirsch J. E. An index to quantify an individual's scientific research output //Proceedings of the National academy of Sciences. – 2005. – Т. 102. – №. 46. – С. 16569-16572.
2. Muhamedyev R. et al. New bibliometric indicators for prospectivity estimation of research fields //Annals of Library and Information Studies (ALIS). – 2018. – Т. 65. – №. 1. – С. 62-69.
3. Garfield E. Citation indexes for science //Science. – 1955. – Т. 122. – №. 3159. – С. 108-111.
4. Van Raan A. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments //TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis. – 2003. – Т. 12. – №. 1. – С. 20-29.
5. Mokhnacheva Y. et al. Bibliometric analysis of patent and document information flows of Moscow Region organizations in the nanotechnological sphere //NAUCHNYE I TEKHNICHESKIE BIBLIOTEKI-SCIENTIFIC AND TECHNICAL LIBRARIES. – 2016. – №. 2. – С. 55-69.
6. Abramo G., D'Angelo C., Pugini F. The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology //Scientometrics. – 2008. – Т. 76. – №. 2. – С. 225-244.
7. Debackere K., Glänzel W. Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key //Scientometrics. – 2004. – Т. 59. – №. 2. – С. 253-276.
8. Daim T. U., Rueda G. R., Martin H. T. Technology forecasting using bibliometric analysis and system dynamics //A Unifying Discipline for Melting the Boundaries Technology Management:. – IEEE, 2005. – С. 112-122.
9. EV C. M. V. Van Der Weyden MB. Life and times of the impact factor: retrospective analysis of trends for seven medical journals (1994-2005) and their editors* views //JR Soc Med. – 2007. – С. 142-150.

10. Lundberg J. Lifting the crown—citation z-score //Journal of informetrics. – 2007. – T. 1. – №. 2. – C. 145-154.
11. Garfield E. The history and meaning of the journal impact factor //jama. – 2006. – T. 295. – №. 1. – C. 90-93.
12. Gauthier É. STATISTICS CANADA, S., TECHNOLOGY REDESIGN, P., STATISTICS CANADA. SCIENCE, I. & ELECTRONIC INFORMATION, D.(1998) Bibliometric analysis of scientific and technological research: A user's guide to the methodology //Science and Technology Redesign Project, Statistics Canada.
13. Bornmann L., Daniel H. D. What do citation counts measure? A review of studies on citing behavior //Journal of documentation. – 2008.
14. Garfield E. Is citation analysis a legitimate evaluation tool? //Scientometrics. – 1979. – T. 1. – №. 4. – C. 359-375.
15. http: //findarticles.com/p/articles/mi_m1387/ is_3_50/ai_88582623.
16. Abramo G., D'Angelo C. A., Cicero T. What is the appropriate length of the publication period over which to assess research performance? //Scientometrics. – 2012. – T. 93. – №. 3. – C. 1005-1017.
17. Wang J. Citation time window choice for research impact evaluation //Scientometrics. – 2013. – T. 94. – №. 3. – C. 851-872..
18. Lundberg J. Bibliometrics as a research assessment tool: impact beyond the impact factor. – Institutionen för lärande, informatik, management och etik, LIME/Department of Learning, Informatics, Management and Ethics (Lime), 2006.
19. http://www.inria. fr/inria/organigramme/documents/ce indicateurs.pdf .
20. Moed H., De Bruin R., Van Leeuwen T. H. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications //Scientometrics. – 1995. – T. 33. – №. 3. – C. 381-422.

21. Garfield E. Citation indexes for science. A new dimension in documentation through association of ideas //International journal of epidemiology. – 2006. – Т. 35. – №. 5. – С. 1123-1127.
22. Durieux V., Gevenois P. A. Bibliometric indicators: quality measurements of scientific publication //Radiology. – 2010. – Т. 255. – №. 2. – С. 342-351.
23. Lundberg J. Lifting the crown—citation z-score //Journal of informetrics. – 2007. – Т. 1. – №. 2. – С. 145-154.
24. http://scientifi c.thomsonreuters. com/support/patents/patinf/terms
25. Chew F. S., Relyea-Chew A. How research becomes knowledge in radiology: an analysis of citations to published papers //American Journal of Roentgenology. – 1988. – Т. 150. – №. 1. – С. 31-37.
26. Bergstrom C. T., West J. D. Assessing citations with the Eigenfactor™ metrics. – 2008.
27. Bergstrom C. T., West J. D., Wiseman M. A. The eigenfactor™ metrics //Journal of neuroscience. – 2008. – Т. 28. – №. 45. – С. 11433-11434.
28. http://www.eigenfactor.org/methods.html
29. Batista P. D., Campiteli M. G., Kinouchi O. Is it possible to compare researchers with different scientific interests? //Scientometrics. – 2006. – Т. 68. – №. 1. – С. 179-189.
30. https://stats.oecd.org/glossary/detail.asp?ID=198
31. Ronald R., Fred Y. Journal of the Association for Information Science and Technology. – 1998. -№59.11. – p.1853–55
32. Barakhnin V. B. et al. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area //Journal of Physics: Conference Series. – 2018. – Т. 1117. .
33. Kosyakov Denis. The State Public Scientific Technological Library of Siberian Branch of the Russian Academy
34. Qasem A. et al. 2007 Index IEEE Transactions on Knowledge and Data Engineering Vol. 19.
35. Sun S., Luo C., Chen J. Information Fusion. – 2017. - №36. – p.10–25

36. Goldberg Y.Journal of Artificial Intelligence Research. – 2016. - №57 - p. 345–420
37. Dale R. NLP meets the cloud //Natural Language Engineering. – 2015. – Т. 21. – №. 4. – С. 653-659.
38. Zamanov I. et al. Voltron: A hybrid system for answer validation based on lexical and distance features //Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). – 2015. – С. 242-246.
39. Chen Y., Conroy N. J., Rubin V. L. News in an online world: the need for an " automatic crap detector&quot //Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community. – 2015.
40. Wikipedia.org Natural language
41. LeCun Y., Bengio Y., Hinton G. Deep learning //nature. – 2015. – Т. 521. – №. 7553. – С. 436-444.
42. Hogenboom F. et al. A survey of event extraction methods from text for decision support systems //Decision Support Systems. – 2016. – Т. 85. – С. 12-22.

Potthast M., Hagen M., Stein B. Author Obfuscation: Attacking the State of the Art in Authorship Verification //CLEF (Working Notes). – 2016. – С. 716-749.

43. Larivière V. et al. The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities //Journal of the American Society for Information Science and Technology. – 2006. – Т. 57. – №. 8. – С. 997-1004.

# 8. Приложения

## 8.1. Загрузка данных

```
#%matplotlib notebook
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
csv_name='BOOK3.xlsx'
sterm='Search terms'
sterm='category'
sc_df = pd.read_excel(csv_name, encoding='latin1')
keys=np.array(sc_df.keys())
ka=np.array(keys.shape)
labels=np.array(sc_df[sterm])
sc_df.head()
```

- Транспонируем и группируем данные в разрезе года и научного раздела

```
NLP_N_TEXT=sc_df.groupby(['category','year']).count()
NLP_N_TEXT_cit=sc_df.groupby(['category','year']).sum()
transposed=NLP_N_TEXT.unstack('year')
transposed_cit=NLP_N_TEXT_cit.unstack('year')
transposed_cit
transposed
```

- На выходе получаем следующую табличку

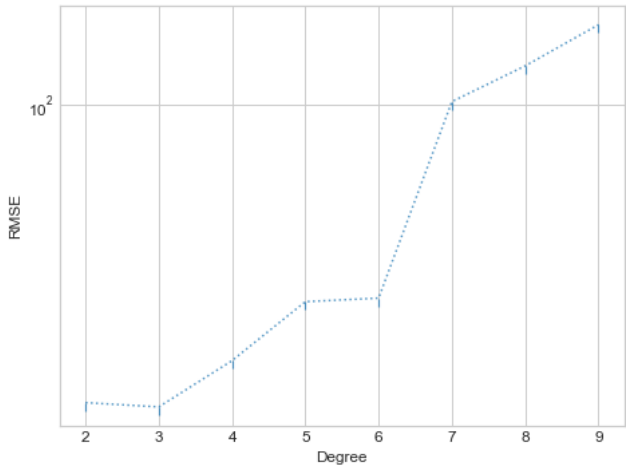
| | cites | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| year | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| category | | | | | | | | | | | | | |
| Fuzzy logic | 2.0 | 3.0 | 6.0 | 7.0 | 4.0 | 9.0 | 7.0 | 13.0 | 12.0 | 5.0 | 6.0 | 19.0 | 14.0 |
| Grammar checking | 3.0 | 2.0 | 1.0 | 3.0 | NaN | 1.0 | 4.0 | 3.0 | 5.0 | 6.0 | 3.0 | 8.0 | 10.0 |
| Knowledge representation | 30.0 | 40.0 | 49.0 | 73.0 | 48.0 | 93.0 | 72.0 | 96.0 | 111.0 | 119.0 | 175.0 | 221.0 | 267.0 |
| NLP & Automatic summarization | 13.0 | 10.0 | 23.0 | 25.0 | 22.0 | 43.0 | 20.0 | 36.0 | 54.0 | 62.0 | 58.0 | 69.0 | 57.0 |
| NLP & Deep Learning | 8.0 | 11.0 | 13.0 | 17.0 | 15.0 | 26.0 | 28.0 | 27.0 | 40.0 | 60.0 | 102.0 | 158.0 | 236.0 |
| NLP & Dialog Systems | 14.0 | 16.0 | 16.0 | 19.0 | 15.0 | 35.0 | 30.0 | 38.0 | 39.0 | 41.0 | 37.0 | 51.0 | 58.0 |
| NLP & Information Extraction | 76.0 | 91.0 | 98.0 | 133.0 | 126.0 | 212.0 | 156.0 | 205.0 | 273.0 | 264.0 | 334.0 | 373.0 | 388.0 |
| NLP & Information Retrieval | 91.0 | 94.0 | 131.0 | 140.0 | 145.0 | 200.0 | 152.0 | 198.0 | 260.0 | 282.0 | 311.0 | 344.0 | 356.0 |
| NLP & Lexical affinity | NaN | NaN | 1.0 | NaN | 2.0 | 3.0 | NaN | NaN | 1.0 | NaN | NaN | 4.0 | 2.0 |
| NLP & Machine Learning | 70.0 | 87.0 | 106.0 | 131.0 | 124.0 | 207.0 | 174.0 | 225.0 | 283.0 | 305.0 | 386.0 | 505.0 | 521.0 |
| NLP & Machine Translation | 22.0 | 51.0 | 39.0 | 68.0 | 48.0 | 109.0 | 65.0 | 101.0 | 128.0 | 133.0 | 141.0 | 218.0 | 162.0 |
| NLP & Opinion Mining | 4.0 | 3.0 | 9.0 | 13.0 | 10.0 | 33.0 | 26.0 | 44.0 | 64.0 | 67.0 | 82.0 | 111.0 | 137.0 |
| NLP & Question Answering | 28.0 | 28.0 | 53.0 | 43.0 | 30.0 | 53.0 | 38.0 | 54.0 | 55.0 | 66.0 | 80.0 | 115.0 | 104.0 |
| NLP & Sentiment analysis | 2.0 | 2.0 | 5.0 | 9.0 | 11.0 | 34.0 | 28.0 | 48.0 | 75.0 | 96.0 | 125.0 | 188.0 | 225.0 |
| NLP & Speech Recognition | 27.0 | 30.0 | 39.0 | 57.0 | 52.0 | 91.0 | 72.0 | 95.0 | 96.0 | 117.0 | 140.0 | 175.0 | 165.0 |
| NLP & Statistical methods | 72.0 | 65.0 | 99.0 | 97.0 | 93.0 | 166.0 | 125.0 | 151.0 | 201.0 | 193.0 | 226.0 | 268.0 | 267.0 |
| NLP & Text Categorization | 29.0 | 17.0 | 33.0 | 25.0 | 34.0 | 47.0 | 39.0 | 65.0 | 74.0 | 68.0 | 89.0 | 83.0 | 104.0 |
| Neural networks | 17.0 | 27.0 | 23.0 | 30.0 | 37.0 | 58.0 | 44.0 | 60.0 | 77.0 | 88.0 | 146.0 | 231.0 | 323.0 |
| Ontology | 72.0 | 65.0 | 99.0 | 97.0 | 93.0 | 166.0 | 125.0 | 151.0 | 201.0 | 193.0 | 226.0 | 268.0 | 267.0 |
| Rule based system | 31.0 | 51.0 | 52.0 | 75.0 | 74.0 | 117.0 | 80.0 | 108.0 | 118.0 | 161.0 | 185.0 | 212.0 | 221.0 |
| Supervised learning | 11.0 | 11.0 | 16.0 | 22.0 | 38.0 | 67.0 | 53.0 | 81.0 | 99.0 | 96.0 | 135.0 | 185.0 | 185.0 |

## 8.2. Визуализация и интерполяция

```
#Print publication activity
#interctive plot
%matplotlib notebook
import pylab as plb
#plt.ion()
plt.style.use('seaborn-whitegrid')
labels=np.array(sc_df[sterm])
keys=np.array(sc_df.keys())

fig, ax = plt.subplots()
fig.set_size_inches(12,7)
fig.suptitle('Publication activity (number of papers)', fontsize=12)

fig.suptitle('Publication activity (number of papers) ', fontsize=12)

ax.set_xlabel('Years', fontsize='medium' ) #fontsize=10)
ax.set_ylabel('Papers', fontsize='medium')   # relative to plt.rcParams['font.size']

markers=np.array(['o', '.', ',', 'x', '+', 'v', '^', '<', '>', 's', 'd'])
linestyless=np.array(['-','--','-.',':','-','--','-.',':','-','--','-.',':',])
markers_size=markers.size
count=0
for lab in labels:

    ka=np.array(keys.shape)
    scnum=sc_df.loc[count,keys[1:ka[0]]]
    y=np.array(scnum)
    x=keys[1:ka[0]]
    j= count% markers_size
    #print(j)
    alpha=0.75
    ln= linestyless[j]
    ax.plot(x_axis,de)
    ax.plot(x_axis,de,'-p',label=lbl, marker=j, markersize=5,linestyle =ln, linewidth=1,alpha=alpha) #, linestyle=':')
    print(y)
    print(x)
    plt.plot(x,y, label= labels[count])
    plb.plot(x,y, label= labels[count])
    plb.plot(x,y, label= labels[count], marker=j, markersize=4,linestyle =ln, linewidth=1.5,alpha=alpha)
    plt.xticks(np.arange(5, 18, step=1))
    ax.plot(x_axis,de,'.',label=lbl, marker=j, markersize=3)
    count=count+1
    plt.grid(b=True, which='major', color='#666666', linestyle='-')

    #plt.show()
   # print(count)

plt.legend(loc='upper left',prop={'size': 9})
plt.show()
```

```
%matplotlib notebook
count=9      #number of row
dotPerYear=3    #dots per year after regression
print_curves=1  #Yes
degree=1   #regression dergree
predictYears=2  #Years of predictions
markers=np.array(['o', '.', ',', 'x', '+', 'v', '^', '<', '>', 's', 'd'])
linestyless=np.array(['-','--','-.',':','-','--','-.',':','-','--','-.',':',])
markers_size=markers.size
def regCurves3(count,dotPerYear,sc_df,degree,predictYears,print_curves):
    #count=2
    keys=np.array(sc_df.keys())
    ka=np.array(keys.shape)
    #Labels=np.array(sc_df['Field'])
    labels=np.array(sc_df[sterm])
    scnum=sc_df.loc[count,keys[1:ka[0]-1]]
    #print(scnum)
    y=np.array(scnum)
    #x=np.arange(ka[0]-2)
    x=keys[1:ka[0]-1]
    x = [int(i) for i in x]  #convert to integer
    x=np.array(x)
    #x=x-2000
    X=x[:,np.newaxis]

    x_train, x_test, y_train, y_test = train_test_split(X, y, shuffle=True,test_size=0.3)

    rmses = []
    degrees = np.arange(1, 10)
    min_rmse, min_deg = 1e10, 0

    for deg in degrees:

        # Train features
        poly_features = PolynomialFeatures(degree=deg, include_bias=False)
        x_poly_train = poly_features.fit_transform(x_train)

        # Linear regression
        poly_reg = LinearRegression()
        poly_reg.fit(x_poly_train, y_train)

        # Compare with test data
        x_poly_test = poly_features.fit_transform(x_test)
        poly_predict = poly_reg.predict(x_poly_test)
        poly_mse = mean_squared_error(y_test, poly_predict)
        poly_rmse = np.sqrt(poly_mse)
        rmses.append(poly_rmse)

        # Cross-validation of degree
        if min_rmse > poly_rmse:
            min_rmse = poly_rmse
            min_deg = deg

    # Plot and present results
    print('Best degree {} with RMSE {} for {}'.format(min_deg, min_rmse,labels[count]))
```

Выборка разделена на обучающую и тестовую. Спрогнозированы последующие два года. Для выбора наиболее подходящего порядка регрессии, были построены регрессии до 10 порядка, затем посчитаны RMSE. Для каждой линии регрессии порядок был выбран в соответствии с минимальным RMSE.




На графике изображена линия регрессии и истинные значения с таблицы

```
    fig = plt.figure()
    s= count% markers_size
    alpha=0.75
    ln= linestyless[s]
    ax = fig.add_subplot(111)
    ax.plot(degrees, rmses,label= 'Regression of '+labels[count],marker=s, markersize=5,  linestyle =ln,linewidth=1.25,alpha=alp
    ax.set_yscale('log')
    ax.set_xlabel('Degree')
    ax.set_ylabel('RMSE')
    poly=make_pipeline(PolynomialFeatures(min_deg),Ridge())
    poly.fit( x_train,y_train)
    y_pred=poly.predict(x_test)

    ###################################################

    x2=np.linspace(x[0],x[-1]+1+predictYears,((x[-1]+1)-x[0])*dotPerYear)
    X2=x2[:,np.newaxis]
    poly.fit(X,y)
    y_pred=poly.predict(X)
    y_pred2=poly.predict(X2)
    mse=mean_squared_error(y,y_pred)

    if (print_curves==1):
        j= count% markers_size
        alpha=0.75
        ln= linestyless[j]
        #print('ln=',ln)
        fig2 = plt.figure()
        ax2 = fig2.add_subplot(111)
        #plt.xticks(np.arange(5, 20, step=1))
        ax2.plot(X2,y_pred2,label= 'Regression of '+labels[count],marker=j, markersize=5,  linestyle =ln,linewidth=1.25,alpha=al
        ax2.scatter(x_train, y_train,  color='green', marker='*', s=25) #, label='Train dots')
        #plt.scatter(X_test, y_test, color='black',marker='o',s=20) #, label='Test dots')
        plt.legend(loc='upper left',prop={'size': 8})

        #ax.plot(x_axis,de)
        #ax.plot(x_axis,de,'-p',label=lbl, marker=j, markersize=5,linewidth=0.5,alpha=alpha) #, linestyle=':')


    return X2, y_pred2,dotPerYear,labels[count]

x, y,dotPerYear,lbl=regCurves3(14,dotPerYear,sc_df,degree,predictYears,1)
```

### 8.3. Расчет индикаторов

```
#calculate of derivatives
def deriv(a,b):
    return((b-a)/2)
def calcDeriv2(y,x, dotPerYear,lbl):
    de=np.zeros(np.array(y.shape))
    num=np.arange(np.array(y.size)-2)
    for i in num:
        de[i+1]=deriv(y[i],y[i+2])

    de=de[0:np.array(de.shape)[0]-1]
    x_axis=(x[0:np.array(x.shape)[0]-1])

    return de,x_axis
```

На этом шаге рассчитывается производная с помощью библиотеки numpy.

NumPy - это библиотека для языка программирования Python, в которую добавлена поддержка больших многомерных массивов и матриц, а также большой набор математических функций высокого уровня для работы с этими массивами. NumPy является программным обеспечением с открытым исходным кодом и имеет много участников.

NumPy нацелен на эталонную реализацию CPython Python, который является неоптимизирующим интерпретатором байт-кода. Математические алгоритмы, написанные для этой версии Python, часто работают намного медленнее, чем скомпилированные эквиваленты. NumPy частично решает проблему медлительности, предоставляя многомерные массивы, а также функции и операторы, которые эффективно работают с массивами, что требует переписывания некоторого кода, в основном внутренних циклов, с использованием NumPy.

Использование NumPy в Python дает функциональность, сравнимую с MATLAB, поскольку они оба интерпретируются [17], и они оба позволяют пользователю писать быстрые программы, если большинство операций работают с массивами или матрицами вместо скаляров. Для сравнения, MATLAB может похвастаться большим количеством дополнительных наборов инструментов, в частности Simulink, тогда как NumPy неразрывно интегрирован с Python, более современным и полным языком программирования. Кроме того,

доступны дополнительные пакеты Python; SciPy - это библиотека, которая добавляет больше функциональности, подобной MATLAB, а Matplotlib - это пакет для черчения, который обеспечивает функциональность, подобную MATLAB. Внутренне и MATLAB, и NumPy полагаются на BLAS и LAPACK для эффективных вычислений линейной алгебры.

Привязки Python широко используемой библиотеки компьютерного зрения OpenCV используют массивы NumPy для хранения и обработки данных. Поскольку изображения с несколькими каналами просто представлены в виде трехмерных массивов, индексация, разрезание или маскирование с помощью других массивов являются очень эффективными способами доступа к определенным пикселям изображения. Массив NumPy как универсальная структура данных в OpenCV для изображений, выделенных характерных точек, ядер фильтров и многого другого значительно упрощает рабочий процесс программирования и отладку.

```
%matplotlib notebook
plt.style.use('seaborn-whitegrid')
#plt.ion()

fig, ax = plt.subplots()
fig.set_size_inches(12,8)
fig.suptitle('D1 (Speed)', fontsize=12)
ax.set_xlabel('Years', fontsize='medium' ) #fontsize=10)
ax.set_ylabel('D1', fontsize='medium')   # relative to plt.rcParams['font.size']
predictYears=2 #2

markers=np.array(['o', '.', ',',  '+','v', '^', '<', '>', 's', 'd','x','1'])
linestyless=np.array(['-','--','-.',':','-','--','-.',':','-','--','-.',':',])
de_cit=0
de=0

markers_size=markers.size
degree=5
for i in np.arange(labels.size):
    #plt.figure(1)
    if i>5 and i<=10:
        x,y,dotPerYear,lbl=regCurves3(i,dotPerYear,sc_df,degree,predictYears,0) # by count

        x_cit,y_cit,dotPerYear,lbl=regCurves3(i,dotPerYear,sc_df_cit,degree,predictYears,0) #by citations

        de,x_axis=calcDeriv2(y,x, dotPerYear,'****')

        de_cit,x_axis=calcDeriv2(y_cit,x_cit, dotPerYear,'****')

        de=de+0.9*de_cit    #vichislenie D1
        j= i% markers_size
        alpha=(i) /labels.size
        alpha=0.75
        ln= linestyless[j]

        ax.plot(x_axis,de,'-p',label=lbl, marker=j, markersize=5,linestyle =ln, linewidth=2,alpha=alpha) #, linestyle=':')
       # ax.plot(x_axis,de_cit,'-p',label=lbl, marker=j, markersize=5,linestyle =ln, linewidth=1,alpha=alpha) #, linestyle=':')
        plt.xticks(np.arange(5, 24, step=1))
        ax.set_ylim([-50,50])
        plt.legend(loc='upper left',prop={'size': 12})
#print black zero line
zero_line=np.zeros(x_axis.size)
ax.plot(x_axis,zero_line,'-k')
#end print zero line
print(de)
plt.draw()
```

В данном отрывке рассчитываются индикаторы. В аргумент функции regCurves3 передается порядковый номер отвечающей за название научного раздела. На выходе получаем дополнительные точки. Далее рассчитываются первые производные по количеству публикаций и цитирований, суммируется и выводится на графике.

Расчет индикатора D2 производится аналогичным путем применения функции calcDervi2 дважды.

```
#plt.ion()
plt.style.use('seaborn-whitegrid')
fig, ax = plt.subplots()
fig.set_size_inches(12,8)
fig.suptitle('D2 (Acceleration)', fontsize=10)
ax.set_xlabel('Years', fontsize='medium' ) #fontsize=10)
ax.set_ylabel('D2', fontsize='medium')   # relative to plt.rcParams['font.size']
predictYears=2 #2

markers=np.array(['o', '.', ',', 'x', '+', 'v', '^', '<', '>', 's', 'd'])
linestyless=np.array(['-','--','-.',':','-','--','-.',':','-','--','-.',':',])
markers_size=markers.size
degree=1
for i in np.arange(0,6):

    x,y,dotPerYear,lbl=regCurves3(i,dotPerYear,sc_df,degree,predictYears,0)
    x,y_cit,dotPerYear,lbl=regCurves3(i,dotPerYear,sc_df_cit,degree,predictYears,0)

    de,x_axis=calcDeriv2(y,x, dotPerYear,'****')
    de,x_axis=calcDeriv2(de,x_axis, dotPerYear,'****')

    de_cit,x_axis=calcDeriv2(y_cit,x, dotPerYear,'****')     # proizvodnaya po citatam
    de_cit,x_axis=calcDeriv2(de_cit,x_axis, dotPerYear,'****')

    de=de+0.9*de_cit #vichislenie D2

    j= i% markers_size
    ln= linestyless[j]
    #ax.plot(x_axis,de)
    #ax.plot(x_axis,de,'-p',label=lbl, marker=j, markersize=5,linestyle =ln, linewidth=1,alpha=alpha) #, linestyle=':')
    #ax.plot(x_axis,de)
    ax.plot(x_axis[2:],de[2:],'-p',label=lbl, marker=j,markersize=5,linestyle =ln,linewidth=2,alpha=0.75)

    ax.set_ylim([-6,7.5])
    #ax.set_ylim([-100,100])
    plt.xticks(np.arange(5, 24, step=1))

    #print(x_axis)
    #ax.plot(x_axis,de,'-',label=lbl)
    #ax.plot(x_axis,de,'o',label=lbl)
#print zero line
zero_line=np.zeros(x_axis.size)
ax.plot(x_axis,zero_line,'-k')
#end print zero line

plt.draw()
print(de_cit)
print(de_cit*5)
plt.legend(loc='upper left',prop={'size': 12})
```

Далее описана имплементация градиентного спуска. Мы отображаем функцию стоимости как функцию оценок параметров, то есть диапазон параметров нашей функции гипотезы и стоимость, полученную в результате выбора определенного набора параметров. Мы движемся вниз к ямкам на графике, чтобы найти минимальное значение. Способ сделать это - взять производную функции стоимости, как объяснено выше. Градиентный спуск понижает функцию стоимости в направлении самого крутого спуска. Размер каждого шага определяется параметром α, известным как скорость обучения.

```python
import numpy as np
import matplotlib.pyplot as plt

class Linear_Regression:
    def __init__(self, X, Y):
        self.X = X
        self.Y = Y
        self.b = [0, 0]

    def update_coeffs(self, learning_rate):
        Y_pred = self.predict()
        Y = self.Y
        m = len(Y)
        self.b[0] = self.b[0] - (learning_rate * ((1/m) *
                                    np.sum(Y_pred - Y)))

        self.b[1] = self.b[1] - (learning_rate * ((1/m) *
                                    np.sum((Y_pred - Y) * self.X)))

    def predict(self, X=[]):
        Y_pred = np.array([])
        if not X: X = self.X
        b = self.b
        for x in X:
            Y_pred = np.append(Y_pred, b[0] + (b[1] * x))

        return Y_pred

    def get_current_accuracy(self, Y_pred):
        p, e = Y_pred, self.Y
        n = len(Y_pred)
        return 1-sum(
            [
                abs(p[i]-e[i])/e[i]
                for i in range(n)
                if e[i] != 0]
        )/n
    #def predict(self, b, yi):

    def compute_cost(self, Y_pred):
        m = len(self.Y)
        J = (1 / 2*m) * (np.sum(Y_pred - self.Y)**2)
        return J

    def plot_best_fit(self, Y_pred, fig):
                f = plt.figure(fig)
                plt.scatter(self.X, self.Y, color='b')
                plt.plot(self.X, Y_pred, color='g')
                f.show()
```

```python
def main():
    X = np.array([i for i in range(11)])
    Y = np.array([2*i for i in range(11)])

    regressor = Linear_Regression(X, Y)

    iterations = 0
    steps = 100
    learning_rate = 0.01
    costs = []

    #original best-fit line
    Y_pred = regressor.predict()
    regressor.plot_best_fit(Y_pred, 'Initial Best Fit Line')


    while 1:
        Y_pred = regressor.predict()
        cost = regressor.compute_cost(Y_pred)
        costs.append(cost)
        regressor.update_coeffs(learning_rate)

        iterations += 1
        if iterations % steps == 0:
            print(iterations, "epochs elapsed")
            print("Current accuracy is :",
                  regressor.get_current_accuracy(Y_pred))

            stop = input("Do you want to stop (y/*)??")
            if stop == "y":
                break

    #final best-fit line
    regressor.plot_best_fit(Y_pred, 'Final Best Fit Line')

    #plot to verify cost fuction decreases
    h = plt.figure('Verification')
    plt.plot(range(iterations), costs, color='b')
    h.show()

    # if user wants to predict using the regressor:
    regressor.predict([i for i in range(10)])

if __name__ == '__main__':
    main()
```