Составление корпуса является неотъемлеммой частью создания семантической сети. Считается, чем лучще и качесвеннее корпус, тем меньше временных затрат уходит на создание сети. Для того, чтобы выбрать цсе необзодимые и релевантные слова, касающиеся темы Природные катаклизмы, было принято решение ограничеится корпусом в приблизительно 4 Мб. Тут как раз и были применены все навыки поиска, сортировки и экстракции информации, которые касаются сферы компютерных лингвистов. Данный корпус был создан при помощи разных видов кровлеров и парзеров. Так как некоторые веб страницы уникальны и имеют собственную разметку, которая вклбючает в себя все возможные html элементы, почти для каждой были созданы собственные кровлеры, которые помогли вытянуть именно чистый текст, которые почти не содержал в себе рекламы, описания картинок и тд. Хорошим моментом стал так же тот факт, что многие страницы имеют одинаковою структуру, что значительно облехчило поиск нужной информации и элементов для получение текстов. Например, для страниц википедии было достоточно создать небольшую программу на базе парзера BeautifulSoup, где было достаточно указать ссылки желаемых статей и они были беспроблемно и быстро загружены и записаны в указаный файл.

url = **'https://de.wikipedia.org/wiki/Naturkatastrophe'**response = requests.get(url)  
soup = BeautifulSoup(response.content, **"html.parser"**)

Лишь указав нужный html-тег, программа сама переходила на указанный ресурс, находила его на странице, очищала от ненужных элементов и информации и добавляла его в наш корпус.

content = soup.find(**'div'**, class\_=**'mw-parser-output'**)

text = content.get\_text()

Для получения иформации со старниц BR24 поребовалось задействование модифицрованой версии выше описаного алгоритма. Ситуация заключалась в следующем: многие страницы генерируют свой контент динамическиб который изменяется или загружается после начальной загрузки страницы, что немного усложняет поиск и соответственно стягивание информации. В таких ситцациях зхорошим помощником являются инструменты для автоматизации действий веб-браузера. В данной работе был использован Selenium. Он позволяет выполнить скрипты и действия, чтобы получить актуальную информацию. Его задачей было имитация действия пользователя, такие как щелчки по элементам, заполнение форм и прокрутка страницы, что быть полезно при сборе данных. Сначала программа выполняла поиск по сайту, используя ключевые слова (которые были переданы вручную). Затем все найденные ссылки (URLs) по даной теме были записаны в документ в формате json, где наш алгоритм переходил по указаным адрессам и собирал необзодимую информацию.

Багодаря таким действиям вся ручная работы по сбору информации была минимирована и время составления корпуса значительно уменьштлось.

Так же было принято решение о разделении корпуса на подтемы, так сказать по принципу «разделяй и властвуй». Идеей этого было уменьшение размера обрабатываемого корпуса, потому что некоторые модели (например как лемматизация в spacy) не могут обрабатывать большие объемы текстов. Кроме этого тазгруппирование облегчило поиск ключевых слов по каждой теме.

После подготовки корпуса слудующей целью работы было углубление в тему природных катаклизмов. Это было необходимым для постоения семантической сети, чтобы иметь общую картину о данной теме и создать такую цепочку концептов, которая не будет с самого начала утяжелять понимание пользователя и позволит расскрывать данную тему шаг за шагом. Для этой цели были взяти основы из страниц википедии, а так же некоторые статьи из интернета, которые позволили сгруппировать концепты. Это послужило тому, что сеть изначально представляет собою обобщенное значение и дает краткое описание центрального концепта.

Следующим шагом составления сети был поиск концептов в корпусе. План состоял в том, чтобы создать списки разного рода слов, биграммы, триграммы и четырехграммы, с их частотами. Цель этого было отслеживание частоты появления тех или иных словосочетаний с нашими ключевыми словами. Чтобы получить максимально точные, но при этом не огромное колличество данных, все подкорпуса были приведены в начальную форму, то есть лемматизированы. Для этого была создана небольшая программа на основе библиотеки для обработки естественного языка spacy, которая предоставляет удобные инструменты для токенизации, лемматизации, извлечения сущностей и других операций с текстом. Она предлагает различные модели языка, которые обучены на больших объемах данных и способны эффективно обрабатывать тексты на разных языках. Кроме этого весь текст был сведен олько к нижнему регистру и очищен от знаков пунктцации и пустых строчек. Так же были убраны не только стопслова, которые уже включены в модули библиотеки nltk, но и вручную дописаны все нерелевантные глаголы, имена прилагательные, обобщенные имена существительные и так далее. Преимуществом такого переделаного текста было так же возможность быстро создавать не только биграммы, но и разного выида н-граммы, при этом изменяя в самой программе лишь одну строчку. Таким образом удалось сформировать списки слов, объем которых для биграмм был примерно 500 пар на каждую тему, что значительно уменьшило время их обработки.

В следующем этапе процесса формирования семантической сети был проведен поиск концептов в корпусе текстов. Планом работы было создание списков слов разного рода, включая биграммы, триграммы и четырехграммы, с указанием их частоты встречаемости. Главная цель состояла в отслеживании частотности использования данных словосочетаний в контексте наших ключевых слов.

Для достижения наибольшей точности без избыточного объема данных, все подкорпусы были приведены к начальной форме слов, т.е. лемматизированы. Для этой цели была разработана небольшая программа на базе библиотеки обработки естественного языка spacy, которая предоставляет удобные инструменты для токенизации, лемматизации, извлечения сущностей и других операций с текстом. Библиотека spacy включает в себя различные языковые модели, обученные на больших объемах данных, что позволяет эффективно обрабатывать тексты на разных языках.

Затем тексты были приведены к нижнему регистру и очищены от знаков пунктуации и пустых строк. Для удаления стоп-слов, т.е. часто встречающихся и малозначимых слов (например, предлоги, союзы и артикли), была использована библиотека NLTK. Дополнительно был составлен список нерелевантных глаголов, прилагательных и обобщенных существительных, которые также были удалены из текста.

Следующим шагом было формирование списков словосочетаний разных длин. Были созданы биграммы (пары слов), триграммы (тройки слов) и четырехграммы (четверки слов). Для этого была разработана программа с использованием библиотеки NLTK (Natural Language Toolkit). NLTK предоставляет широкий набор инструментов для обработки естественного языка, включая функционал для создания словосочетаний на основе заданных текстовых данных.

Списки слов и словосочетаний, такие как биграммы, содержали примерно 500 пар на каждую тему. Это позволило сократить объем данных, с которыми мы работали, и ускорить последующие этапы обработки. Путем унификации и структурирования текстовых данных на этих этапах, мы смогли более эффективно анализировать информацию для дальнейшего использования.

Имея уже подготовленную основу из наших документов, можно было переходить к самой времени затратной части нашей работы - поиску и выбору нужных и релевантных слов. Тут цель была уже отобрать те слова и словосочетания, которые относятся напрямую к теме Природных катаклизмов. Иными словами, эти концепты должны без целых предложений объяснить читателю тему, исключая многосмысленные трактования и путаницы в терминах. Это своего рода ассоциативная карта или тезисное представление полной темы. Хоть упор и делался на частоту происхождений тех или инных словосочетаний, большая часть работы была все равно связанв с текством. Так как смысл некоторых результатов был не очень понятен, текст давай зорошее объяснение. Зачастую, достаточно было найти искаемое ключевое слово в тексте, а уже его предложение помагало отрисовать картину и решить, будет ли данное выажение частью семантической сети.

Так как наш корпус состоит практически из новостных статей и в нем нет научной теории, было принято решение сгруппировать некоторые концепты. Для этой процедуры был использован гугл как ресурс по поиску непосредственной терминологии, которая могла дать разъяснение каждому концепту в пределах 2-3 слов. Преимуществом такой идеи является то, что пользователь уже в процессе работы с сетью будет знать о чем идет речь на следующем уровне сети и это позволит ему быстро выстроить логические взаимосвязи между концептами.

Заключительным этапом для создания лексикона для семантической сети стала ее кодировка. Для того чтобы наши данные были также читаемы для вычеслительныз машин, их нужно было помустить в соответсвующую струетуру данных. Выбор пал на формат Comma-separated value, потому как жто позволило создать так называемые массивы для каждго концепта. В такой массив взодила информации о id узла, его название, синонимы, все вохможные формы слова, тип (экстерн или интерн) и id его родителя (к какому концепту относится данное слово/словосочетание).

Так как эта работа нацелена прежде всего на лингвистический аспект, формы слова и их значение было неотъемлемой структурой каждого узла сети. Учет разных форм слова позволяет более полно и точно передать грамматические, семантические и лексические аспекты темы, что в свою очередь способствует более точному анализу и пониманию текстовых данных. Типы узлов указывают на принадлежность данного концепта к теме, насколько плотно они связаны с друг другом. Так как все компоненты всегда связаны, но не всегда напрямую принадлежат к теме, они выражены как экстерновские узлы и служат для связи с «внешним миром». Сеть это всегда ацикличееский граф и узлы имеють свои родительские узлы, которые связаны между собою с указанием ids вверх лежащих узлов.

Оценивание семантической сети

После составления нашей семантической сети пред нами постал вопрос о ее качестве. Самым известным методом оценивание стал метод сравнение. Для этой процедуры все наши концепты (узлы сети) были записаны в тхт файл и было выбрано 3 книги о природных катаклизмах. После при помощи базовый строчек кода на питоне наша программа смогла с легкостью найти все слова в книгах, которые так же совпадают с нашей сетью. Во внимание принимальсь только наши концепты, то есть формы слова в единственном числе. Интересным замечанием было то, что каждый из 3 текстов имел совершенно разные рузультаты. Из одного текста было выявлено всего 1/8 слов со всех концептов, когда в то же время инной книге результат был 1/4. Как еще оказалось, эти данные можно было улучшить, добавив в список концептов части слов, такие как «beben» или «versicherung», или же их множественные формы, например «Schäden» или «Wirbelstürme». Этот метод конечсно же не так существенно, но немного ихменил статистику.

Чтобы подвести итог на примере этих трех текстов, было так же принято решение сравнить все три списка из данных текстов и найти хт instersectioт, то есть найти найиболее расправстранненыеслова в этих текстах и выделить все слова, которые встречабтся во всех трех текстах. Тут естественно результат оказался намного хуже: из 368 узлов оказалось только 36, которые присутствовали во всех текстах.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

После составления семантической сети возник вопрос о ее качестве. Для оценки мы применили широко известный метод сравнения, который заключался в анализе присутствия всех узлов сети в различных текстах. Для этой процедуры мы записали все концепты (узлы) сети в текстовый файл и выбрали три книги, посвященные природным катаклизмам. С помощью простого кода на языке Python наша программа легко нашла все слова в этих книгах, которые также совпадали с нашей семантической сетью. При анализе учитывались только концепты в единственном числе.

Интересное наблюдение заключалось в том, что каждая из трех книг дала совершенно разные результаты. Из одной книги было выявлено всего лишь 1/8 слов, соответствующих нашим концептам, в то время как в другой книге этот результат составил 1/3. Это указывает на вариативность использования терминологии в разных источниках и на неоднозначность связей между понятиями в семантической сети.

Кроме того, мы обнаружили, что результаты могут быть улучшены путем расширения списка концептов. Мы добавили частные формы некоторых слов, такие как "beben" или "versicherung", а также их множественные формы, например, "Schäden" или "Wirbelstürme". Хотя эти дополнения не имели существенного влияния на общую статистику, они улучшили совпадение слов в текстах с нашей сетью.

В итоге, для суммарной оценки семантической сети на основе этих трех текстов, мы решили сравнить все три списка слов и найти пересечение, то есть наиболее распространенные слова, которые встречаются во всех трех текстах. Однако результаты оказались намного хуже ожидаемых: из 368 узлов сети только 36 присутствовали во всех трех текстах. Это указывает на то, что наша семантическая сеть может быть дальше усовершенствована и расширена для более точного отражения связей между понятиями в контексте природных катаклизмов.

Одной из возможных причин низкой статистики является то, что в нашей семантической сети мы использовали термины, которые явно указывают на тему природных катаклизмов. Это означает, что при обычном поиске эти слова будут непосредственно связаны с этой темой и иметь высокую релевантность.

Однако, при использовании семантической сети, мы стремились улучшить и уточнить поиск, учитывая не только прямые связи, но и контекстуальные отношения между понятиями. В результате, некоторые более общие или абстрактные термины, которые могут быть связаны с природными катаклизмами, могли быть исключены из списка концептов.

Это ограничение сети может привести к недостаточному покрытию текстов и, как следствие, низкой статистике совпадений. Для улучшения результатов и повышения статистики можно рассмотреть возможность включения таких общих терминов или использования дополнительных методов анализа контекста для расширения и уточнения семантической сети. Это позволит улучшить точность и полноту поиска в разнообразных текстах о природных катаклизмах.



Тема применения семантических сетей для предсказания и оценивания природных катастроф является актуальной исследовательской областью, которая привлекает внимание ученых и специалистов по области стихийных бедствий. Семантические сети представляют собой структуры, моделирующие связи между понятиями и позволяющие обрабатывать и анализировать информацию с учетом их семантической связности.

Одним из основных применений семантических сетей в контексте природных катастроф является предсказание таких событий и оценка их потенциального воздействия. С помощью семантических сетей можно анализировать различные факторы, включая географические, климатические, гидрометеорологические и другие данные, чтобы выявить сигналы и паттерны, предшествующие возникновению природных катастроф. Например, сети могут использоваться для обнаружения корреляций между определенными погодными условиями и вероятностью возникновения определенных катастроф, таких как ураганы, землетрясения или наводнения.

Семантические сети также могут быть полезны при оценке и прогнозировании потенциальных последствий природных катастроф. Они позволяют учитывать различные факторы, такие как местоположение, топографию, населенные пункты, инфраструктуру и другие аспекты, которые могут повлиять на масштаб и воздействие катастрофы. Анализируя информацию из различных источников и учитывая семантические связи между этими факторами, семантические сети могут помочь в оценке рисков и разработке стратегий управления катастрофами.

Кроме того, семантические сети могут использоваться для интеграции различных типов данных, включая географические карты, сателлитарные изображения, статистические данные и др. Это позволяет создавать комплексные модели, которые объединяют информацию из разных источников и позволяют получать более точные и полные прогнозы и оценки природных катастроф.

Однако, несмотря на потенциальные преимущества, применение семантических сетей в предсказании и оценке природных катастроф также сопряжено с некоторыми вызовами. Например, требуется большой объем данных для обучения и настройки семантической сети, а также высокая степень точности и достоверности данных для получения надежных результатов. Кроме того, необходимо постоянно обновлять и адаптировать семантическую сеть, учитывая изменяющиеся условия и новые данные.

В заключение, применение семантических сетей в предсказании и оценке природных катастроф представляет собой перспективную область исследований, которая может принести значительный вклад в снижение рисков и улучшение стратегий управления стихийными бедствиями. Однако, для достижения оптимальных результатов необходимо учитывать сложности, связанные с обработкой больших объемов данных, обновлением моделей и обеспечением высокой достоверности информации.

Крім того семантичні мережі відрізняються за кількістю типів та поділяються на однорідні(з єдиним типом) та неоднорідні (з різними типами). В свою чергу існують бінарні (відношення зв’язують два об’єкти) та н-нарні (зв’язують більше двох понять) відношення. За *типом структури вершин* поділяються на простого (вершини не мають власної структури)та ієрархічного типу (ають власну структуру у вигляді мережі). Ієрархічні можуть поділити мережу на піжмережі.

онтологія – це точна специфікація певної області, що містить у собі словник термінів цієї області і множину логічних зв’язків (типу «елемент-клас», «частина-ціле») [6], що описують, як ці терміни співвідносяться між собою.

<https://core.ac.uk/download/pdf/132546583.pdf>

поняття онтології сильно перетинається з уже давно прийнятим в інформатиці і лінгвістиці поняттям тезауруса. Онтології дозволяють представити поняття в такому вигляді, що вони стають придатними для машинної обробки. На сьогодні існує багато різноманітних онтологій, описаних за допомогою різних мов та пов’язаних із найрізноманітнішими предметними областями. Ці онтології різняться за багатьма властивостями – обсягом, виразними можливостями, призначенням, ступенем формалізації знань тощо.

Genauigkeit: Wie präzise und korrekt sind die Konzepte und ihre Beziehungen im semantischen Netz dargestellt? Gibt es Unstimmigkeiten oder Fehler?

Die Konzepte aus unserem semantischen Netz stammen aus dem selbst gesammelten Korpus und orientierten sich in der ersten Reihe auf Leser der Artikelnachrichten.

Vollständigkeit: Sind alle relevanten Konzepte und ihre Beziehungen im Netz enthalten, oder gibt es Lücken und fehlende Informationen?

Um Lücken oder fehlende Informationen zu vermeiden wurden zusätzliche Rechersche im Internet durchgeführt, indem wir fast jeden Begriff durchgeprüft, analysiert und in das Nezt so angepasst, dass sie eine logische und verständliche Struktur haben.

Skalierbarkeit: Kann das semantische Netz problemlos erweitert und an neue Informationen angepasst werden? Ist es flexibel genug, um zukünftige Anforderungen zu erfüllen?

Das Netz bietet auch die Möglichkeit an Skalierbarkeit. Dafür kann man als Grundlage verschiedener Fachspezifische Literatur verwenden (wie z. B die Bücher aus unserer Evaluierung). In dem Fall wird sich aber die Stuktur bzw Konzepte ändern. Der Grund dafür, dass viele von Begriffen in unserem Netz sind adaptiert direkt zu dem Netz und brauchen keine zusätzlichen Informationen und Beschreibungen: „Opfer“ ist allgemein, in unserem Netz existieren verschieden Bezeichnungen dafür wie „Erdbebenopfer“ oder „Flutopfer“.

Nutzbarkeit: Ist das semantische Netz für Benutzer leicht zugänglich und verständlich? Sind die Konzepte und Beziehungen intuitiv und logisch aufgebaut?

Das Netz ist für jeden Benutzer leicht zugänglich und verständlich und bietet einen Gesamtüberblick über das Thema „Naturkatastrophen“ und liefert Informationen über Ursachen, Folgen, Maßnahmen und Maßstaben den größten Naturkatastrophen.

Effizienz: Wie schnell und effizient kann auf das semantische Netz zugegriffen und damit gearbeitet werden? Gibt es Engpässe oder Leistungseinbußen?

Die Effizienz des Netzes ist hoch und bietet dank der passenden Gruppierung eine schnelle Suche zu jedem Konzept.

Robustheit: Ist das semantische Netz widerstandsfähig gegenüber Störungen, Fehlerquellen oder Änderungen in den Eingabedaten?

Um ein vollständiges und zu 100 % korrektes Netz zu erstellen muss man in dem Idealfall mit Experten aus diesem Domäne arbeiten, da sie sich mit dem Thema auskennen und mit Verbindungen zwischen einzelnen Knoten helfen können. Da unser Netz aus allgemeinden Daten ausgegangen ist, kann man nicht zu 100% ihre Robustheit und Genauigkein garantieren.

Erweiterbarkeit: Lässt sich das semantische Netz leicht erweitern, um neue Konzepte, Beziehungen oder Domänen abzudecken?

Da das semantische Netz tief mit anderen Domänen wie zB Geologie überschneidet, kann ses auch erweitert werden. In dem fall wird es als ein Bestandteil (Frame) eines großes Netz sein, das zB Naturereignisse beschreiben würde.

Vergleich mit anderen Ansätzen: Wie schneidet das entwickelte semantische Netz im Vergleich zu anderen ähnlichen Ansätzen oder vorhandenen Ontologien ab?

При составлении нашей онтологии мы решили так же взять в расчет и частоту тех или иннных слов и словосочетаний в нашем корпусе. И как показал анализ, частота хоть и дала нам некоторые свойства значимости, но тем не менее не сыграла ключевую роль. Просмотрев все списки слов до самого конца, быо обнаружено, что не менее важные и существенные выражение имеют хоть и частоту N=1, но не могут быть исключены, так как являются неотъемлемой частью темы. Более того, списки из четырьохграммами не принесли никакой пользы, так как почти идентичная информация уже содержалась в би- и триграммах.

(Savinkin, 2019)

Это была сеть лишь с 6 узлами, еоторая описывала основные области данной темы,а именно: , и служила началом для развертывание последующих подтем.