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. 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.

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.