**Проект solver**

**Размещение слов из словаря в памяти:**

При парсинге для каждого слова вычисляется свой хэш код, и этот код будет являться индексом для доступа к массиву из слов загруженного словаря. Вариант использовать вместо массива структуру типа map. Это увеличит время доступа к элементу, но развяжет руки с функцией хеширования. Она тогда сможет быть любой.

**Получить словарь из word**

Для использования словарей от OpenOffice необходимо переименовать расширение oxt на rar и открыть обычным архиватором. Файл с расширением aff содержит принципы изменения слов. Файл с расширением dic – сами слова.

Возможно придется распарсить какой-то(какие-то) словари под собственные нужды.

Для начала необходимо определить, что именно, из данных в словарях мне интересно, и лишь потом приводить имеющиеся слова к необходимости. Для начала составить собственные данные для словаря. Такой экспериментальный словарь.

Открыть Базу Ильи Козиева и посмотреть как он это делал! Структуру.

**Еще посмотреть примеры нейронных сетей…**

При составлении словарей может рассмотреть случай, когда в словарь добавляется часть слова и правила формирования полных слов из этой части.

Например, слово – мечта

мечт, сущ., ж., им.-а, род.-ы, дат.-ой,… мн. им.-ы, мн.род.-ы, мн.дат.-ами, глагол – мечтать, я – аю, ты – ешь, вы – ете, они –ют,… прил.им. - ательный,… ср.род им.-ательное

Количество окончаний и приставок не бесконечно!!! Можно всем приставкам и суфиксам и окончаниям раздать свои ID. И уже после этого собирать слова.

ДЛЯ ОПИСАНИЯ ЗНАЧЕНИЙ СЛОВ В СЛОВАРЯХ используется ограниченный набор слов. Необходимо после всего определить этот набор. И сделать определение смысла слов только на основании комбинаций этого набора

БЫТЬ МОЖЕТ для разбора слов и получения их хеш-кода придется давать различные ВЕСА для приставок, корней, суффиксов и окончаний.

**ГЛАГОЛЫ**

Попробовать глаголы в таблице определять как инфинитивы без приставок. Определить возможный перечень приставок и добавить к глаголам признаки возможного наличия определенной приставки у глагола.

Сделать таблицу приставок с описаниями вида и действия приставок. Отдельно может потребоваться таблица приставок исключений.

Можно на первых порах не рассматривать переносное значение глагола.

В начале разработки не записывать в таблицу все мыслимые и немыслимые глаголы. Сосредоточиться на 10. Быть может, стоит взять их из определенного вида глаголов, например, глаголов движения.

Записать в таблицу глаголов и способы формирования предложения с определенной приставкой. Например, ВОЙТИ, далее предлог В, далее ВИНИТЕЛЬНЫЙ ПАДЕЖ СУЩЕСТВИТЕЛЬНОГО.

Там где в основном справочнике слов нет в СВОЙСТВАХ Именительного падежа – значит это слово несклоняемое

Бывают варианты, когда одно и тоже слово пишется по-разному. Например, бриллиант и брильянт. И при этом это не синоним а это же слово!!!

Принято решение на первых порах – слова в базе группировать по group\_part, а не по cell\_part. Это связано с тем, что существуют в ячейках слова основная часть которых полностью отличается от других слов в группе. Например, **поадекватнее** и **адекватно** (всего слов около – 30000); **год** и **лет** (сущ. – 54)

Есть слова, который не имеют group\_part – 84 шт. Это между, меж, промеж; он, его, нему…

Рассматривается решение, те слова ячейки, которых имеют одинаковые символы для поиска – добавить как ячейки. Которые имеют групповые части для поиска, - как группы. Все остальные – добавить пословно.

Для выбора слов выбран метод поиска слов в словаре, а не автоматическое определение окончания, по причине того, что в словаре можно указать информацию индивидуального характера. В будущем , возможно необходимо будет написать код просчитывающий окончания падежей, суффиксы и т.д. для слов, отсутствующих в словаре.

Есть наречия и прилагательные, которые являются сравнительными, так вот они начинаются с префикса «**по»** (7159 шт). Для выделения постоянной части – stem, префикс в БД выносится как отдельная часть.

9251 быть гл

20674 год сущ

36275 идти гл

54845 много нар

56676 мы мест сущ

71354 он мест сущ

71355 он сущ

83596 плохо нар

83602 плохой прил

115987 слать гл

115988 слаться гл

129522 умный прил

134944 хороший прил

134949 хорошо нар

136506 человек сущ

142216 я мест сущ

Будут определены как окончания – из-за особенностей русской речи.

При парсинге tag можно русские названия перекодировать на латинские буквы и сделать переменные типа char во всех классах имплементирующих PartOfSpeech.

При создании классов Verb, Noun, Pronoun… следует сделать анализ на частотность встречаний граммем. Те граммемы, которые встречаются чаще – поставить вначале списка сравнений в операторе switch. Также стоит определить частоту встречаний граммем в каждой подгруппе тега и отсортировать по частоте встречаний в существующей базе. Это все нужно только при загрузке классов из таблицы!