Contents

1 Выделение информации

На входе мы имеем несколько тесктов, в которых содержатся рекомендации, советы и т.д. по поводу достижения некоторой единой для всех текстов цели. На выходе мы хотим получить на основе этих текстов общую схему действий для достижения этой цели.

На данный момент мы предполагаем получение, что все советы прописаны достаточно явно. Мы не предполагаем, что нам нужно читать какой-либо личный рассказ о том, как человек достигал этой цели, и в соответствии с анализом этого текста вносить изменения в схему.

1.1 Основные этапы выделения сценария

Обозначим основные этапы выделения сценария:

1. Выделение из каждого текста информации, возможно относящейся к сценарию.
2. Струтуризация информации для каждого текста согласно единому шаблону.
3. Выбор предложений, удовлетворяющих необходимым условиям для вхождения в сценарий.
4. Формирование окончательного сценария на основе полученных на предыдущем шаге выборок.

Теперь разберем каждый из этих этапов. Анализ, насколько идея и реализация успешны, будет в сецкии Результаты И Анализ.

1.2 Выделение Информации Из Исходных Текстов

Рассмотрим отдельный текст со входа. Мы предполагаем, что полезную информацию могут содержать только те элементы текста, которые относятся к одному из двух типов:

- Предложения, содержащие глаголы и отглагольные части речи,
- Списки внутри текста.

Разберем, почему выделили именно эти два типа.

Схема, которую мы хотим получить в конце и которую мы называем сценарием, по сути есть компиляция из всех советов, рекомендаций и указаний, находящихся в текстах. Эти элементы есть побуждения к

действию, а действие всегда за редким исключением выражаются через глагол и отглагольные части речи. По сути это объясняет и почему мы пренебрегаем остальными предложениями.

Однако есть такие элементы, как списки. В некотором смысле, список достаточно часто есть продолжение предыдущих предложений и содержит особо важную выделенную информацию. Это может быть список того, что нужно взять, или того, какие вопросы следует задать. И потому нам не следует ими пренебрегать.

О реализации. Нахождение предложений с глаголами сводится к задаче сегментации текста на предложение и опреднию того, является ли слово глаголом. Для этого существует достаточно большое разнообразие парсеров, в нашем случае мы использовали isanlp. Далее любой найденный глагол или отглагольная часть речи вместе со своими зависимостями будет называться действием.

Теперь поговорим о списках. В нашем представлении список выглядит следующем образом: это череда абзацев, которые начинаются с определенных символов - для нумерованного списка это числа или буквы, для ненумерованного - различные символы, например, '*', '-', '+' и т.д. Так же мы учитываем возможность, что после каждого маркированного абзаца, являющегося элементом списка, идет еще один абзац, который является разъясняющим к первому. Использование этих критериев позволяет легко выделять списки.

Здесь стоит заметить, что список в тексте это необязательно несколько однотипно начинающихся обзацев и он может быть более сложно устроен. В частности, каждый элемент списка может быть представлен более, чем двумя абзацами, или список может содержать вложенные списки. Однако мы пренебрегаем данным фактом, поскольку мы предполагаем, если такие сложные структуры имеют место быть, то каждый элемент списка и так должен содержать некоторые дополнительную информацию, которая будет учтена, как предложения с глаголами. И в результате мы не проиграем сильно от того, что не знаем, что это список.

1.3 Структуризация Информации

Данный этап является надстройкой над предыдущим. В результате предыдущего этапа мы научились выделять действия и списки. В результате этого этапа мы получаем следующую структуру данных: каждый элемент это либо абзац, либо список. Элементы упорядочены согласно встречанию в

тексте. Каждый элемент-абзац содержит все действия, содержащиеся в нем, и название секции, к которой он принаджежит. Каждый элемент-список содержит составляющие его элементы-абзацы и название секции, к которой он принадлежит.

Немного обсудим название секций. Большинство текстов содержат некоторую внутренную структуру, которая выражается в первую очередь за счет заголовков разных уровней (заголовки, подзаголовки, подзаголовки для подзаголовков и т.д.). И логично предположить, что эти заголовки смогут помочь при дальнейшем объединении извлеченной из разных текстов информации в единый сценарий. Для элемента-абзаца, который является частью списка, мы считаем названием секции первое предложение соответствующего элемента списка. Для элементов-списков и всех остальных элементов-абзацев мы считаем названием секции заголовок самого нижнего уровня.

Выделение заголовков построено на крайне простой идеи: каждый заголовок представляет собой абзац, состоящий из единственного предложения. Причем это предложение удовлетворяет двум условиям: оно не очень длинное (эмперически подобранно ограничение не более 10 слов) и оно либо заканчивается обычными для окончания предложения знаками припинания (тока, вопросительный или восклицательный знак), либо в конце не стоит никакого знака припинания.

1.4 Необходимые Условия Вхождения В Сценарий

В результате предыдущих этапов мы получили все действия, которые есть в тексте, и структуризировали их. На этом этапе мы определимся как, используя морфологическую и синтаксическую информацию, оставить то, что нам с крайне высокой вероятностью подходит.

За время анализа текстов были выделены следующие условия:

- Глагол в повелительной форме
- Глагол в форме инфинитива, причем этот глагол зависит от таких слов, как 'можно', 'нужно' и т.д.
- Глаголы во втором лице

Все эти условия так же проверяются, используя инструменты из библиотеки isanlp.

В результате отработки всех предыдущих и этого этапов для каждого текста мы имеем экстракт, который отчасти уже является отдельным сценарием для этого текста. Он имеет два важных следующих недостатка:

- содержит некоторые действия, которые не несут никакой смысловой информации
- не рассматривает возможность ветвления сценария

Устранение первого недостатка, предположительно, произойдет при сравнении сценариев, извлеченных из разных текстов. В частности этому посвящены следующие этапы.

1.5 Результаты И Анализ

Мы работали с выборкой текстов из интернета, которые есть инструкции по покупке автомобиля. Всего в выборке находится 35 документов. Вся выборка состоит из трех частей: инструкции для подержанных автомобилей, инструкции для новых и остальные тексты. Каждая из этих выборок содержит 13, 17 и 5 документов соответственно.

В данном разделе, мы обсудим качество информации, выделяемой выше методом. Заметим, что эти результаты во многом зависят от качества используемого парсера.

О выделении списков. Среди всех 35 документов не было обнаружено такого, что текст в нем содержит список и он не был бы обнаружен. Поэтому критерии выделения списка, описанные в Выделение Информации Из Исходных Текстов, можно считать сформулированными корректно.

О качестве получаемой выборки из текста. Для начала, давайте определимся, как исследовать это качество. Результат первых трех этапов есть выборка, состоящая из действий и списков. Давайте считать каждый элемент списка отдельной единицей информации. Так же пусть у нас имеется выделенный в ручную сценарий. Нас будут интересовать две величины:

- сколько из вручную выделенного сценария элементов было выделено алгоритмически,
- сколько из выделеных алгоритмически элементов входит во множество вручную выделенных элементов.

Т.е. какую часть истинного сценария мы смогли найти и какая часть выборки является нужной. Легко заметить, что это хорошо известные метрики Recall и Precision. Было выбрано два текста и результаты вы можете найти в таблице 1.

col1	Recall,%	Precision, %
Text 1	97.4	97.6
Text 2	83.7	87.8
Mean	90.6	92.7

Table 1: Качество выбираемой информации

Как можно видеть, наш метод показывает достаточно хорошие результаты. При непосредственном ознакомлении с текстами, можно заметить, что больше всего выделяется ненужной информации в начале (предисловии, которое достаточно часто содержится в текстах и не несет сильной смысловой нагрузки) и в конце (пожелания, напутственные слова и т.д.).

2 Об идентификации шагов

В данном разделе мы хотим изучить, насколько хорошо возможно отличить один шаг от другого и насколько хорошо можно определить, что два шага из разных документов есть один и тот же шаг.

Мы сделаем следующие гипотезы:

- Каждому тексту соответствует некоторый вектор в $\mathbb{R}^n$;
- Каждому шагу соответствует некоторый центральный вектор;
- Вектора текстов, которые воплощают именно этот шаг, ближе всего к центрального вектора именно этого шага.

Обсудим эти гипотезы. Соответствие из первой гипотезы строится следующим образом: возьмем некоторую обученную модель word2vec (в нашем случае мы использовали готовые модели RusVectores), каждому слову в тексте поставим в соответствие вектор из этой модели, и тогда вектор, соответствующий этому тексту, есть среднее арифметическое векторов всех слов входящих в него.

Подтверждение же последних двух гипотез равносильно тому, что все шаги легко идентифицируемы. Проверим насколько это верно при помощи следующего эксперимента:

1. Каждый текст вручную разбиваем на множество текстов, каждый из которых есть воплощение одного шага. В результате этого пункта, у нас есть множество текстов, каждый из которых есть отдельный шаг и мы знаем, что это за шаг;

2. При помощи word2vec для каждого текста мы находим соответствующий ему вектор;

3. Для каждого типа шага сценария, мы находим центр соответствующих векторов. В результате, для каждого типа шага сценария у нас есть центральный вектор;

4. Определим для каждого текста-шага тип шага по ближайшему центру и сравним полученную разметку с исходной разметкой.

№ шага	Имя шага	Количество текстов
1	Ваши деньги	5
2	Цены	4
3	Объявлеия	4
4	Телефонный разговор	7
5	Мониторинг Сайтов	9
6	Документы на машину	6
7	ДКП	8
8	Осмотр	8
9	Тест-драйв	5
10	Определиться с маркой и моделью	7
11	Дигностика	4
12	Год выпуска и пробег	3
13	Постановка на учет	1

Table 2: Информация о выделенных шагах

Информацию о выделенных шагах вы можете найти в таблице 2.Результаты сравнения представлены в таблицах 3 и 4.

№ шага	Recall,%	Precision, %
1	100.0	100.0
2	100.0	100.0
3	100.0	80.0
4	85.7	66.7
5	66.7	75.0
6	83.3	83.3
7	75.0	85.7
8	87.5	87.5
9	80.0	100.0
10	100.0	100.0
11	100.0	100.0
12	100.0	100.0
13	100.0	100.0

Table 3: Качество идентификации шагов

Метрика	Значение, %
Mean Recall	90.6
Mean Precision	90.6
Accuracy	87.3

Table 4: Качество идентификации шагов, средние показания