электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.91

Применение Apache UIMA при решении задачи выделения имён из текстов документов

Гречищев К.М.

Студент, кафедра «Компьютерные системы и сети»
МГТУ им. Н.Э. Баумана, г. Москва, Россия
Научный руководитель: Самарев Р.С., к.т.н., доцент кафедры «Компьютерные системы и сети» МГТУ им. Н.Э. Баумана

МГТУ им. Н.Э. Баумана
KMGrechishchev@ya.ru

В ряде случаев при автоматизированной обработке текстов на естественном языке достаточно иметь некоторую ограниченную информацию о документе - метаинформацию, примером которой может быть заголовок документа, его автор, дата создания. Использование метаинформации ускоряет поиск документа, а также позволяет решать задачу классификации документов. Многие системы автоматизации документооборота, например ECM Alfresco[1], поддерживают возможность добавления такой информации к документу и работы с ней.

Во многих случаях пользователя интересует о ком и о чем говорится в конкретном тексте. В этом случае ему удобно было бы обратиться к соответствующему метаинформационному полю документа. Следовательно, информация об именах собственных, упоминаемых в документе, является актуальной. Возникает задача автоматического выделения таких данных из документа.

Данная работа посвящена использованию средств семантического фреймворка Apache UIMA[2] для автоматического выделения из документа информации о фамилии, имени и отчеству лиц, упоминаемых в документе.

Особенности Apache UIMA

Apache UIMA (Unstructured Information Management) - каркас, позволяющий создавать системы, анализирующие большие объемы неструктурированной информации. Для обработки информации средствами UIMA необходимо определить процесс анализа (Analysis Engine - AE) - модуль, анализирующий документы и получающий их них метаинформацию. Каждый такой процесс может состоять из одного или нескольких составных блоков - аннотаторов. Аннотаторы выполняют определенную обработку документа и выдают результат в виде структур признаков. Примечанием (annotation) в терминологии UIMA называется структура признаков конкретного типа, создаваемая аннотатором, которая содержит начальную и конечную позиции примечания в тексте. Кроме того, примечание также может включать в себя некоторую дополнительную информацию, полученную в процессе анализа, например, часть речи или начальную форму слова. Таким образом, UIMA позволяет определить составной процесс анализа, последовательно применяющий несколько аннотаторов, которые могут использовать результаты друг друга.

Схема разработанного анализатора представлена на рисунке 1.

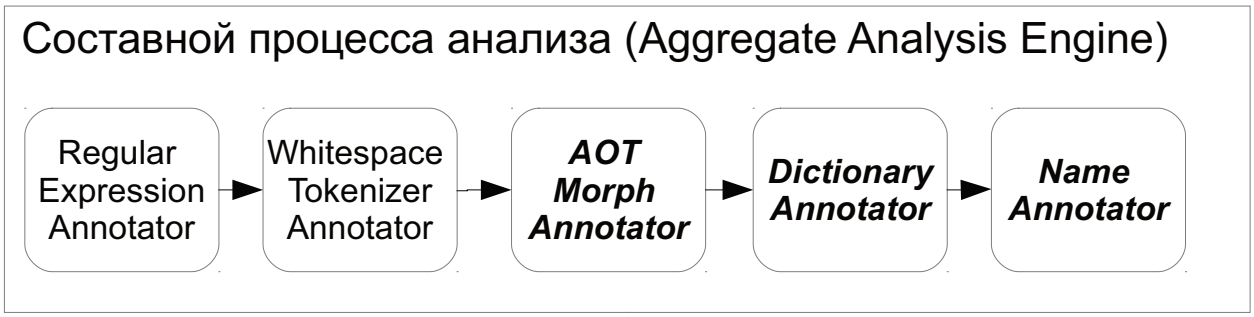

Рис.1. Составной процесс анализа

Процесс состоит из трёх определенных авторами аннотаторов ("AOT Morph Annotator" и "Name Annotator", "Dictionary Annotator") и двух аннотаторов, входящих в состав фреймворка UIMA. Аннотатор "Regular Expression" позволяет выделять в документе аннотации, соответствующие набору правил, заданных регулярными выражениями. "Whitespace Tokenizer" выполняет разбиение документа на составные части в соответствии с заданными разделителями. "Dictionary" аннотатор выполняет построение примечаний на основе некоторого словаря.

Принцип проведения процесса анализа

Первым этапом выделения имени из документа является обнаружение "кандидата" на имя с помощью шаблона на основе регулярного выражения. Было сформулировано 8

шаблонов, для которых написаны регулярные выражения. Шаблоны предусматривали как полную форму задания имени (Фамилия Имя Отчество), так и сокращенную (Фамилия И.О.). С помощью аннотатора "Regular Expression" выделяются аннотации, соответствующие шаблонам. Особенность данного аннотатора заключается в том, что он позволяет пользователю определять собственные регулярные выражения в качестве переменных, которые потом могут входить в другие переменные или в конечное регулярное выражение.

После этого необходимо проверить, что все выявленные на предыдущем шаге примечания действительно являются именами. Для этого, с помощью "Whitespace Tokenizer" аннотатора из обнаруженных примечаний выделяются их составные части - слова. Следующим этапом является морфологический анализ выделенных токенов.

Для морфологического анализа использовался "AOT Morph" аннотатор, разработанный авторами с использованием библиотеки морфологического анализа русского языка, созданной рабочей группой aot.ru[3]. Аннотатор для каждого слова формулирует примечание, содержащее в себе часть речи и базовую форму слова. Для существительных под базовой формой понимается именительный падеж единственного числа.

После получения базовой формы имени становится возможным проверить его по словарю имен. Среди базовых форм слов "Dictionary" аннотатор выделяет те, которые входили в подключенный к нему словарь имен и отчеств. На финальном этапе "Name" аннотатор анализирует выделенные с помощью "RegExp" примечания полных имен на вхождение в них примечаний типа "Dictionary" и формирует примечания типа "Name".

Эксперимент и результаты

Проверка применимости метода проводилась на произведениях русской литературы. Были проанализированы романы Л.Н. Толстого «Война и мир», А.Н. Толстого «Хождение по мукам», Б.Л. Пастернака «Доктор Живаго», М.А. Булгакова «Мастер и Маргарита» и другие, и подсчитано количество вхождений каждого из выявленных имен в документ. Имена были отранжированны по числу вхождений, после чего было построено распределение имен, которое представлено на рисунках 2 и 3.

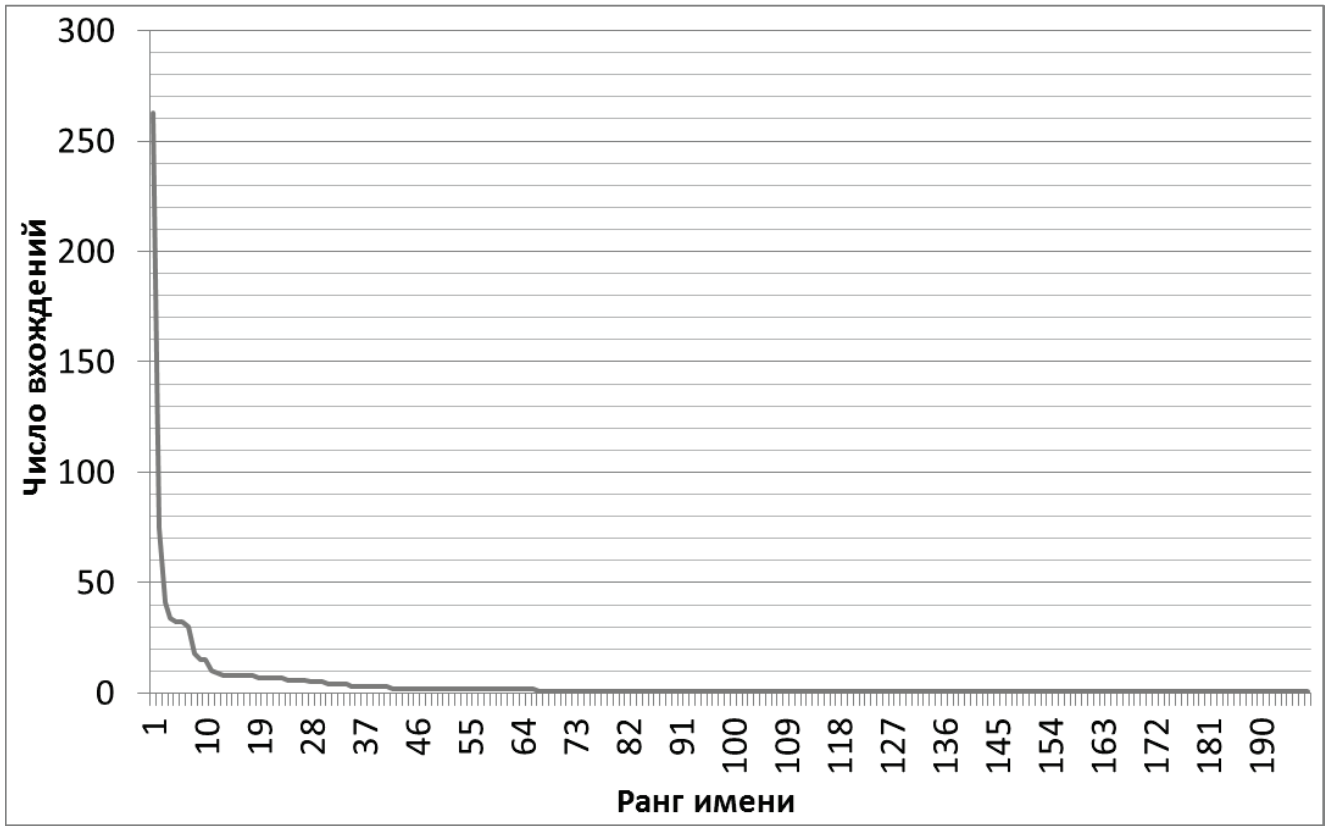

Рис.2. Распределение имен в романе «Доктор Живаго»

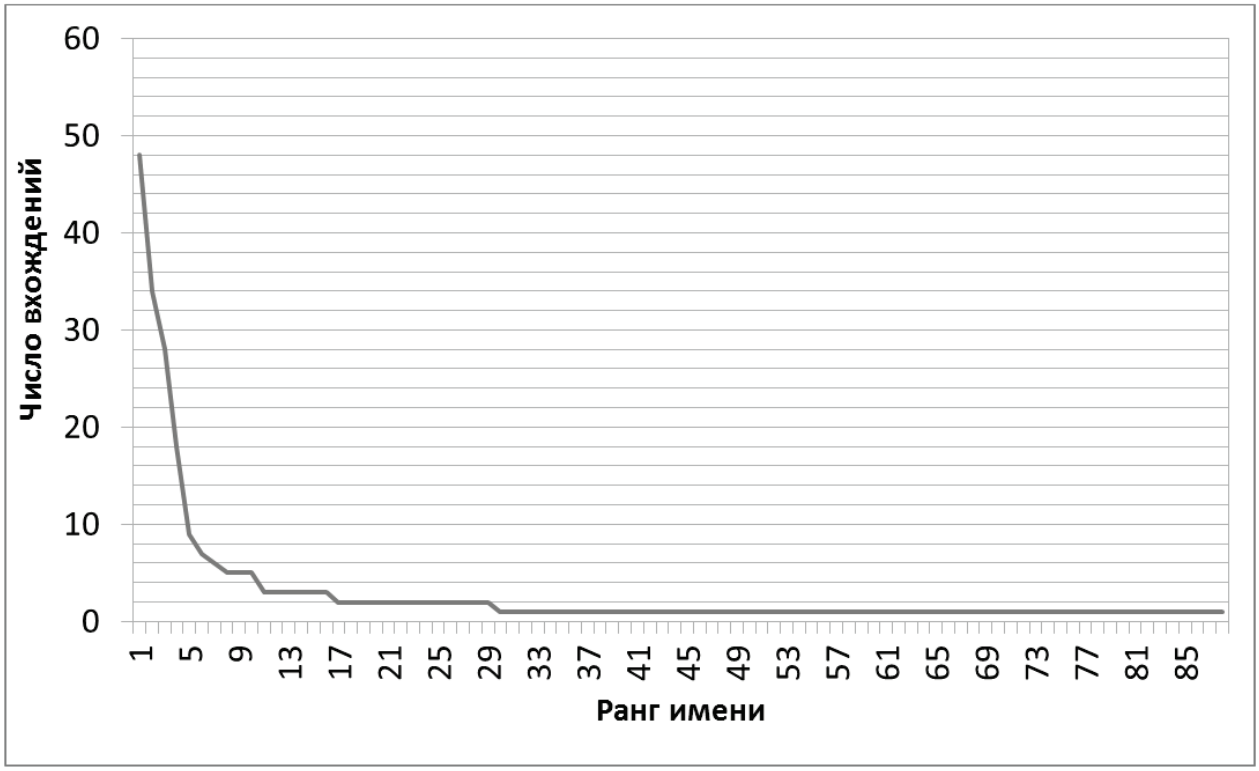

Рис.3. Распределение имен в романе «Война и мир»

В процессе анализа графика была выдвинута гипотеза, что распределение имен является распределением Парето [4]:

$$p(x,\theta) = \begin{cases} \frac{\alpha}{\theta} \cdot \left(\frac{\theta}{x}\right)^{\alpha+1}, & x \geq \theta, \\ 0, & x < \theta \end{cases}$$

где $\alpha > 0$ и $\theta > 0$ - параметры распределения. Распределение Парето (известное также в лингвистике под именем закона Ципфа) описывает зависимость абсолютной частоты слов в достаточно длинном тексте от ранга, а также кривую для популярности имен в больших группах населения. Для проверки гипотезы использовался критерий согласия Колмогорова-Смирнова. Статистика критерия для функции распределения превысила процентную точку распределения Колмогорова, следовательно, выдвинутая гипотеза была не верна и данное распределение не является распределением Парето. Полученный результат объясняется тем, что выборку имен в произведении отдельного автора нельзя считать случайной, поскольку выборка отражает субъективные предпочтения автора. Нельзя не принять во внимание и тот факт, что в выборке могли присутствовать ложные отсчеты из-за ошибочно выделенных имен.

Для проверки точности был поставлен эксперимент на сгенерированном тексте. За основу был взят роман Л.Н. Толстого «Война и мир». В произвольные места текста между словами с заданной вероятностью добавлялись части имен, затем анализировалось количество найденных имен в измененном тексте. Все новые имена, которые появились в результате этой операции, считались ошибочными. Объем анализируемого текста составлял порядка 110 тысяч слов. Результат эксперимента приведен на рисунке 4. В худшем случае на примерно 90 тысяч добавленных фрагментов имен число ложных срабатываний составило около 9 тысяч, т.е. менее 10 %.

Выводы и дальнейшая работа

Авторами был предложен способ автоматического выделение имен собственных из текстовых документов, особенностями которого является проверка полных имен по словарю, что уменьшает число ложных срабатываний по сравнению с методами, основанными исключительно на шаблонном поиске.

Задача отождествления имен, написанных в разных формах, представляет значительную сложность и пока не решена. Например, описанный подход не отождествляет имена Николай Андреевич и Николай Андреич или Анна Павловна, Анна Павловна Шерер и

Анна Шерер. Поскольку проверка фамилий с использованием словаря невозможна, то велик процент ложных срабатываний при обработке предложений вида:

Рис.4. Зависимость числа найденных имен от вероятности добавления части имени в текст

«Иногда Лариса Федоровна приподнималась на локте, подпирала подбородок ладонью и, разинув рот, смотрела на Юрия Андреевича». В этом случае слово «Иногда» выделяется как фамилия Ларисы Федоровны. Для решения указанных проблем необходимо реализовать механизм выделения имен с учетом контекста.

Кроме того, к недостаткам предложенной технологии следует отнести чувствительность метода к орфографическим ошибкам, знакам точки после сокращения имени.

В процессе анализа результатов эксперимента была опровергнута гипотеза о том, что распределение имен в произведениях художественной литературы является распределением Парето.

При этом следует отметить то, что средства семантического фреймворка Apache UIMA продемонстрировали принципиальную возможность их использования для выявления имён из текстов на естественном языке, простоту настройки и построения процесса обработки

текста в рамках предложенного способа. Несмотря на простоту реализации разработанного способа, получен удовлетворительный результат выделения имён.

Список литературы

1. Professional Alfresco/David Caruana, John Newton, Mike Farman. – Indianapolis, Indiana, USA: Wiley Publishing inc., 2010. – 576 с.
2. UIMA Overview & SDK Setup/Apache UIMA Development Community. Forest Hill, Maryland, USA: The Apache Software Foundation, 2009. – 58 с. [Электронный ресурс]. – Режим доступа: http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/pdf/overview_and_setup.pdf – Загл. с экрана. htm (дата обращения: 17.02.2013).
3. Сокирко А.В. Морфологические модули на сайте www.aot.ru [Электронный ресурс]. – Режим доступа: http://www.aot.ru/docs/sokirko/Dialog2004.htm (дата обращения: 17.02.2013).
4. Математическая статистика: Учеб. Для вузов/В.Б. Горяинов, И.В.Павлов, Г.М. Цветкова и др.; Под ред. В.С. Зарубина, А.П. Крищенко. -М.: Издательство МГТУ им. Баумана, 2001. - 424 с.