МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования

"МИРЭА - Российский технологический университет"

РТУ МИРЭА

Институт информационных технологий (ИТ)
Кафедра практической и прикладной информатики (ППИ)

Доклад
по дисциплине
«Анализ и концептуальное моделирование систем»

Выполнили студенты группы ИНБО-10-21 Четырин Б.П.

Приняла ассистент Свищёва И.В.

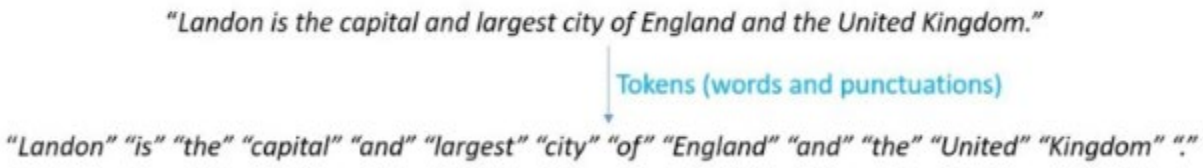
Способы автоматизированного извлечения знаний о предметной области из текстов электронных документов

Способы извлечения знаний из текстов электронных документов

1) Токенизация
2) Стемминг и лемматизация
3) Пометка частью речи (POST)
4) Распознавание именованных объектов (NER)
5) Извлечение словосочетаний
6) Контролируемые методы извлечения ключевых фраз

Токенизация

Токенизация относится к процессу разбиения текста на более мелкие токены. Это может включать разбиение абзацев на предложения, а предложений на слова, части слов или даже символы. Это фундаментальный шаг к созданию словаря, необходимого для выполнения любой задачи в НейроЛингвистическогоПрограммирования:



Одним из наиболее важных соображений при токенизации является определение границ. Например, в английском языке слова разделяются пробелами, но это может быть неверно в других языках. В зависимости от уровня, на который мы хотим разбить текст, мы можем выбрать токенизацию слов, токенизацию символов или токенизацию вложенных слов.

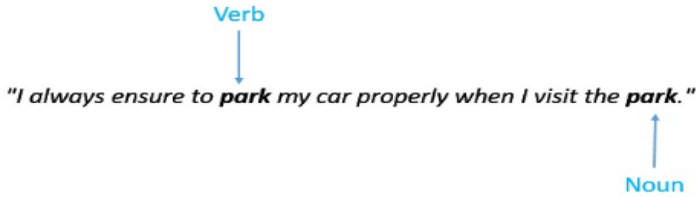
Стемминг и Лемматизация

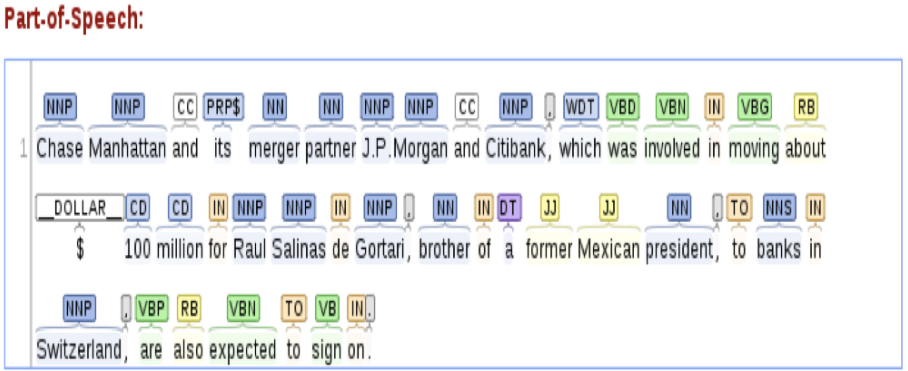
В естественных языках слова могут принимать различные формы, которые изменяют их грамматическое употребление, но не их семантическое значение. Например, travel, travelling, travels и travelled имеют разное употребление, но схожие значения. В области НЛП эти формы слова известны как словоизменительные формы. Для IR часто желательно привести все эти слова к их базовой форме:

Целью как стеммирования, так и лемматизации является генерация базовой формы изменяемых слов в тексте. Использование эвристического подхода с использованием стемминга является болеегрубым путем отсечения окончания слова для получения его базовой формы. Лемматизация более сложна и для достижения того же результата использует словарный запас и морфологический анализ слов.

Пометка частью речи (POST)

Часть речи, или просто PoS, - это категория слов со схожими грамматическими свойствами. В английском языке мы обычно выделяем девять частей речи, таких как существительное, глагол, артикль, прилагательное и другие. Мы можем использовать одно и то же слово в предложении как существительное или как глагол. Например, посмотрите на использование слова парк в этом предложении

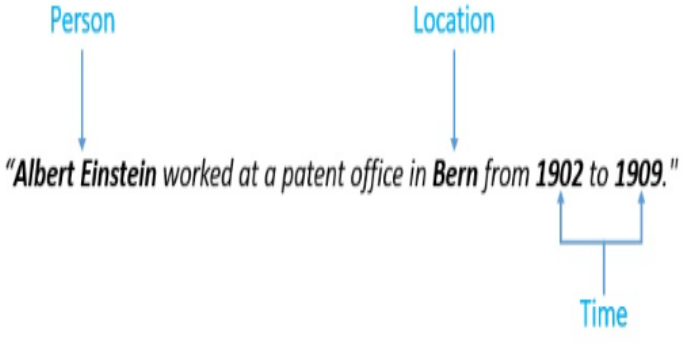


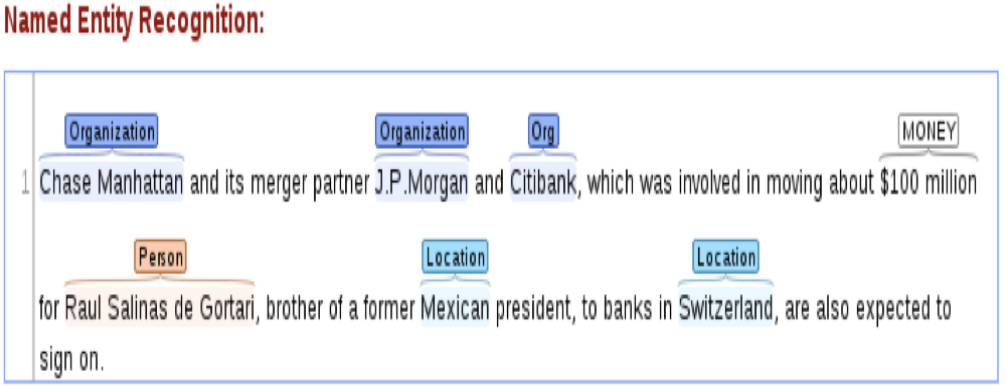


По сути, пометка частью речи - это процесс выделения слова в тексте как соответствующего определенной части речи в зависимости от его использования. Обычно мы используем основанный на правилах или стохастический алгоритм тегирования для достижения этого автоматически. PoS-теги имеют несколько применений в NLP

Распознование именованных объектов (NER)

Именованный объект относится к объекту реального мира, такому как города, люди или организации. Например, Baeldung, London, Jack Daniel, все они могут быть названы именованными объектами в любом тексте. В некоторых случаях именованные объекты могут также включать временные и числовые выражения, например 400 морских миль или 2020 год. Например, давайте проанализируем следующее предложение:

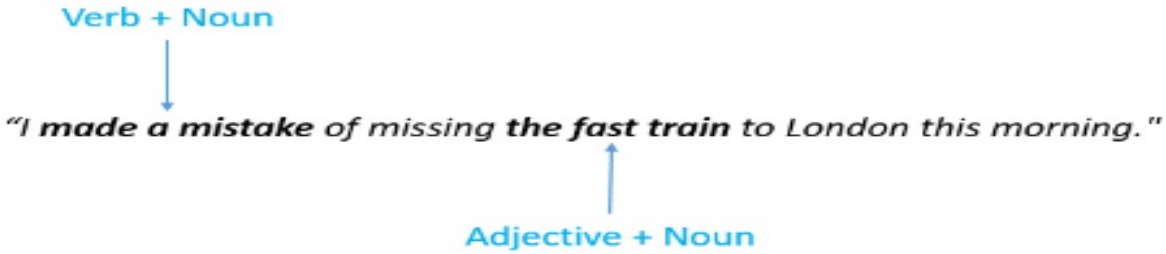




Распознавание именованных объектов - это процесс идентификации именованных объектов в неструктурированных текстах. Существует несколько основанных на правилах и статистических алгоритмов для автоматического распознавания именованных объектов. Статистическое распознавание именованных объектов обычно включает контролируемые и полууправляемые модели машинного обучения. Это имеет множество применений в IR, таких как классификация контента, индексирование и рекомендации.

Извлечение словосочетаний

Коллокация в основном относится к последовательности токенов, которые встречаются вместе в корпусе чаще, чем то, что мы можем считать совпадением. Это имеет скорее культурное значение, чем грамматическую ориентацию. Например, в обычной практике мы довольно часто используем такие фразы, как “крепкий чай” и “мощный компьютер”. Существует шесть основных типов словосочетаний:



Извлечение словосочетаний - это задача автоматического определения всех словосочетаний в тексте с использованием алгоритма. Один из простейших алгоритмов для определения словосочетаний использует их частоту в корпусе. Более сложные алгоритмы могут использовать методы, основанные на средних значениях и отклонениях, а также точечные взаимные информационные показатели. Это может иметь полезные применения в задачах IR.

Контролируемые методы

Методы извлечения ключевых фраз под наблюдением обычно работают, переформулируя проблему извлечения ключевых фраз в задачи классификации или ранжирования. При использовании подхода классификации нам интересно знать, подходит ли ключевая фраза-кандидат для представления текста или нет. Однако, поскольку это непростая задача, подход ранжирования пытается ранжировать кандидатов попарно на основе их релевантности

$$P[yes] = \frac{Y}{Y+N} P_{TF-IDF}[t|yes] P_{distance}[d|yes]$$

$$P[no] = \frac{Y}{Y+N} P_{TF-IDF}[t|no] P_{distance}[d|no]$$

KEA использует метод наивного Байеса для генерации модели

Существует несколько традиционных алгоритмов машинного обучения для обучения под наблюдением, которые мы можем использовать здесь. Например, наивный Байесов, деревья принятия решений, машины опорных векторов и многое другое. Однако некоторые конкретные реализации соответствуют всем требованиям больше, чем другие. Например, KEA - это метод бинарной классификации, который использует TF-IDF (t) и должность (d) первое появление для выбора ключевых фраз

Заключение

В этой статье мы обсудили фундаментальные концепции IR и NLP. В первую очередь мы сосредоточились на методах и алгоритмах, которые наиболее часто используются для автоматического извлечения ключевых фраз. Наконец, мы рассмотрели процесс и алгоритмы, которые мы можем использовать для идентификации и выбора ключевых фраз.

Table 1: Translating data into "Documents" in IR

Document	Field	Term
concept C	text	tokens in textual properties
relation R	text	tokens in textual properties
individual i	type	concepts that i belongs to
	subjOf	all relations R that $(i, R, ?)$ is a triple in data
	objOf	all relations R that $(?, R, i)$ is a triple in data
	text	tokens in textual properties of i

Спасибо за внимание!