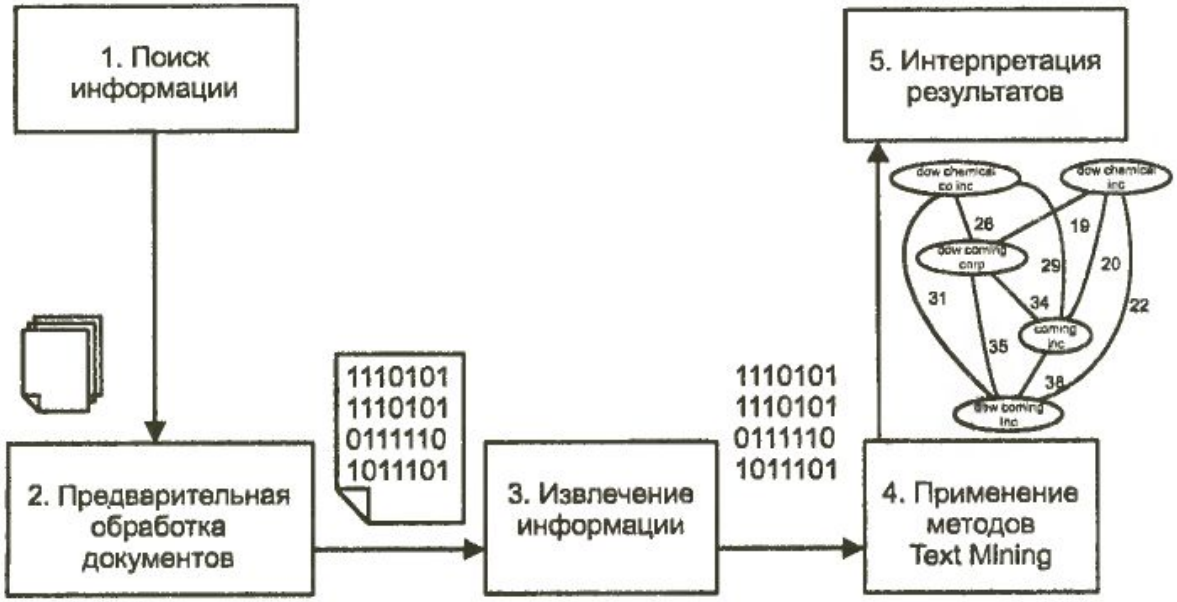
Text Mining. Анализ текстовой информации

Text Mining- методы анализа неструктурированного текста

Обнаружение знаний в тексте - это нетривиальный процесс обнаружения действительно новых , потенциально полезных и понятных шаблонов в неструктурированных текстовых данных **(набор документов, представляющих собой логически объединённый текст без каких либо ограничений на его структуру**:

- **web-страницы,**
- **электронная почта,**
- **нормативные документы и т.д.)**

Этапы Text Mining



Предварительная обработка текста

- **Удаление стоп-слов.**

Стоп- слов – вспомогательные слова, которые несут мало информации о содержании документа («так как», «кроме того»).

- **Стэмминг - морфологический поиск.**

Преобразование каждого слова к его нормальной форме.

(«сжатие», «сжатый» **->** «сжимать»)

- **Приведение регистра.**

«ТЕКСТ», «Текст» **->** «текст»

Задачи Text Mining

- **Классификация-** определение для каждого документа одной и нескольких заранее заданных категорий, к которой этот документ относится.
- **Кластеризация-** автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества.
- **Автоматическое аннотирование** - позволяет сократить текст, сохраняя его смысл
- **Извлечение ключевых понятий-** идентификация фактов и отношений в тексте
- Навигация по тексту – позволяет перемещаться по документам относительно тем и значимых терминов
- Поиск ассоциаций- идентификация ассоциативных отношений между ключевыми понятиями

Извлечение ключевых понятий из текста

Интерес представляют некоторые сущности, события, отношения. Извлечённые понятия анализируются и используются для вывода новых.

Извлечение ключевых понятий – фильтрация больших объёмов информации:

- отбор документов из коллекции ,
- пометка определённых терминов в тексте

Подходы к извлечению информации из текста

Определение частых наборов слов и объединение их в ключевые понятия (Apriori)

Идентификация фактов в текстах и извлечение их характеристик

Факты-некоторые события или отношения

Идентификация производится с помощью набора образцов.

Образцы-возможные лингвистические варианты фактов

Применение шаблонов

Извлечение ключевых понятий с помощью шаблонов

Локальный анализ

Петр Сергеевич Иванов покинул должность вице-президента известной фабрики ООО "Анкор". Его заменил Иван Андреевич Сидоров.

Лексический анализ. Текст делится на предложения и лексемы.

Словарь должен включать специальные термины, имена людей, названия городов, префиксы компаний…(«ООО», «ЗАО», «АО»)

Лексемы: «Петр», «Иван» - имена, «ООО» - префикс фирмы

Извлечение имён собственных (даты, денежные выражения). Имена идентифицируются с помощью образцов (регулярных выражений), которые строятся на основе частей речи, синтаксических и орфографических свойств.

- *Петр Сергеевич Иванов* с типом *"человек"*;
- *Иван Андреевич Сидоров* с типом *"человек"*;
- *ООО "Анкор"* с типом *"фирма"*.

В результате получим следующую структуру:

[имя собственное тип: человек Петр Сергеевич Иванов] покинул должность вице-президента известной фабрики [имя собственное тип: фирма ООО "Анкор"]. Его заменил [имя собственное тип: человек Иван Андреевич Сидоров].

Локальный анализ

Синтаксический анализ. Построение структур для групп имён существительных (имя сущ. + его модификации) и глагольных групп (глагол+ вспомогательные части)

1.Помечаются все основные группы имён сущ. меткой «сущ.»

2.Помечаются глагольные группы меткой «гл.»

[сущ. сущность: e1 Петр Сергеевич Иванов] [гл.: покинул] [сущ. сущность: e2 должность вице-президента] [сущ. сущность: e3 известной фабрики] [[сущ. сущность: e4 ООО "Анкор"]. [сущ. сущность: e5 Его] [гл.: заменил] [сущ. сущность: e6 Иван Андреевич Сидоров].

Для каждой группы имён существительных создаётся сущность. В нашем примере их 6.

- *e1* — тип: *человек*, имя: *"Петр Сергеевич Иванов"*;
- *e2* — тип: *должность*, значение: *"вице-президент"*;
- *e3* — тип: *фирма*;
- *e4* — тип: *фирма*, имя: *"ООО "Анкор"*;
- *e5* — тип: *человек*;
- *e6* — тип: *человек*, имя: *"Иван Андреевич Сидоров"*.

Локальный анализ

Наборы образцов используют для укрупнения групп имён существительных.
Образцы объединяют 2 группы имён существительных и промежуточные слова в большую группу

Образцы: описание фирмы, имя фирмы (фирма)

[сущ. сущность: e1 Петр Сергеевич Иванов] [гл: покинул] [сущ. сущность: e2 должность вице-президента известной фирмы ООО "Анкор"]. [сущ. сущность: e5 Его] [гл: заменил] [сущ. сущность: e6 Иван Андреевич Сидоров].

e3

Таким образом, список сущностей обновится следующим образом:

- e1 — тип: *человек*, имя: *"Петр Сергеевич Иванов"*;
- e2 — тип: *должность*, значение: *"вице-президент"* фирмы: *e3*;
- e3 — тип: *фирма*, имя: *"ООО "Анкор"*;
- e5 — тип: *человек*;
- e6 — тип: *человек*, имя: *"Иван Андреевич Сидоров"*.

Стадия интеграции и вывода понятий

- Для извлечения событий и отношений используются образцы, которые получаются за счёт расширения образцов, описанные ранее.

Событие преемственности должности извлекается с помощью следующих образцов : человек покинул должность, человек заменяется человеком

Группа имён сущ. «человек», «должность»

Выделяют две структуры событий

Глагольные группы. «покинул», «заменяется»

[событие: e7 Петр Сергеевич Иванов покинул должность вице-президента известной фирмы ООО "Анкор"]. [событие: e8 Его заменил Иван Андреевич Сидоров].

Список сущностей обновляется следующим образом:

- *e1* — тип: *человек*, имя: *"Петр Сергеевич Иванов"*;
- *e2* — тип: *должность*, значение: *"вице-президент"* фирмы: *e3*;
- *e3* — тип: *фирма*, имя: *"ООО "Анкор"*;
- *e5* — тип: *человек*;
- *e6* — тип: человек, имя: *"Иван Андреевич Сидоров"*;
- *e7* — тип: *покинул*, человек: *e1*, должность: *e2*;
- *e8* — тип: *заменил*, человек: *e6*, человек: *e5*.

Анализ ссылок. Разрешение ссылок , представленных местоимениями и описываемыми группами имён сущ.

«Его»(сущность е5).

Для разрешения этой ссылки будет выполняться поиск первой предшествующей сущности с типом "человек". В нашем примере такой сущностью является *е1*. В результате ссылки на *е5* должны быть заменены ссылками на *е1*. Таким образом, список сущностей и событий обновится следующим образом:

- *е1* — тип: *человек*, имя: *"Петр Сергеевич Иванов"*;
- *е2* — тип: *должность*, значение: *"вице-президент"* фирмы: *е3*;
- *е3* — тип: *фирма*, имя: *"ООО "Анкор"*;
- *е6* — тип: *человек*, имя: *"Иван Андреевич Сидоров"*;
- *е7* — тип: *покинул*, человек: *е1*, должность: *е2*;
- *е8* — тип: *заменил*, человек: *е6*, человек: *е1*.

Результат извлечения ключевых понятий из текста

В результате последовательности действий можно получить следующие извлечённые ключевые понятия.

	Событие	Человек	Должность	Фирма
1	Покинул	Петр Сергеевич Иванов	Вице-президент	ООО "Анкор"
2	Вступил	Иван Андреевич Сидоров	Вице-президент	ООО "Анкор"