Модели и методы формирования и использования баз знаний для интегрированных мультипредметных информационных систем

на примере обработки массива данных социальной сети ВКонтакте

ЦЕЛИ РАБОТЫ

- Разработка инструментов сбора больших данных
- Разработка инструментов оптимального хранения больших данных
- Учёт особенностей работы с большими данными соцсетей
- Разработка алгоритмов подготовки и анализа социальных данных
- Разработка алгоритмов визуального представления информации
- Разработка механизма подключения сторонних модулей обработки

ЧЕМ ИНТЕРЕСНЫ СОЦСЕТИ (ВОЗМОЖНОСТИ)

- Выявление настроений в обществе (Срез эмоционального фона)
- Прогнозирование реакции к будущим событиям.
- Поиск кадров - “автоматическое собеседование”.
- Автоматические соц.опросы. (более искренние и честные ответы)
- Создание новостных агрегаторов на базе социальной заинтересованности.
- Выявление деструктивной активности.
- Поиск связанных событий методикой поиска без ключевых слов.
- Карта активности (гео-метки, связанные с активностью в соц.сети)
- Выявление ботов

ОСОБЕННОСТИ ТЕКСТА СОЦИАЛЬНЫХ СЕТЕЙ

Текст:

- текст короткий
- частое применение нетекстовых символов
- сообщение может целиком состоять из нетекстовых символов
- частое намеренное искажение слов
- распространены различные синтаксические ошибки
- различная тематика обсуждения

Время жизни

- время жизни документов (публикация, социальная реакция, забвение).
- возможность отслеживания по времени

НЕОБХОДИМЫЙ ИНСТРУМЕНТАРИЙ

- сбор социальных данных
- хранение собранных и обработанных социальных данных
- хранение информации о динамике изменения отслеживаемых документов
- подготовка текста
- средства графического отображения полученных результатов
- средства получения различных выборок
- генерация наборов данных (датасетов) для инструментов тематического моделирования
- выявление тем (методики тематического моделирования)
- возможность быстрой интеграции пользовательский модулей обработки

core > sys_config.py > ...

```
'''
{'коллекция на сервере мониторинга':    {   entity_name     коллекция на сервере обработки
                                            metrics         какую инфу для этой коллекции собирать и её тип }
}

INFO_SET:
    entity_name     Имя сущности на сервере обработки
    fields          Поля, которые нужно хранить на сервере обработки
                    id          поля id, записываются единожды при создании документа на СВ, больше не меняются
                    counters    целочисленные счётчики (просмотры, лайки, ...)
                    text        текстовые поля
                    complex     часть полей является документом, нужно выгрызать из них вложенные поля и размещать на одном
                                уровне с остальными. Пример - City в группах. def proc_complex забирает нужное вложенное поле
                                на основании этого правила. Пока сделана обработка одного вложенного уровня.
                                vs_store_name   имя поля под которым сохранять на CBsdsd
                                inner_name      вложенное имя внутри поля container
                                Вложенная метрика строго целочисленная. Правило позволяет извлекать несколько вложенных полей
                                (*** можно добавить параметр-маршрут или уровень для работы с многоуровневыми составными полями)
                    variable_fields Поля, не всегда встречающиеся, которые можно заполнить просканировав другие поля документа.
    na_if           Not Active If 'field' == 'value'
    idife           Ignory Document if Empty поле. Например, нам ни разу не упукались посты или коменты с пустым полем text
    creation_date   Дата создания документа. 'ms' - из какого поля брать из документа ВК, 'vs' - под каким именем сохранать на СВ

Z_SYMBOLS_LEN       Timestamp слева дополняется количеством нулей указанных в этом параметре
'''
INFO_SET = {    'mon.posts':    {   'entity_name':  'posts',
                                    'fields':  {    'id':       ['id','from_id','owner_id','reply_owner_id','postponed_id'],
                                                    'counters': ['reposts','likes','views','comments'],
                                                    'text':     ['text'],
                                                    'raw':      ['attachments','comments','copyright','post_type','geo','signer_id',
                                                                 'copy_history','is_pinned','marked_as_ads',''],
                                                    'str_date': [],
                                                    'complex':  {},
                                                    'variable_fields': []
                                                    },
                                    'na_if':                {},
                                    'idife':                ['text'],
                                    'extract':          {},
                                    'creation_date':        {'ms':'date', 'vs':'date'}
                },
```

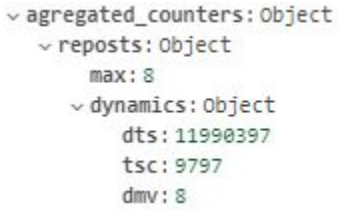
ПАРАМЕТРЫ СЕРВЕРА ОБРАБОТКИ

ПОДСЛУШАНО АПАТИТЫ


```
id: 61904449
› members_count: Object
› name: Object
  name__last: "Подслушано Апатиты"
  name__last_ts: "000001622142004"
  name__last_clean: "Подслушано Апатиты"
› description: Object
  description__last: "У нас вы найдете откровения и новости жителей г. Апатиты и г. Кировск ..."
  description__last_ts: "000001622142004"
  description__last_c… : "У нас вы найдете откровения и новости жителей г. Апатиты и г. Кировск ..."
⌄ city: Object
    000001622142004: "Апатиты"
  city__last: "Апатиты"
  city__last_ts: "000001622142004"
  city__last_clean: "Апатиты"
› country: Object
  country__last: "Россия"
  country__last_ts: "000001622142004"
  country__last_clean: "Россия"
› contacts: Array
› counters: Array
  cover: ""
› links: Array
  place: ""
› site: Array
› status: Array
  trending: ""
› verified: Array
› wall: Array
  public_date_label: ""
  market: ""
› main_section: Array
  fixed_post: ""
› age_limits: Array
  wiki_page: ""
  last_mon_ts: "000001648396805"
› agregated_counters: Object
  rating: 0
  ntp: 1
```

ТИПИЧНЫЙ ПОСТ

```
date: "000001622055649"
id: 2046224
from_id: -61904449
owner_id: -61904449
reply_owner_id: 0
postponed_id: 0
> reposts: Object
> likes: Object
> views: Object
> comments: Object
> text: Object
text__last: "Может кто знает где режут  зеркальную плитку? ( нужно вырезать отверст..."
text__last_ts: "000001622142008"
text__last_clean: "Может кто знает где режут зеркальную плитку? нужно вырезать отверстия ..."
attachments: ""
copyright: ""
> post_type: Array
geo: ""
> signer_id: Array
copy_history: ""
is_pinned: ""
> marked_as_ads: Array
: ""
last_mon_ts: "000001622400008"
> agregated_counters: Object
rating: 67
ntp: 1
```



ХРАНЕНИЕ ДИНАМИКИ СОЦИАЛЬНОЙ АКТИВНОСТИ

Метрики динамики социальной активности:

dts Diff Timestamp Sum
Сумма временных “разрывов” между изменениями метрик

dmv Diff Metrics Values
Сумма изменений метрик между срезами

tsc ts_count
Количество временных меток (срезов) метрики

Расчёт рейтинга документа

```
DOCUMENT_RATING_K = {'reposts':20, 'comments':10, 'likes':5, 'views':1}
```

$$i = \frac{K_R \times Reposts + K_C \times Comments + K_L \times Likes \times V_v}{(Views - 1) \times 100}$$

, где

$$V_v = \frac{\sum ts_{v+1} - ts_v}{n}$$

V - скорость изменения метрик поста. Временные промежутки между изменениями метрик делим на количество этих изменений (n).

ts - TimeStamp поста

Социальная активность ВК

Метрики:

○ likes ● reposts ○ comments ○ views ○ LRCV-grow ○ rating-ss

Сообщества:

× Кольский Север | Апатиты Мурманск Хибины и др × [X] Курорт Большой Вудъявр | BigWood ski resort × [X] Лавина (Апатиты | Кировск | Хибины) × Наши Апатиты × Подслушано Апатиты
× [X] Подслушано Кировск Хибины × [X] Профсоюзная организация "ФОСАГРО-АПАТИТ" × [X] Рабочее время | Апатит × [X] СУРОВЫЕ ХИБИНЯ × [X] Твой Кировск × [X] Типичный Комбинат "Апатит"
× Хибины Апатиты Кировск Народное ТВ × [X] Чёрный | Белый список Апатиты

Get current dataset

Период, дн.:

Label 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29

Документов:

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29

Поделились | Документов: 10 | Период: 3 дня

(5) Кто спорил? Победила "Пятерочка"☺ новый сетевик возле Ре-Монта.

👁18555 | ♡194 | 💬69 | ↩451

Подслушано Апатиты

(1) Кто спорил? Победила "Пятерочка"☺ новый сетевик возле Ре-Монта.

👁18555 | ♡194 | 💬69 | ↩451

Подслушано Апатиты

(2) На BigWoode в Кировске погиб человек Появилось видео с места событий

👁16709 | ♡66 | 💬73 | ↩347

Подслушано Апатиты

(3) На BigWoode в Кировске погиб человек Появилось видео с места событий

👁12324 | ♡52 | 💬46 | ↩218


0,94
0,85
0,57
0,6
0,45
0,23
0,2