Обработка естественного языка и понимание речи

Воронцов Константин Вячеславович

● Лаборатория машинного интеллекта,
Московский Физико-Технический Институт ●
● ООО «Айтея» ●
voron@forecsys.ru

Обработка естественного языка и понимание речи

1. Задачи и методы анализа текстов
   - Задачи обработки естественного языка
   - Что такое «понимание» и что такое «смысл»?
2. Примеры задач классификации текстов
3. Примеры задач тематизации текстовых коллекций

Задачи обработки естественного языка

Вспомогательные лингвистические задачи:

Sequence-to-sequence Learning
Ontology Learning
Topic Modeling
Word Embedding
Word Sense Disambiguation
Semantic Role Labeling
Named Entity Recognition
Automatic Term Extraction
Parsing Syntax Analysis
Part-of-speech Tagging
Lemmatization

Конечные бизнес-задачи:

Conversational Intelligence
Machine Translation
Question Answering
Information Retrieval
Sentiment Analysis
Text Summarization
Text Segmentation
Text Classification
Text Clustering
Intent Recognition
Fact Extraction

Выделение смысла? Понимание речи?

- Поиск «смысла» бессмысленен
- Что такое «понимание», не понятно
- Бизнес, технологии и математика работают только с чётко определяемыми понятиями и чётко поставленными задачами
- Задача чётко поставлена, если для неё описано **ДНК**:
  «что **Д**ано – что **Н**айти – **К**ритерий качества решения»
- Измеримый критерий появляется, когда цели прагматичны:
  – автоматизация рутинных операций
  – повышение производительности труда
  – снижение издержек

Обработка естественного языка и понимание речи

1. Задачи и методы анализа текстов
2. Примеры задач классификации текстов
   - Распознавание шаблонных фрагментов
   - Тегирование звонков в контактный центр
   - Классификация отзывов клиентов по известным категориям
3. Примеры задач тематизации текстовых коллекций

#1: Выделение значений параметров

Цель: автоматизировать анализ конкурсной документации по госзакупкам:

Задача: находить и выделять в текстах значения параметров:

- Дата начала выполнения работ
- Дата окончания действия контракта
- Размер обеспечения заявки
- Наличие аванса

Критерий: точность распознавания размеченных полей

Метод: фиксированные или обучаемые правила

Результат: точность близка к 100%

#1: Сложность задачи – разнообразие фраз

Пример. Встречающиеся способы описания *даты начала работ*:

…контракт вступает в силу с момента заключения контракта и действует по 31 марта 2019 г. включительно
…срок выполнения работ: начало: 15 декабря 2017 года; окончание – 31 декабря 2018 года
…настоящий договор действует до исполнения обязательств сторонами, но не позднее 31 декабря 2019 года
…срок действия которых истекает не ранее 25 сентября 2019 г
…срок оказания услуги: с момента заключения договора (но не ранее 09.01.2017 г.) по 31.12.2017 г.

#2: Тегирование звонков в контакт-центр

Цель: оценивание результативности маркетинговых акций,
оценивание рекламных площадок,
оценивание качества работы операторов

Задача: определение намерений клиента и результата разговора

Критерий: точность, полнота, F1-мера по размеченной выборке записей разговоров

Методы: логистическая регрессия с отбором признаков,
кросс-валидация со стратификацией классов

Результат: F1-мера от 50% до 90% в зависимости от класса

#2: Звонки в автосалоны

Классификация разговоров:

- запись на тех. обслуживание
- автомобиль в кредит
- договор о встрече
- trade-in

Вспомогательная задача:

- определение марки и модели автомобиля

Категория	Precision	Recall	F-1 score	Accuracy
Договор о встрече	53,6%	89%	66,9%	75,5%
Trade-In	46,5%	58,6%	51,9%	80,7%
Тех. Обслуживание	66,6%	92,6%	77,5%	93,2%
Марка автомобиля	86,6%	93,7%	90%	81,7%
Модель автомобиля	66,3%	78%	71,7%	55,8%

#2: Звонки в риэлторские компании

Классификация разговоров:

- договорённость о встрече
- договорённость о перезвоне
- готовность к оплате
- ипотека

Вспомогательная задача:

- параметры объекта недвижимости
- наличие нецензурной лексики

Категория	Precision	Recall	F-1 score
Договор о встрече	45%	60%	51%
Договор о перезвоне	79%	73%	76%
Ипотека	61%	66%	64%
Квартира в аренду	66%	80%	73%
Покупка квартиры	87%	90%	88%

#2: Интерпретируемость отбора признаков

Пример 1: Авто в кредит

- *взнос кредит кредитный ставка процент кредитование процентный условие программа рассчитать посчитать платёж* встреча *банка* сожаление клиент зачёт сдавать визитка подать сервис акция хотеться ожидание срок смочь знать покупка самый выбрать отправить записаться брать чёрный встретиться новое обсудить

Пример 2: Ипотека

- *ипотека банк* компания справка площадь отделка условие *ндфл* документ *втб* отдел собственность принцип проблема одобрение апартамент история номер комнатка контакт *сбербанк* этаж объект станция сдача адрес дело улица размер знать ремонт лицо ставка планировка новое консультация процесс смска недвижимость координата

#3: Классификация отзывов по категориям

Цель: анализ отзывов потребителей по каналам обратной связи (горячая линия, ВК, telegram, mail, форум…)

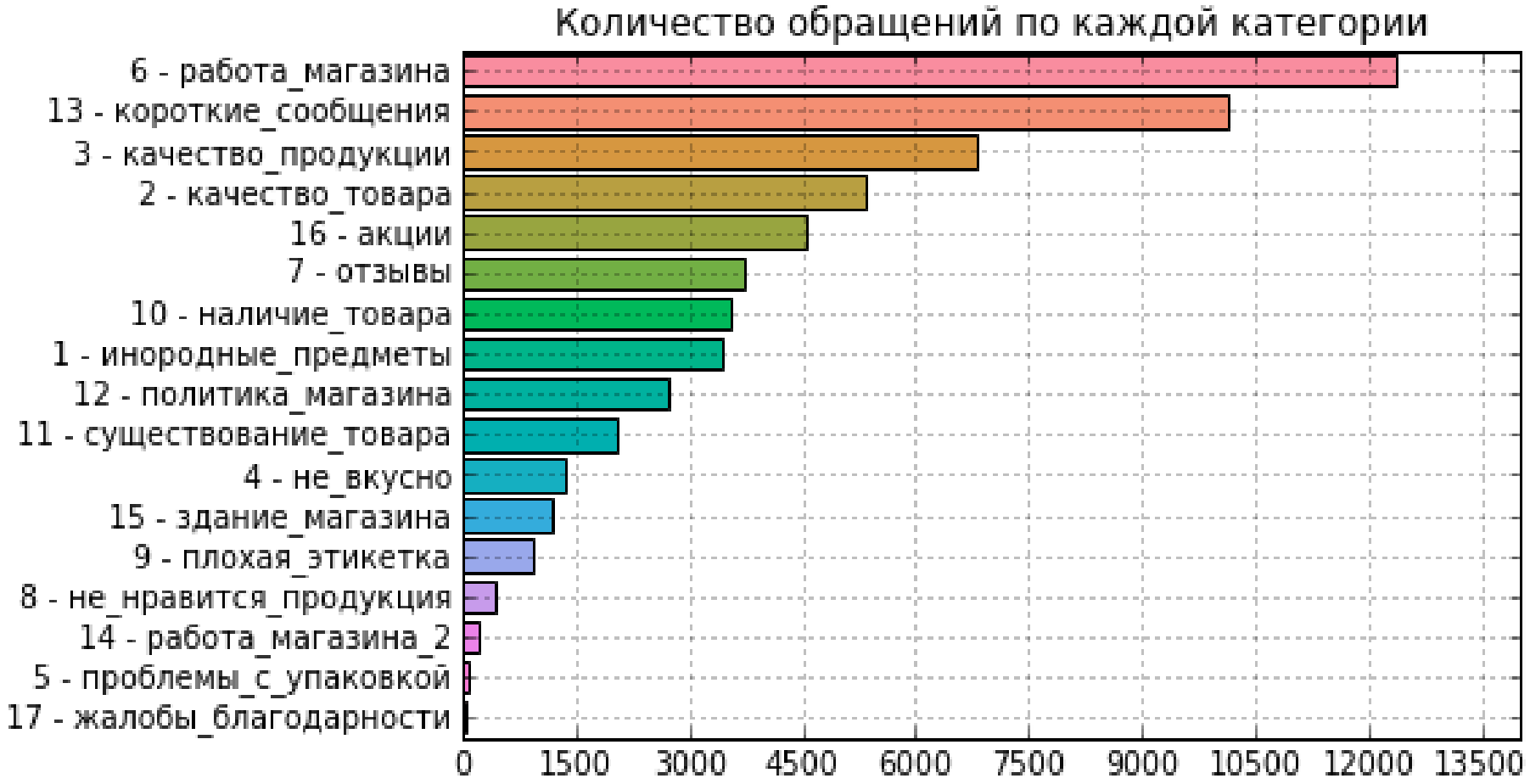
Задача: классификация отзывов по 17 *известным категориям*

Критерий: точность, полнота, F1-мера по размеченной выборке

Методы: логистическая регрессия,
градиентный бустинг,
отбор признаков,
кросс-валидация со стратификацией классов

Результат: F1-мера от 50% до 90% в зависимости от класса

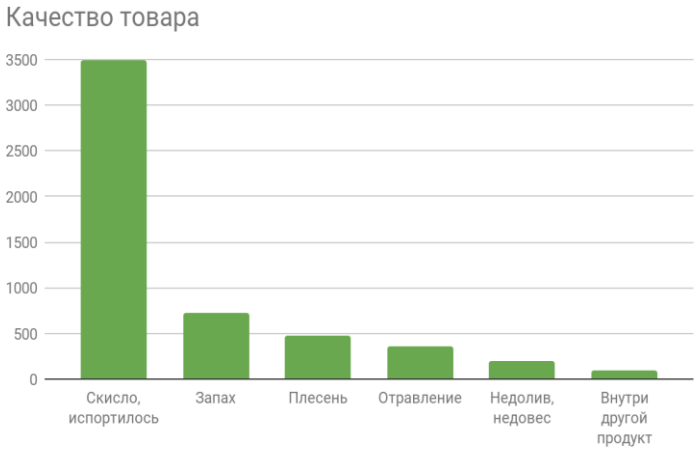
#3: Сложность задачи – несбалансированность классов

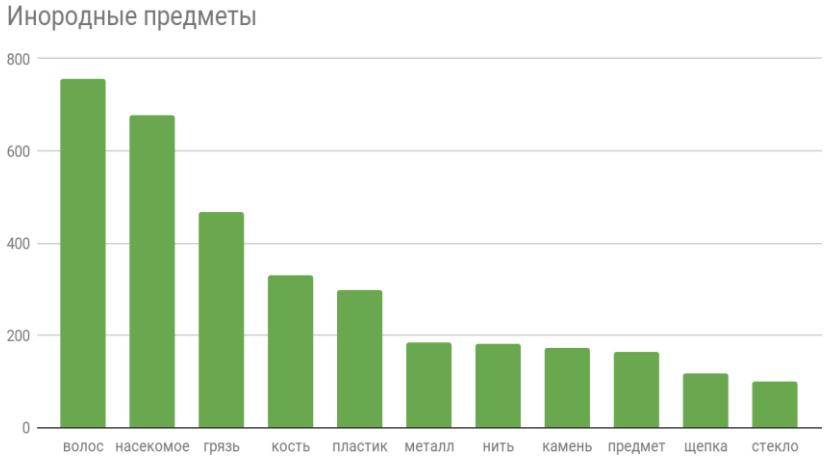


#3: Примеры отчётов

Отчёты строятся

- по категориям проблем
- по времени
- по каналам коммуникации





Обработка естественного языка и понимание речи

1. Задачи и методы анализа текстов
2. Примеры задач классификации текстов
3. Примеры задач тематизации текстовых коллекций
   - Классификация отзывов клиентов по заранее неизвестным темам
   - Тематическая сегментация записей разговоров контакт-центра
   - Разведочный тематический информационный поиск

#4: Кластеризация отзывов по темам

Цель: анализ отзывов пользователей приложения для оформления заказов в ресторанах быстрого питания

Задача: кластеризация (тематическое моделирование) отзывов по кластерам (темам), которые *заранее не известны*

Критерий: точность, полнота, F1-мера по размеченной выборке

Методы: тематическое моделирование (BigARTM), анализ тональности

Результат: F1-мера улучшилась от 61% (простейшая модель) до 81% (+модальности +частичное обучение)

#4: Результат тематического моделирования

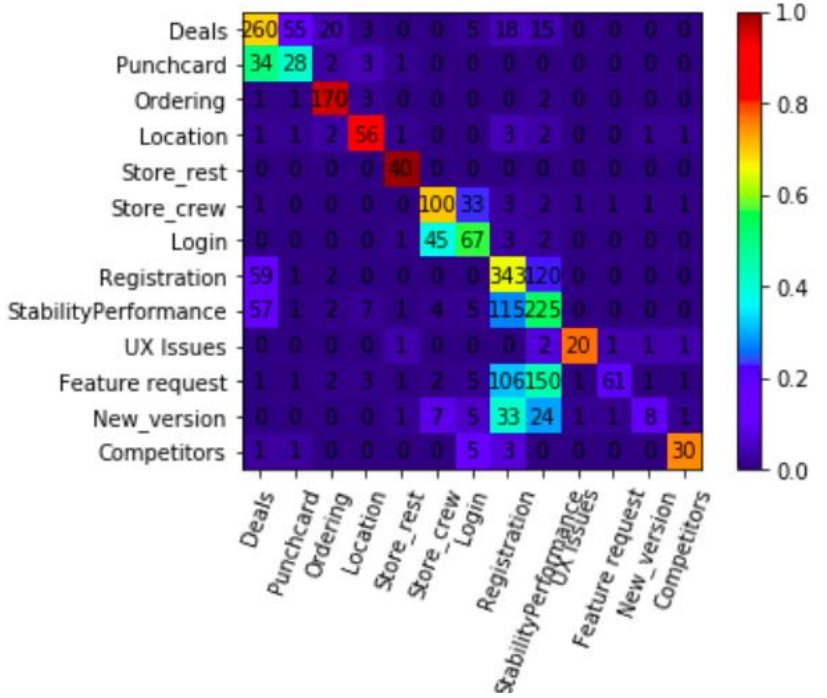
Deals • *deal, coupon, menu, offer, drink, meal, sandwich, price, option, beverage*	**Punchcard** • *punchcard, punch, coupon_code, show, save, free, coffee, price, drink, reward*	**Competitor mentions** • *restaurant, glitch, pass,log_in, get_passed, download, waste, crash, uninstall*
Ordering • *menu, order, mobile, card, drink, meal, price, discount, purchase, cashier*	**Stability & Performance** • *buggy, load, freeze, slow, terrible, fix, install, down, version, app*	**Location** • *location, gps, place, gps_location, address, map, place, store, work, find_location*
UX Issues • *crash, usability, trouble, terrible, password, install, review, login, get, uninstall*	**Store Issues (Crew)** • *employee, work, train, fix_issues, scan, uninformed, problem, use_app, customer, app*	**Store Issues (Restaurant)** • *order, food, lunch, update, favourite, open, place, menu, card, mobile*
Registration • *register, log_in, get_passed, download, install, account, load, password, notification*	**Login** • *log, glitch, pass, log_in, get_passed, waste, crash, uninstall, account, sign_in*	**Feature Request** • *find, price, device, calculator, deal, location, permission, password, new, app*
New version/Release Issue • *update, version, new, change, bad, work, release, new_app, bug, memory*	**Other** • *food, lovin_it, love, like, wonderful, awesome, delicious, work, save_money*	**These notable keywords were automatically detected by the topic model for each top-level category. No manual filtering was done.**

#4: Меры по улучшению качества модели

- Ручная пост-фильтрация словарей тем
- Фиксация тем для некоторых (10%) размеченных документов
- Использование модальности для размеченных документов

Объединение трёх подходов
дает наилучшее качество:

- *Accuracy*: **81%**
- *Precision*: **78.2%**
- *Recall*: **67.3%**
- *F1:* **72.3%**



#5: Тематическая сегментация записей разговоров контактного центра

Цель: мониторинг качества работы операторов,
выявление лучших практик,
генерация подсказок операторам

Задача: разбиение разговора на короткие тематические сегменты,
построение графа переходов между темами

Критерий: качество тематической сегментации

Методы: выделение терминов,
синтаксический анализ,
тематическое моделирование (BigARTM)

Результат: качество сегментации (доля правильно выделенных сегментов) возросло от 40% до 75%

#5: Тематическая сегментация
(подчёркиванием выделена асессорская разметка)

Оформление заявки | Индивидуальный подход | Решение банка | Доставка
Бонусная программа | Бесплатная доставка/оформление

вот на данный момент я звоню предлагаю только составить заявку чтобы банк изучил вашу кредитную историю и подобрал под вас индивидуальный тарифный план после чего на ваш мобильный поступит уведомление в котором будет указано каким образом в случае положительного ответа будут доставлены бумаги у нас есть два способа доставки это либо курьерская доставка либо заказным письмом почтой России

ну вот бонусы значит на все абсолютно покупки один процент а если вы совершаете покупки у банка будет полный перечень магазинов у вас в личном кабинете до тридцати процентов бонусов можете то есть вот две тысячи что то купили а ориентировочно шестьсот вернулось вам на это уже плюсов согласитесь что это довольно таки это одна покупка вот так и хочу сказать что вы абсолютно ничего не теряетесь соглашаясь оформить заявку ничего за что не платите потому что вам карту выпускают доставляют абсолютно бесплатно вам либо представитель банка привозит либо по почте она приходит

#6: Разведочный информационный поиск

Цель: поиск документов по смыслу, а не по ключевым словам, разведочный информационный поиск

Задача: многокритериальное тематическое моделирование, ранжирование документов по тематическому сходству

Критерий: точность и полнота поиска по размеченной выборке (коллекции habrahabr.ru и techcrunch.com)

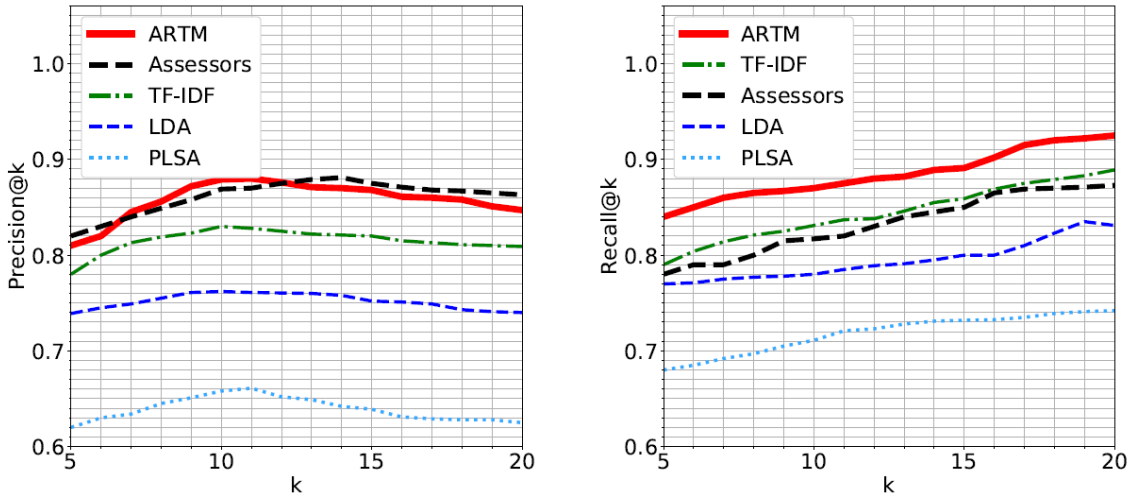
Методы: выделение терминов (TopMine), тематическое моделирование (BigARTM)

Результат: точность и полнота поиска возросли с 65% до 89%, автоматизация задач поиска, требующих около 30 мин.

#6: Разведочный информационный поиск

- Длинные запросы (1 стр. А4)
- 100 запросов
- 3 асессора на каждый запрос
- 30 минут в среднем на запрос
- Разметка на Яндекс.Толока
- Коллекции техно-новостей:

Результат: *точность* (precision) и *полнота* (recall) поиска



A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сухой остаток

- Невозможно поставить задачу «понимания речи» или «понимания смысла текста», когда нет конкретной цели
- Легче ставить задачи автоматизации обработки текстов, когда есть конкретные бизнес-цели, формализуемые с помощью измеримых критериев качества решения
- Для решения таких задач не обязательно создавать сложные универсальные инструменты
- Они решались, решаются, и будут решаться различными специализированными методами NLP и ML
- Наиболее критичны – объём и чистота обучающих данных