Выявление новизны в потоках сложно структурированных данных для задач обнаружения аномалий

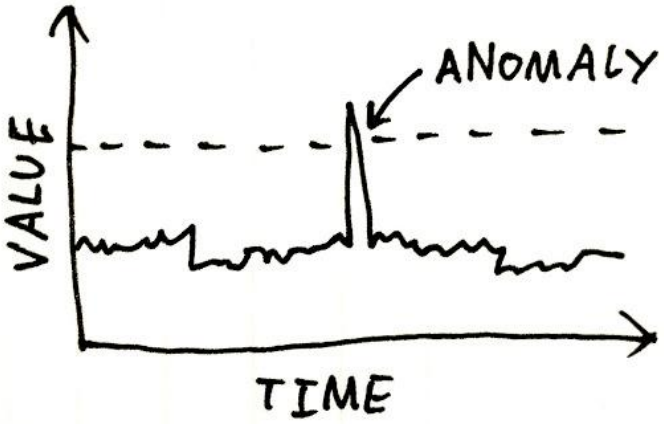
Студент 3 курса ВМК МГУ
Калашников Дмитрий Павлович

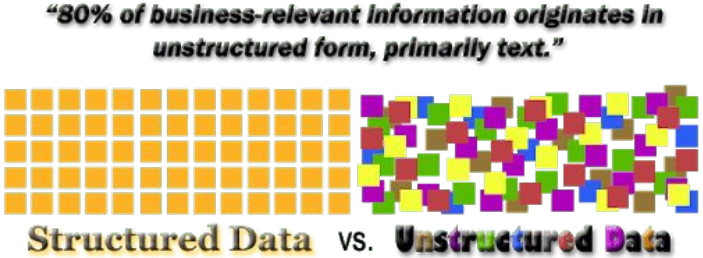
Научный руководитель:
Горохов Олег Евгеньевич

Введение

Ключевые понятия:

- Аномалии
- Потоки данных
- Новизна
- Сложно структурированные данные



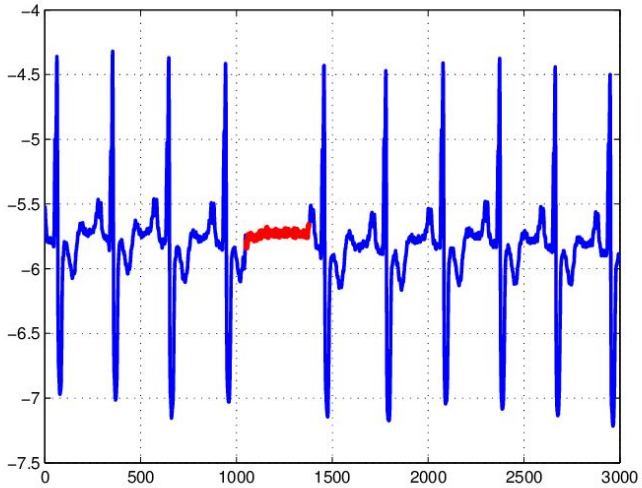


Введение. Аномалии

Аномалии - это закономерности в данных, которые не соответствуют четко определенному понятию нормального поведения.

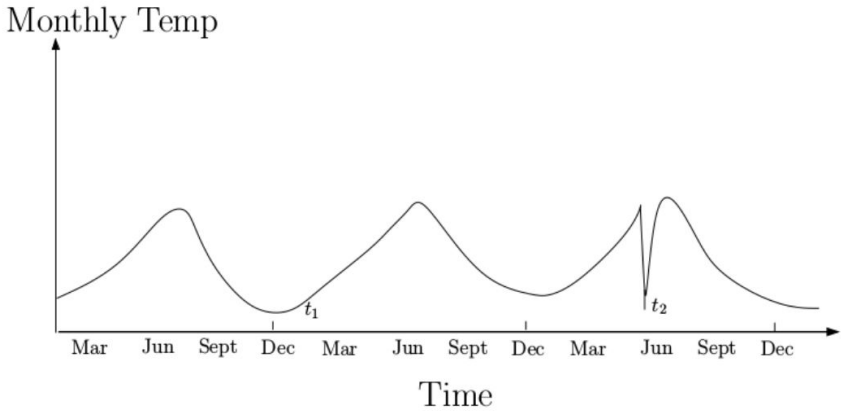
- точечные
- контекстуальные
- коллективные [1]

- ❏ Выброс в измерениях температуры
- ❏ Большая сумма трат в обычный день
- ❏ Статичный регион в кардиограмме
- ❏ Последовательность использованных протоколов в системе

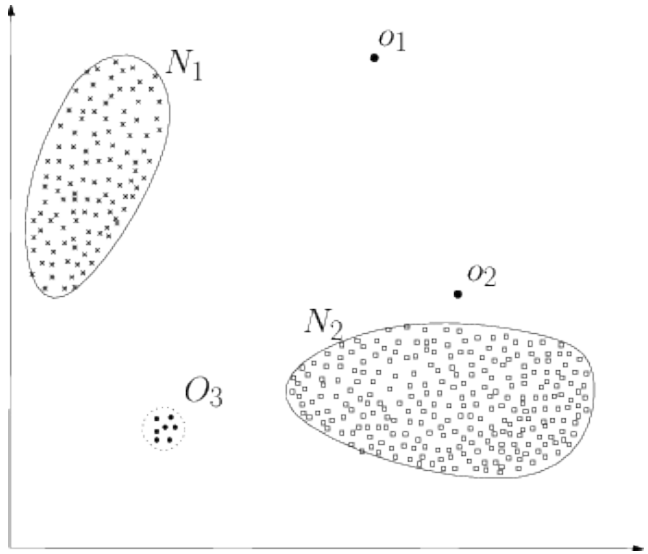


Введение. Аномалии

- точечные
- контекстуальные
- коллективные



контекстуальные аномалии



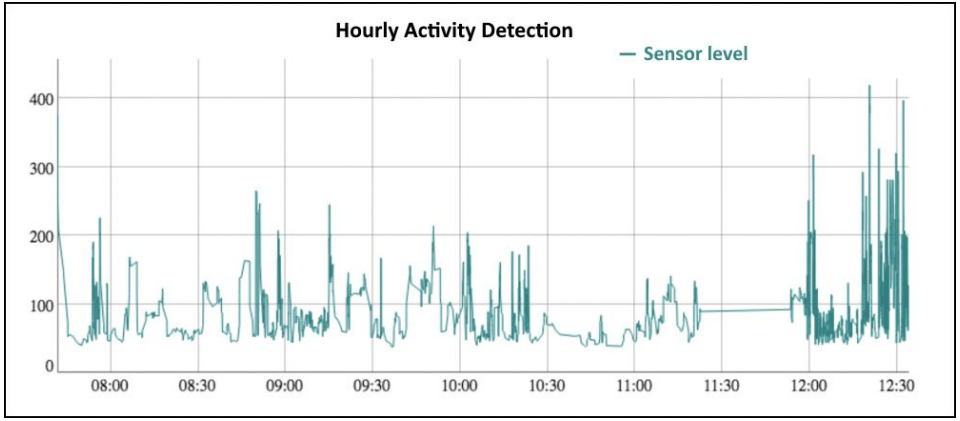
точечные и коллективные аномалии

Введение. Поток данных

Потоки данных представляют собой объемные, непрерывные, неограниченные, упорядоченные последовательности данных, поступающие с высокой скоростью и меняющиеся с течением времени.

Не могут поместиться в память и сканироваться несколько раз, как традиционные данные[2].

- транзакции в банке
- поисковые запросы
- данные сенсоров

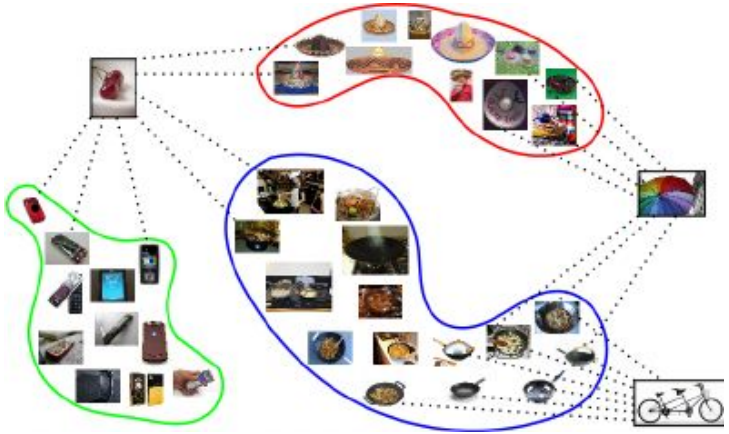


Введение. Новизна

Выявление новизны - обнаружения отличий в полученных на текущий момент данных от полученных ранее.

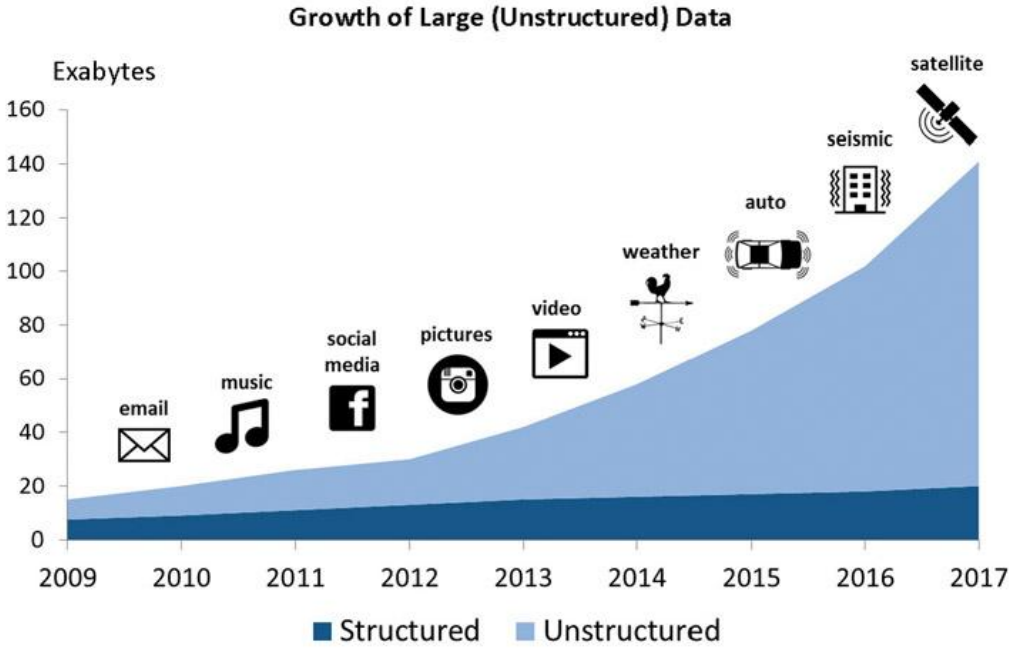
Термин “новизна” указывает на только появляющееся или совершенно новое понятие, которое должно быть включено в нормальный паттерн.

Проблема: обучающая выборка содержит в себе примеры только нормального класса[3].



- новые темы
- новые события
- новостные сюжеты в коллекции документов

Введение. Неструктурированность



Введение. Неструктурированность

Классификация данных по структуре[4]:

- Неструктурированные, или сложно структурированные данные. Не организованы предопределенным образом или не имеют предопределенной модели данных. Например, текстовые документы, PDF, изображения видео.
- Полуструктурированные данные. Обладают некоторыми организационными свойствами, облегчающими анализ. Например, XML-файлы.
- Структурированные данные. Имеют определенную модель данных, формат и структуру. Например, база данных.

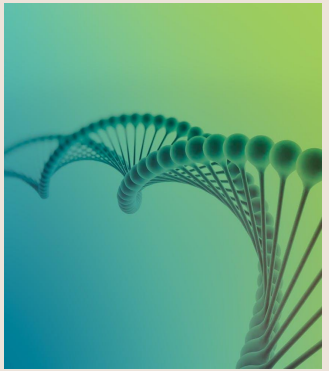
Введение. Неструктурированность

Сложно структурированные данные - данные, не имеющие предопределенной модели данных.

Сложно структурированные данные можно представить в виде текста. Возникает задача предобработки текстовых данных[5].

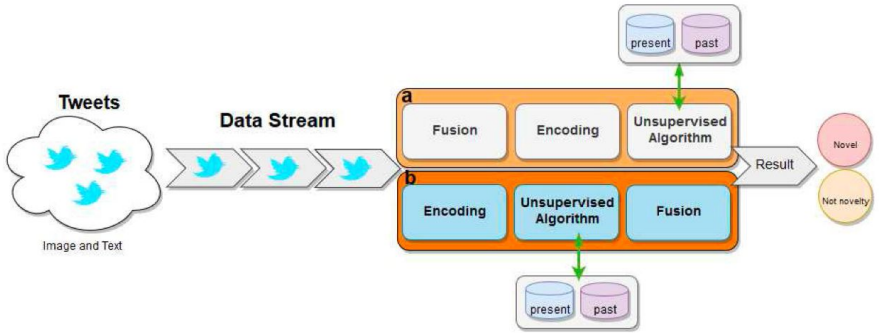
- электронные письма
- книги
- информация с датчиков
- системные журналы

Актуальность



Актуальность

- *Биоинформатика*: изменение нормального паттерна в генах с течением времени[8].
- *Медицина*: имеем набор снимков, знаем некоторые визуальные паттерны заболевания, но могут существовать и другие[7].
- *Производство*: выявление неисправностей двигателя - отсутствие подобных прецедентов в прошлом[7].
- *Общество*: появление новой темы для обсуждения в социальных сетях[1].
- *ИИ и робототехника*: развитие модели биологического существа. Способность отделять известные паттерны от принципиально новых[9].



Актуальность

Задача обнаружения новизны имеет широкое применение в разных областях[1, 6, 7, 8, 9].

- Существует широкий класс задач, где приходится иметь дело не с фиксированным датасетом, а с потоком данных.
- Большинство данных приходится собирать в сложно структурированном виде.
- Классические алгоритмы классификации не позволяют корректно обнаруживать новизну в потоках сложно структурированных данных.
- Часто, данные содержат мало объектов и много признаков(что характерно для сложно структурированных данных). В таких случаях задача выявления новизны становится существенно сложнее.

Постановка задачи

- Исследование и разработка методов выявления новизны в потоках сложно структурированных данных.

Постановка задачи

Требуется:

- изучить существующие подходы к каждому из этапов решения задачи выявления новизны в потоках сложно структурированных данных;
- провести сравнение эффективности алгоритмов внутри каждой подзадачи для различных наборов данных;
- выявить комбинацию рассмотренных методов решения подзадач, дающую наилучший результат в основной задаче для различного набора данных.

Этапы решения задачи

- Исследование существующих подходов обнаружения новизны в потоках сложно структурированных данных.
  - Предобработка данных.
  - Построение признакового пространства.
  - Сокращение признакового пространства(для определенных моделей).
  - Построение моделей обнаружения новизны.
- Разработка алгоритма выявления новизны в данных.
- Экспериментальная оценка предлагаемого подхода.

Исследование существующих методов обнаружения новизны

Цель обзора:

Исследовать существующие подходы к решению поставленной задачи на каждом из этапов, провести их сравнительную характеристику и выбрать наиболее подходящие решения для дальнейшего практического анализа.

Этапы исследования задачи:

- предобработка
- векторизация
- сокращение размерности
- построение модели

Предобработка текста

Стемминг

Лемматизация

Аугментация

Дедубликация

Извлечение отношений*

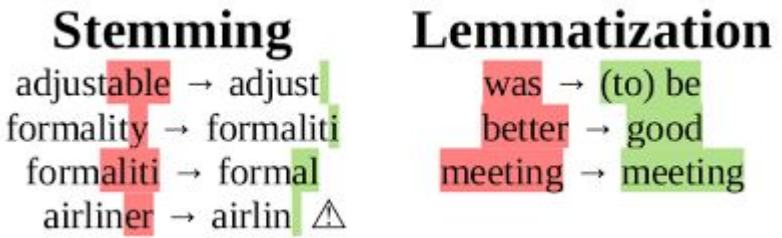
Обзор. Предобработка текста

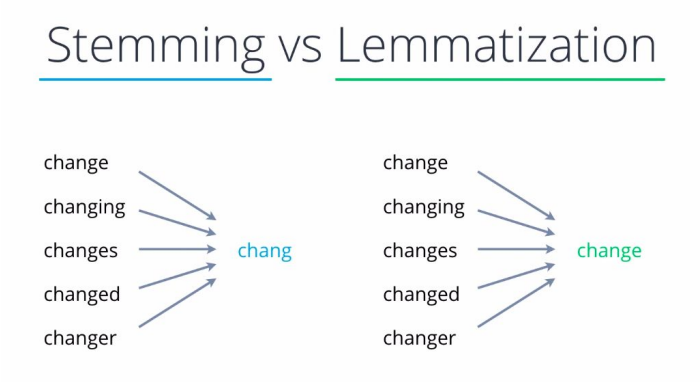
Стемминг - выделение из слова псевдоосновы, заключается в отсечении с начала и конца слова его частей.

Эффективно выделяет ключевые слова в тексте VS трудно расширяется на другие языки[10,11].

Лемматизация - приведение слова к нормальной словарной форме. Например - инфинитив; ед.ч. им.п м.р.

Сохраняет больше информации о слове VS порождает очень большой словарь корпуса[12].



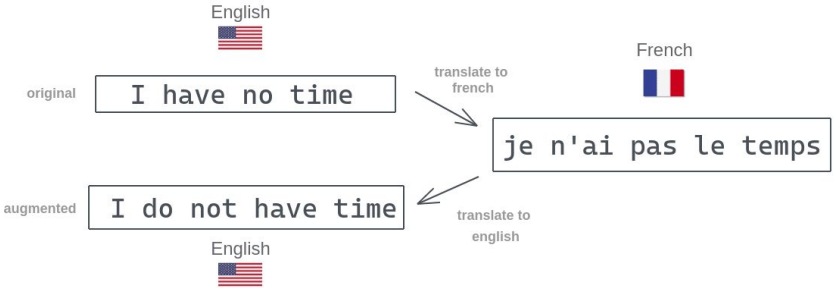


Обзор. Предобработка текста

Аугментация

Проблема: имеем мало объектов и много признаков[18].

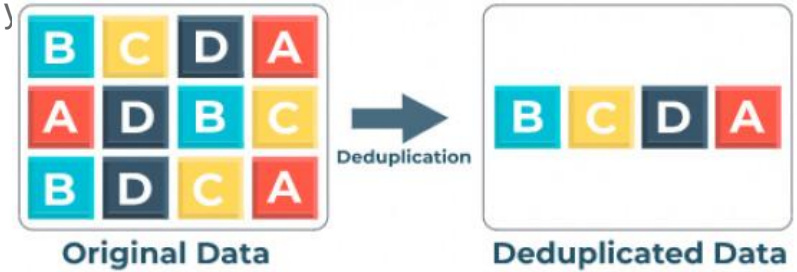
Аугментация текстовых данных: случайная замена синонимов, перестановка слов в предложении, случайное удаление слов, обратный перевод, комбинация методов.



Дедубликация

Проблема: дубликаты искажают информацию о распределении слов в корпусе; переобучение [15].

Дедубликация текстовых данных: выбирается метрика близости документов друг к другу, подбирается порог для метрики, производится



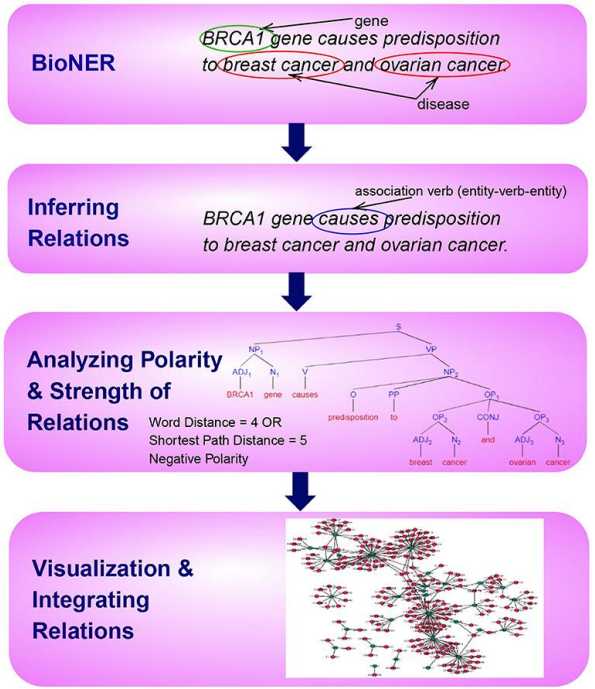
Обзор. Предобработка текста*

Извлечение отношений

Подходы: составление правил вручную, использование справочников; скрытые марковские модели, решающие деревья, метод максимальной энтропии, метод опорных векторов и другие[17].

=> можем сформировать списки синонимов.

+ для аугментации(расширить словарь);
+ для мешка слов(сузить словарь).



Векторизация

Частотное кодирование

TF-IDF

Word2Vec

Мешок N-грамм

NNLM, RNNLM*

Обзор. Векторизация

Bag-of-words & Bag-of-ngramms

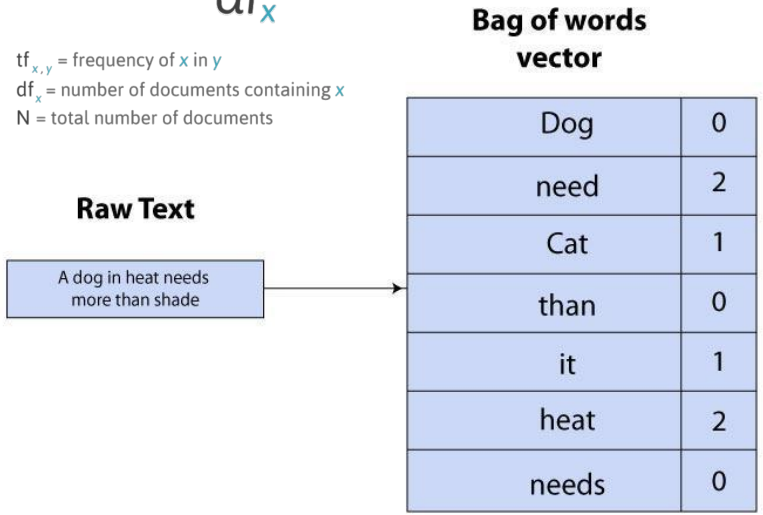
- Частотное кодирование
- One-hot кодирование
- TF-IDF кодирование

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents



Ориентированы на ключевые слова в тексте, а не на текст в целом[19,20,21].

\+ простые и быстрые
\+ часто хорошая точность
\+ промежуточный этап векторизации
\- словарь большой размерности
\- не учитывает контекст
\- и порядок слов* в предложении

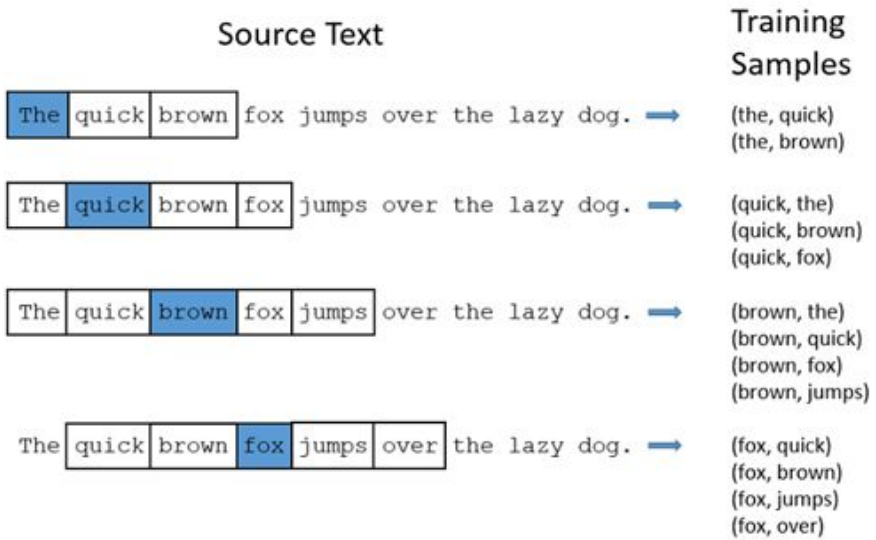
Обзор. Векторизация

Word2Vec: “контекст слова определяется окружением!”. Подход: скользящее окно + нейросети[22, 24, 25].

CBOW - предсказание слова по контексту

Skip-gram - предсказание контекста по центральному слову

- \+ Учитываем контекст
- \+ Учитываем относительный порядок слов
- \- Скорость
- \- Память
- \- Фиксированный словарь



Word2Vec - улучшение NNLM, RNNLM[22].

Обзор. Векторизация

Модели семейства Bag Of Words простые, быстрые, дают интерпретируемый результат, но работают достаточно наивно. Возникают проблемы с точностью.

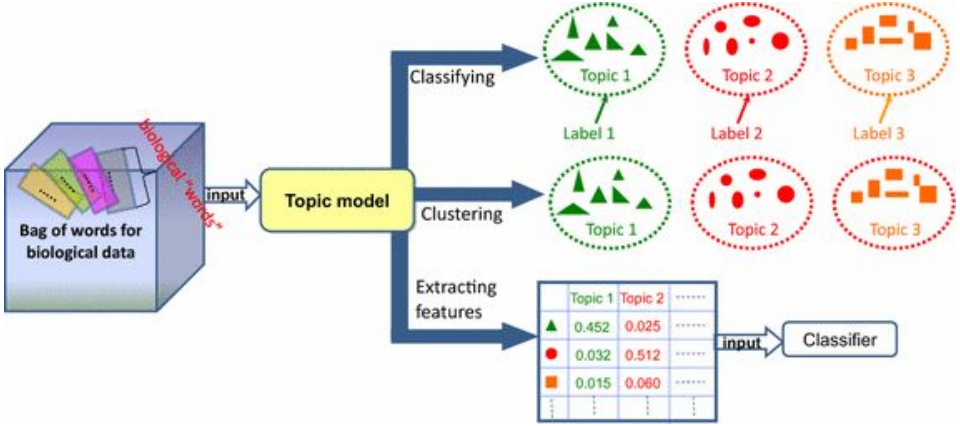
Word2Vec и их модификации - модели, хорошо учитывающие семантику текста, в связи с чем могут давать высокую точность на задачах выявления аномалий. Работают значительно медленнее других рассмотренных алгоритмов.

Обзор. Отбор признаков

LSA - Латентно Семантический Анализ. Вероятностные модели; тематические модели.

Позволяет работать с тематической составляющей документов[23].

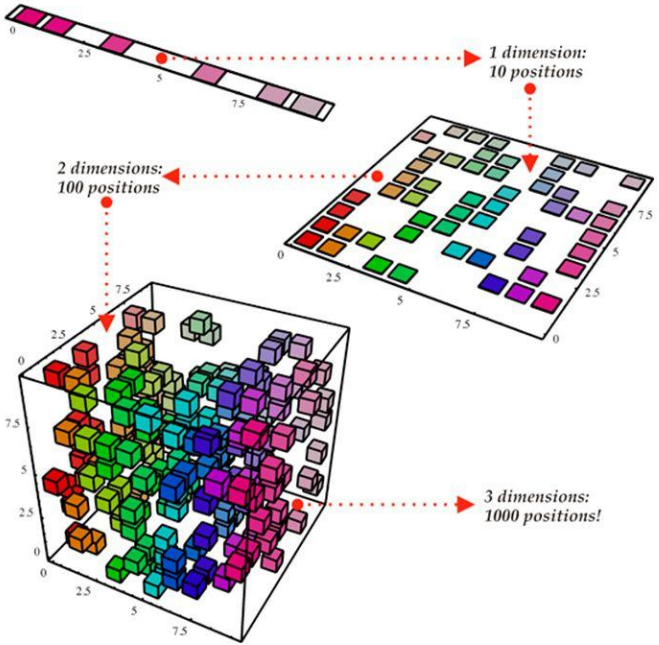
+ можем учитывать семантику текста
+ количество признаков - гиперпараметр
- неинтерпретируемость признаков
- тематики - абстрактные



Обзор. Отбор признаков

Методы снижения размерности:

- LSA
- PLSA - разложение матрицы использует предположения о вероятностной модели.
- GLSA - комбинация LSA и методов поиска информации. Использует N-граммы слов.
- LDA - предположение о вероятностной модели: распределение Дирихле
- SVM - построение разделяющей гиперплоскости



Обзор. Модели

Режимы работы алгоритмов выявления аномалий[1]:

- ***Supervised***. Есть как нормальные, так и *аномальные* помеченные экземпляры.
- **Semi-Supervised**. Есть нормальные помеченные экземпляры
- **Unsupervised**. Не требуют размеченных обучающих данных.

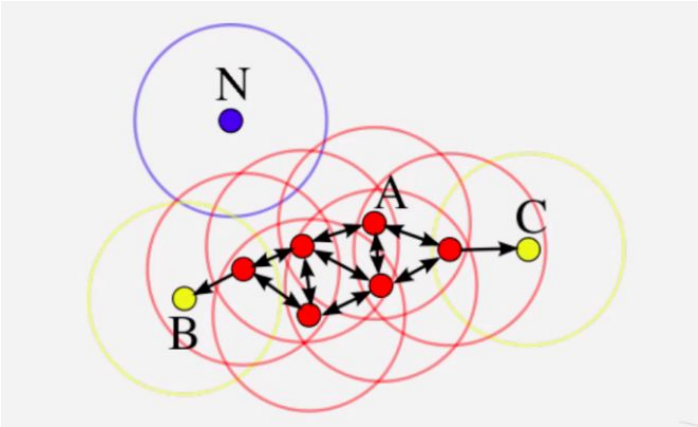
Используются методы, основанные на[1]:

- Кластеризации
- Классификации
- Рассмотрении ближайших соседей
- Статистике
- Теории информации
- Спектральных подходах

Обзор. Модели

Примеры распространенных моделей для решения задач обнаружения аномалий в потоках данных[26]:

- Модели, основанные на кластеризации:
  - CURE, K-means, CLARANS, BIRCH;
  - DBSCAN - плотностной алгоритм
- STORM. Использует скользящее окно потока. Состоит из менеджера потока и менеджера запросов, использует информацию о соседях в окне.
- Abstract-C. Аналогично STORM, но использует оптимизированную информацию о соседях.
- COD, MCOD



Результаты обзора

В процессе обзора была рассмотрена общая задача обнаружения новизны в потоках сложноструктурированных данных, задача была сведена к выявлению аномалий в текстах, были рассмотрены основные методы предобработки и векторизации текстовых данных, описаны их идеи, достоинства и недостатки.

В практической части планируется использовать: **стемминг** и **лемматизацию**(со сравнением точности), **аугментацию** и **дедубликацию**(по отдельности и вместе); из методов векторизации - **TF-IDF**(для сравнения точности - также частотное и one-hot кодирование), **Word2Vec** - также со сравнением результата, и, в конечном счёте, различные комбинации данных методов. Среди методов отбора признаков планируется использовать различные модели **LSA**.

Охвачены все основные этапы решения поставленной задачи.

Результаты обзора

По итогу:

- прочитано 36 статей. Суммарно 336 страниц.
- исследованы особенности реализации 5 методов предобработки
- исследованы особенности реализации 8 методов векторизации

Программный стенд. Архитектура.

Программный стенд. Программные особенности.

- Язык: Python 3
- Библиотеки: sklearn, tensorflow, gensim, pymorphy2
- Написано: 1355 строк кода
- Реализация и знакомство с интерфейсом:
  - CountVectorizer, TF-IDF
  - LDA
  - Word2Vec
- Реализация: knn, k-means
- Реализация: линейная и логистическая регрессия

Результаты

- Исследованы методы предобработки, векторизации, снижения размерности и выявления новизны.
- Реализован программный стенд.
- Подготовлен план проведения экспериментальной оценки исследованных алгоритмов.

Дальнейшие планы.

- Цель:
  - Сравнить отобранные в процессе обзора алгоритмы с использованием разработанного программного стенда.

- План проведения экспериментов:
  - Отбор наборов данных.
  - Определение метрики оценки качества алгоритмов.
  - Подбор параметров алгоритмов, оптимизирующих метрику качества.
  - Экспериментальное сравнение алгоритмов.

Дальнейшие планы.

- Постановка экспериментальной оценки задачи:
  - Проведение экспериментов
    - Подробно изучить все основные алгоритмы обнаружения аномалий.
    - Изучить различные виды данных, возникающих в поставленной задаче.
    - Приступить к реализации различных методов выявления новизны в тексте.
    - Провести сравнительную характеристику различных моделей.
    - Найти комбинацию методов, дающую наилучшую точность на различных наборах данных.
  - Завершение исследования и реализации алгоритмов обнаружения новизны.

Источники

1. Arindam Banerjee, Vipin Kumar. Anomaly Detection: A Survey, 2009
2. Manish Gupta, Jing Gao, Outlier Detection for Temporal Data: A Survey, 2014
3. Elaine R. Faria1, Isabel J. C. R. Gonçalves2, André C. P. L. F. de Carvalho3, João Gama. Novelty detection in data streams.
4. Octavian Rusu; Ionela Halcu; Oana Grigoriu; Giorgian Neculoiu. Converting unstructured and semi-structured data into knowledge, 2013
5. Robert Blumgerg and Shaku Atre.The Problem with Unstructured Data, 2003.
6. Agnar Aamodt, Enric Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, 1994.
7. Barry Schiffman and Kathleen R. McKeown. Context and Learning in Novelty Detection, 2005.
8. Spinosa, E.J.: André Carlos Ponce Leon Ferreira de Carvalho. SVMs for novel class detection in Bioinformatics, 2004.
9. Stephen Marsland, Ulrich Nehmzow and Jonathan Shapiro. Novelty Detection for Robot Neotaxis, 2000.
10. Julie Beth Lovins. Development of a Stemming Algorithm, 1968.
11. Ilia Smirnov. Overview of Stemming Algorithms, 2008
12. Joël Plisson, Nada Lavrac, Dunja Mladenic. A Rule based Approach to Word Lemmatization
13. lKaren Spärck Jones. A statistical interpretation of term specificity and its application in retrieva, 1972.
14. Д.В. Климов. Предобработка текстовых сообщений для метрического классификатора, 2017
15. Jared Dinerstein, Sabra Dinerstein, Parris K. Egbert. Learning-based Fusion for Data Deduplication.
16. Thomas Hofmann. Probabilistic Latent Semantic Analysis
17. Астахова Д.И. Извлечение именованных сущностей с использованием Википедии, 2015
18. Jason Wei, Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, 2019

Источники:

19. Федоренко В.И., Киреев В.С. Использование методов векторизации текстов на естественном языке для повышения качества контентных рекомендаций фильмов, 2018
20. Н.А.Федюшкин, С.А.Федосин. О выборе методов векторизации текстовой информации, 2018.
21. В.В. Попов, Т.В. Штельмах. Естественный текст: математические методы атрибуции, 2019. К.Д. Жук. Особенности использования некоторых алгоритмов для классификации текстов.
22. Tomas Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.
23. А. В. Платонов, И. А. Бессмертный, Ю. А. Королёва. Векторное представление слов при помощи аппарата квантовой теории вероятностей, 2019.
24. Jay Alammar. The illustrated word2vec, 2019.
25. Thushan Ganegedara. Intuitive Guide to Understanding Word2vec, 2018.
26. by JinitaTamboli, Madhu Shukla. A Survey of Outlier Detection Algorithms for Data Streams, 2014.

Спасибо!