УДК 004.6

Разработка программного комплекса для обработки больших объемов текстовой информации

Безверхий Олег Александрович
Амурский государственный университет
магистрант

Самохвалова Светлана Геннадьевна
Амурский государственный университет
к. т. н, доцент кафедры информационных и управляющих систем
факультета математики и информатики

Аннотация
В статье рассмотрены вопросы и проблемы использования различных технологии для автоматизированной обработки текстовых документов, а также их визуализация.
Ключевые слова: слова: методы анализа текста, автоматическая обработка текстов, кластеризация, снижение размерности, python

Development of software for processing big volumes of textual information

Bezverhii Oleg Aleksandrovich
Amur State University
Undergraduate student

Samohvalova Svetlana Gennadievna
Amur State University
Candidate of Technical Sciences

Abstract
The article discusses the issues and problems of using different technologies for the automated processing of text documents, as well as their visualization.
Keywords: text analysis methods, automatic word processing, clustering, dimensional reduction, python

Объектом исследования являются слабоструктурированные текстовые данные.

Предметом исследования являются математические и компьютерные модели кластеризации текстовых сообщений, методы снижения размерности.

Целью работы является исследование задачи кластеризации большого объёма текстовых данных и создание программного комплекса, позволяющего выполнять кластеризацию и представлять данные в графическом виде.

Для выполнения цели работы были поставлены следующие задачи:

1. анализировать методы обработки и кластеризации текстовых сообщений для освоения предметной области, а так же найти пути повышения эффективности обработки и кластеризации текстов;

2. построить математическую модель, которая позволит исследовать закономерности между тематиками (кластерами) и динамику (частоту) поисковых запросов;

3. разработать алгоритм обработки текстовых запросов, позволяющий производить оперативную кластеризацию текстовых документов;

4. анализировать алгоритмы, предназначенные для снижения размерности;

5. разработать программный комплекс, выполняющий предложенные алгоритмы.

При решении поставленных задач использовались методы системного анализа, математического и компьютерного моделирования, обработки естественного языка, теории вероятностей, искусственного интеллекта, нейронных сетей, проектирования информационных систем и языки программирования.

Научная новизна. В данной работе были объединены несколько этапов – от предварительной обработки до графического представления. Так же был опробован метод из смежной области – вероятностное тематическое моделирование. Результаты работы показывают основные направления для дальнейшего развития в области кластерного анализа текстовых документов.

Практическая и теоретическая значимость исследований.

Результаты работы могут найти применение в информационном и его разновидности – разведочном поиске, в анализе коллекций научных статей, новостных потоков. Применение тематического моделирования позволит решать такие задачи, как определение тематики авторов, различных изданий (журналов, материалов конференций). Использовать разработанный программный комплекс можно в составе рекомендательных систем, например, для определения интересов пользователей социальных сетей на основе их постов.

Количество источников и само количество информации в наше время лавинообразно возрастают. Это и сообщения из социальных сетей, теле и аудио информация, показания различных приборов и так далее, которые каждую минуту генерируют огромный поток данных [11]. Поэтому всё чаще возникает необходимость в том, чтобы оперативно структурировать приходящую информацию по определенным группам.

Большая часть информации, сохраняемой людьми в мире, существует в виде текста. Данные на естественном языке составляют особую разновидность неструктурированных данных; обработка таких данных достаточно сложна, потому что требуют знания, как лингвистики, так и методов машинного обучения.

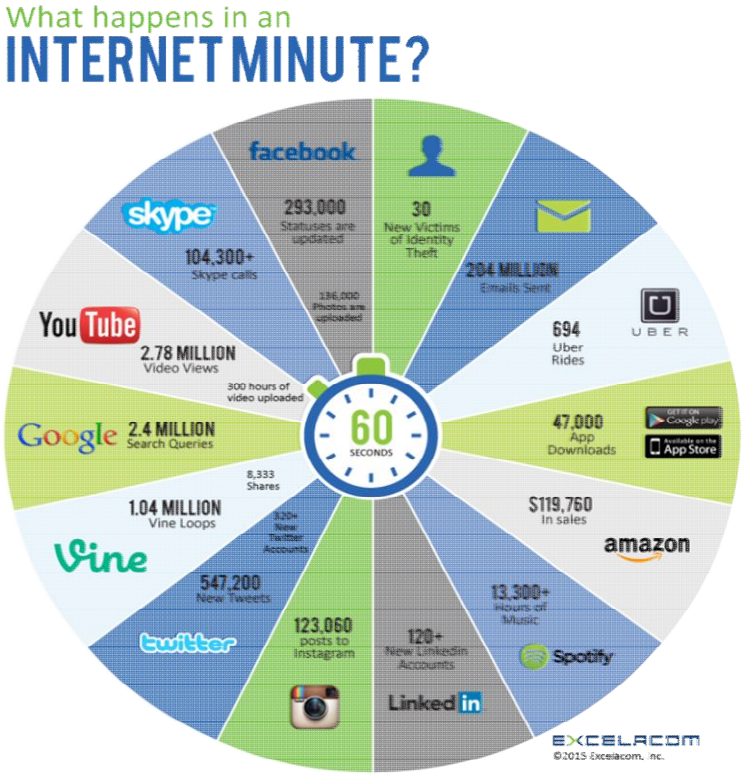

Рисунок 1 – Количество генерируемой информации за минуту в различных соцсетях

Статистические закономерности функционирования языка и текста являются предметом лингвистики. Наряду с единичными текстами объектом лингвистики становятся коллекции текстов и информационные потоки.

Существует два способа анализа текстовых данных: кластеризация и классификация.

Кластеризация является единственным решением задачи, когда нет точного представления о составе и структуры данных, а ручной отбор сложен либо не соответствует временным и человеческим ресурсам. Классификация занимает больше времени – необходимо собрать и создать большую обучающую выборку. Поэтому было решено решать задачу кластеризации.

Кластеризация - это процесс выделения подгрупп объектов с близкими свойствами. Система должна найти самостоятельно атрибуты и сгруппировать объекты по группам. Группы формируются только на основе попарной схожести описаний документов, причем характеристики этих групп заранее не известны, в отличие от классификации документов [3].

Кластеризация применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, упрощения процесса просмотра при поиске необходимой информации, нахождения уникальных документов из коллекции, выявления дубликатов или очень близких по содержанию документов.

В качестве единицы анализа текста в работах используются стандартные единицы, как лексема и словоформа [4]. Когда и какая из этих

единиц важнее – решает исследователь, и выбор задается целью и задачами работы.

Процесс анализа текстов необходимо начать с предварительной обработки. Она включает в себя следующие этапы:

- Фильтрация – удаление спецсимволов и пунктуации;
- Токенизация – разбивание текста на термины – слова или словосочетания;
- Стемминг;
- Удаление стоп-слов;
- Сокращение – удаление низкочастотных слов (является необязательным параметром);
- Создание взвешенной матрицы терм-документ – переход к векторной форме документа. В работе было применено преобразование TD-IDF.

В общем алгоритм обработки данных можно представить так:

Рисунок 2 – Алгоритм обработки данных

Нормализация – приведение слова к начальной форме. Пример: Кошек – кошка, бежал – бежать и т.д. Русский язык характерен морфологической сложностью, свободным порядком слов. Поэтому многие разработчики используют либо готовые словари, либо вручную прописывают правила для каждой словоформы. На данный момент не существует точного алгоритма для правильного и однозначного выбора леммы.

Судя по научным конференциям в сфере компьютерной лингвистики, наблюдается тенденция к разделению процесса нормализации на стемминг и синтез леммы.

Стемминг–определение стеммы (основа слова). Русский язык относится к группе флективных синтетических языков – языков, в которых преобладает словообразование с использованием аффиксов, сочетающих сразу несколько грамматических значений [6]. Поэтому для языка допускается использование алгоритмов стемминга. Были опробованы стеммер Портера, PyMorphy, а также инструмент от Яндекс – Mystem.

Для работы с текстовыми документами обычно используют векторное представление (Vector Space Model), т.е. отображение текста в вектор. Данную процедуру можно осуществить несколькими способами. Наиболее популярными являются мешок слов (Bag of Words) и учет взаимного положения слов.

В работе использовалась модель мешка слов, под которыми понимаются n-граммы, т.е. словосочетания длины не более 3, т.к. данное представление хорошо зарекомендовало себя в задачах автоматической классификации. Для подсчета значимости слова в тексте (веса) применяется модель tf-idf.

Мера TF-IDF [1] является произведением двух характеристик: TF (term frequency — частота слова) и IDF (inverse document frequency — обратная частота документа)

$$tfidf(t,d,D) = tf(t,d) * idf(t,D)$$

Над полученной матрицей применялись следующие алгоритмы кластеризации: K-means и его модификацию Mini Batch K-Means, DBSCAN, аггломеративная кластеризация с различными метриками [10].

Для получения графического результата необходимо перевести многомерные вектора в двухмерное пространство. Для этого используются алгоритмы снижения размерности.

Существует множество алгоритмов, например, MDS, SVD, PCA и каждый из них даёт различные результаты. К сожалению, не всегда возможно их применение.

В работе был использован алгоритм Incremental PCA (IPCA) [8]. Алгоритм используется в качестве замены метода главных компонентов (PCA), когда набор данных, подлежащий разложению, слишком велик, чтобы разместиться в оперативной памяти. IPCA создает низкоуровневое приближение для входных данных, используя объем памяти, который не зависит от количества входных выборок данных. Он по-прежнему зависит от функций входных данных, но изменение размера пакета позволяет контролировать использование памяти.

На рисунке 3 показаны необработанные данные, кластеризованные методом K-Means и сжатые методом IPCA. Количество кластеров равняется 25.

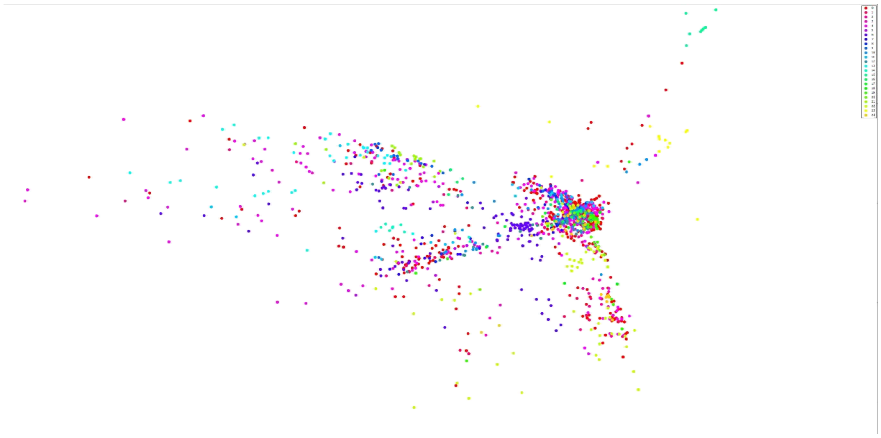

Рисунок 3 – K-Means на «сырых» данных

На рисунке 4 показаны предварительно обработанные данные, кластеризованные методом K-Means и сжатые методом IPCA. Количество кластеров равняется 25.

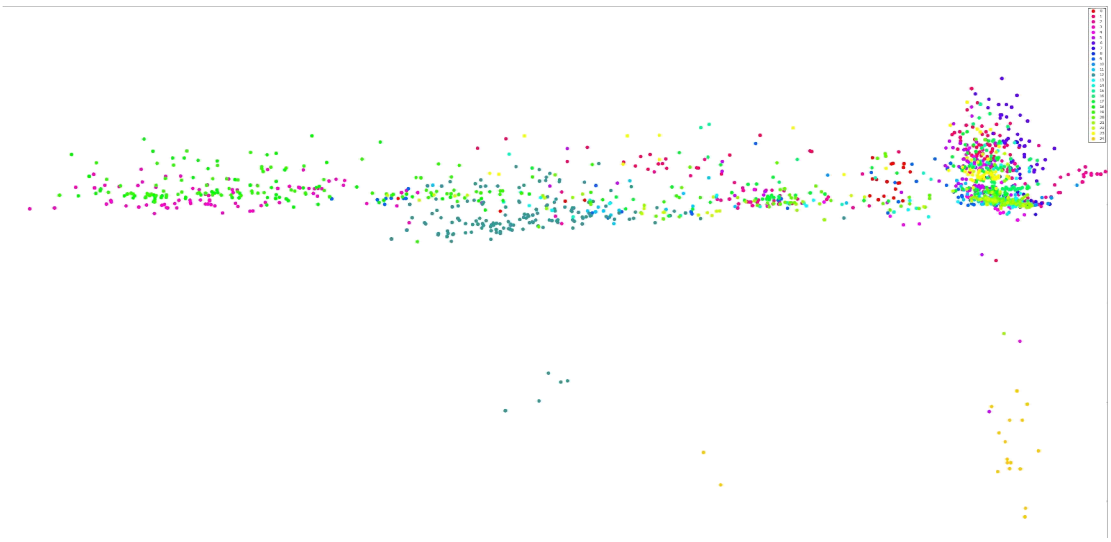

Рисунок 4 – K-Means на «сырых» данных

Для визуализации уже выявленных кластеров можно использовать дендограмму. Дендрограмма позволяет представить кластерную структуру в виде плоского графика независимо от того, какова размерность исходного пространства. Пример такой дендограммы представлен на рисунке 5.

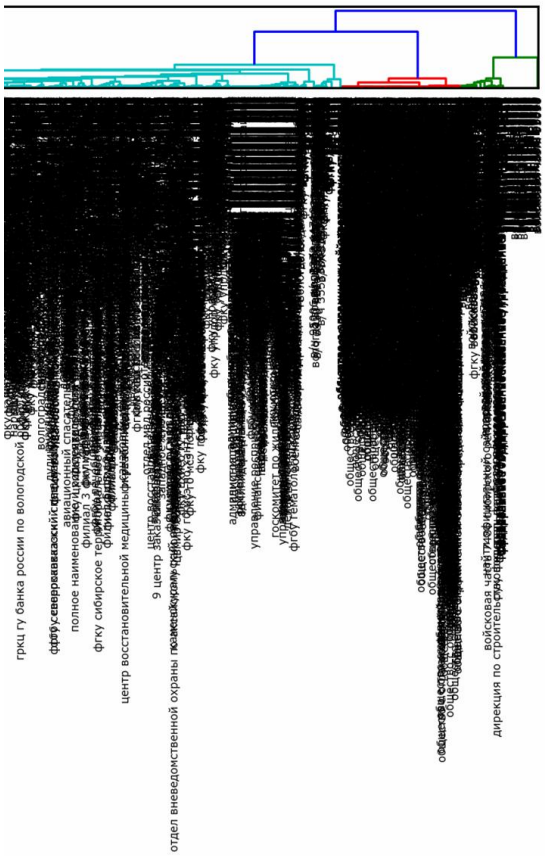

Рисунок 5 – Дендограмма исследуемых данных

Представленные ранее алгоритмы позволяют выполнять грубую кластеризацию, то есть один объект может относиться только к одному классу. В реальных текстах это утверждение не всегда верно, потому конкретный текст может относиться к нескольким классам.

Для решения данной проблемы были придуманы алгоритмы тематического моделирования.

Вероятностное тематическое моделирование – это набор алгоритмов, позволяющих анализировать слова в больших наборах документов и извлекать из них темы, связи между темами и изменение их во времени [7].

Существует несколько моделей и для исследования была выбрана модель латентного размещения Дирихле. Модель латентного (скрытого) размещения Дирихле (Latent Dirichlet Allocation, LDA) — порождающая модель, в которой каждый документ рассматривается как смесь различных тем. Эта модель схожа с PLSA, но отличается тем, что в LDA распределение тем следует распределению Дирихле. Это позволяет оценивать вероятности документов и терминов вне текстовой коллекции.

Результат работы алгоритма с 5 темами и модуля LDAvis для графического представления тематических моделей, представлен на рисунке 6.

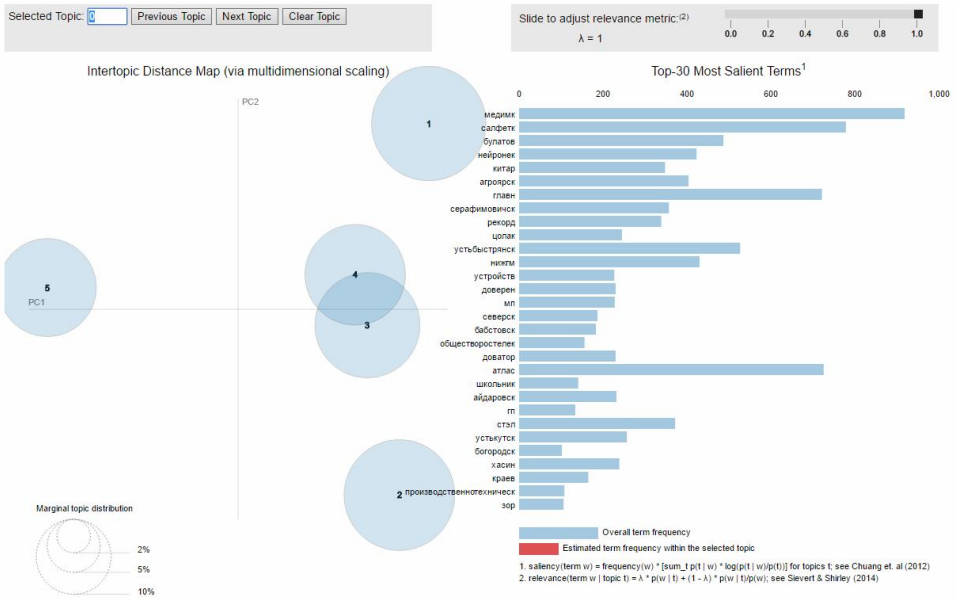

Рисунок 6 – Модель LDA с 5 темами

Проблемы, выявленные при проведении исследования:

Наличие в тексте омонимов. С точки зрения анализа слово будет нести только один смысл, хотя значение слова в различных текстовых данных может быть различным. Фразы клише и фразеологизмы тоже усложняют анализ текстов.

Недостаточность знаний об объекте. Существует трудноформализуемые области, в которых создание модели объекта затруднительно. В таких случаях тяжело применять алгоритмы,

основывающиеся на представлении класса, как набора распределённых в пространстве переменных [9].

Неустойчивость результатов кластеризации. Конечные результаты могут сильно различаться в зависимости от выбранных начальных условий, параметров работы алгоритмов.

Форма кластеров. Чаще всего алгоритмы завязаны под конкретную форму кластеров.

Библиографический список

1. TF-IDF | Virtual Laboratory Wiki | Fandom powered by Wikia [Электронный ресурс]. – Режим доступа: http://ru.vlab.wikia.com/wiki/TF-IDF– 27.05.2017.
2. Кластеризация текстовых документов из электронной базы публикаций алгоритмом FRiS-Tax [Электронный ресурс]. – Режим доступа: www.ict.nsc.ru/jct/getfile.php?id=1580 – 05.04.2017
3. Барахнин В.Б., Ткачев Д.А. Кластеризация текстовых документов на основе составных ключевых термов // Вестник НГУ. Информационные технологии. 2012. Т. 10. № 4. С. 95-103.
4. Методы анализа текста: методологические основания и программная реализация. [Электронный ресурс]. – Режим доступа: http://cyberleninka.ru/article/n/metody-analiza-teksta-metodologicheskie-osnovaniya-i-programmnaya-realizatsiya.pdf – 05.04.2017
5. Столяренко А.В., Анализ методов кластеризации текстов применительно к работе с корпусом научных статей [Электронный ресурс]. – Режим доступа: http://sntbul.bmstu.ru/doc/836973.html – 05.04.2017.
6. Стемминг [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Стемминг – 05.04.2017.
7. Вероятностное тематическое моделирование - MachineLearning.ru [Электронный ресурс]. – Режим доступа: www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf – 05.04.2017.
8. Incremental PCA [Электронный ресурс]. – Режим доступа: http://scikit-learn.org/stable/auto_examples/decomposition/plot_incremental_pca.html – 05.04.2017.
9. Актуальные проблемы кластерного анализа [Электронный ресурс]. – Режим доступа: http://davaiknam.ru/text/aktualenie-problemi-klasternogo-analiza – 05.04.2017.
10. Безверхий О.А., Самохвалова С.Г. Кластеризация большого объема текстовых поисковых запросов// Ученые заметки ТОГУ. 2016. Т.7. № 3-1. С.104–110
11. Безверхий О.А., Самохвалова С.Г. Понятие больших данных и их использование в различных сферах // Содружество. 2016. № 4 (4). С.89-93