АВТОМАТИЗИРОВАННЫЙ СБОР И СТРУКТУРИРОВАНИЕ ДАННЫХ ПЕРЕХВАТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ КОНЦЕПЦИИ МАШИННОГО ОБУЧЕНИЯ

АВТОР РАБОТЫ: АБРИКОСОВ ЕВГЕНИЙ ПАВЛОВИЧ

РУКОВОДИТЕЛЬ: КАНАШ СЕРГЕЙ ЮРЬЕВИЧ

АКТУАЛЬНОСТЬ И ЗНАЧИМОСТЬ РАБОТЫ

→ **Проблема и актуальность**

В настоящее время массивы информации, доступные человеку, многократно выросли благодаря развитию сети Интернет. Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. В огромных информационных объемах имеет смысл говорить только об автоматической рубрикации.

→ **Значимость проекта**

В ходе выполнения работы был создан программный комплекс – автоматический классификатор данных с применением алгоритмов анализа естественного языка, применимый для извлечения структурированной информации из текстов. Данный программный комплекс позволяет автоматически обрабатывать поступающие материалы по выбранным тематикам.

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ

Цель проекта

В данной работе поставлена задача разработки программного комплекса, позволяющего автоматизировать сбор и структурирование информации на естественном языке с тематических интернет-ресурсов – классификатора данных.

Задачи проекта

→Проанализировать современное состояние исследований в области агрегации данных и анализа естественных языков.

→Провести анализ существующих научных и практических решений в выбранной области, изучить методы, принципы и технологии извлечения именованных сущностей, возможности их применения для агрегации данных.

→Спроектировать компонентную реализацию программной системы для классификации тематических данных.

→Разработать программный комплекс.

→Сделать вывод об эффективности созданной системы классификации данных.

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Теоретические положения автоматической классификации данных

→ Классификация или рубрикация информации: отнесение порции информации к одной или нескольким категориям из ограниченного множества, является традиционной задачей организации знаний и обмена информацией.

→ При применении методов машинного обучения для построения классификатора используется набор документов, предварительно отобранная человеком. Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества текстов.

→ Машинное обучение – это научное исследование алгоритмов и статистических моделей, которые компьютерные системы используют для эффективного выполнения конкретной задачи без использования явных инструкций, опираясь на шаблоны и выводы.

РАЗРАБОТКА

МОДУЛЬНОСТЬ СИСТЕМЫ

Разработанная система состоит из четырех модулей:

→ Модуль работы с источниками данных

→ Модуль предварительной обработки текста

→ Модуль оценки подготовленного текста

→ Модуль обработки действий пользователя

РАЗРАБОТКА

РЕАЛИЗАЦИЯ МОДЕЛИ КЛАССИФИКАЦИИ

- → Импорт библиотек
- → Импорт набора данных
- → Предварительная обработка текста
- → Преобразование слов текста в коэффициенты
- → Обучающие и тестовые наборы
- → Обучение модели классификации текста и прогноз
- → Оценка модели
- → Сохранение и загрузка модели

РАЗРАБОТКА

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ТЕКСТА

European losses hit GM's profitsGeneral Motors (GM) saw its net profits fall 37% in the last quarter of 2004, as it continued to be hit by losses at its European operations.The US giant earned $630m (Â£481.5m) in the October-to-December period, down from $1bn in the fourth quarter of 2003.

European losses hit gm profitsgeneral motors gm saw its net profits fall 37 in the last quarter of 2004 as it continued to be hit by losses at its european operations the us giant earned 630m 481 5m in the october to december period down from 1bn in the fourth quarter of 2003

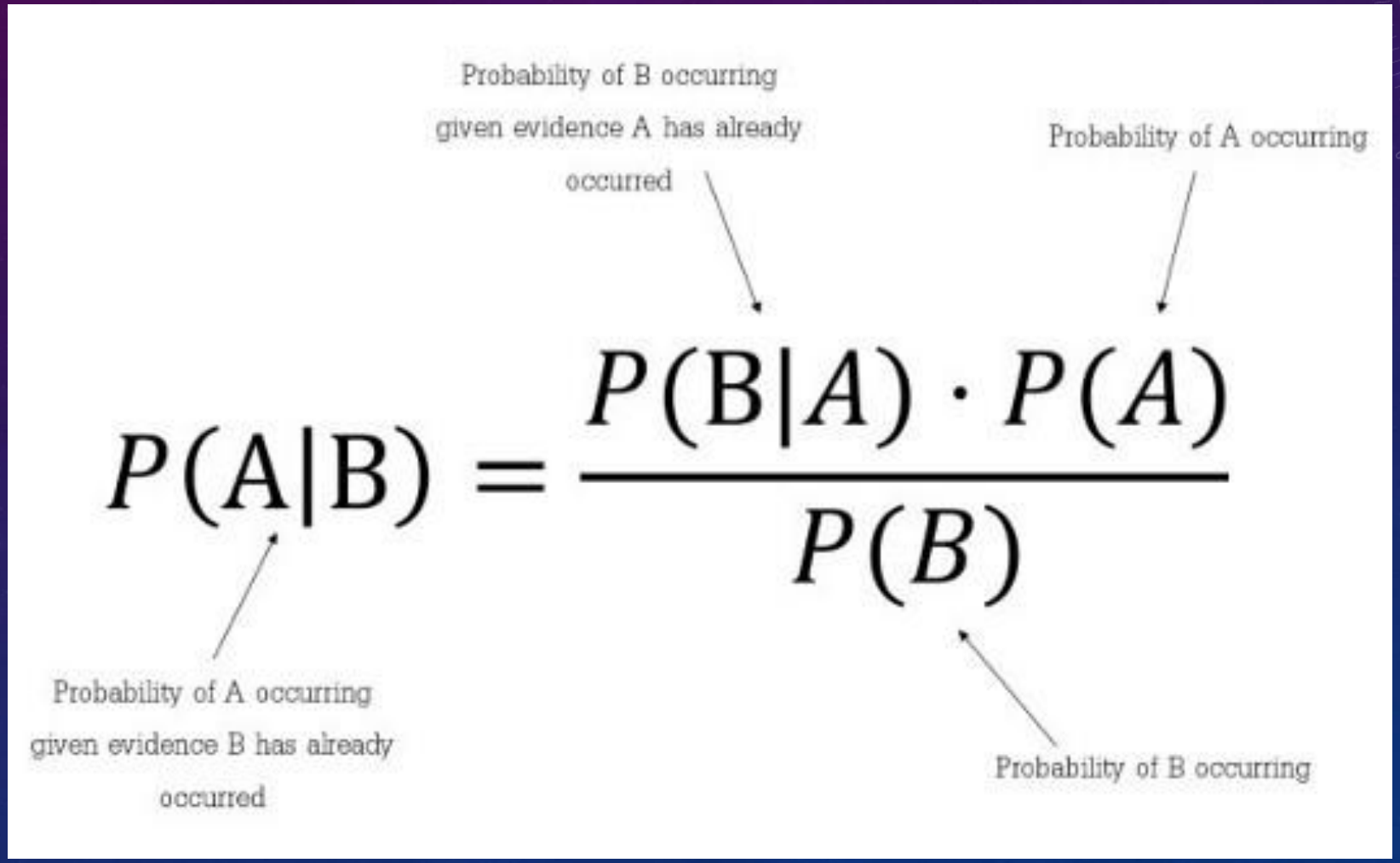
Удаление спецсимволов, цифр, одиночных символов

```
business [('said', 1618), ('us', 753), ('year', 571), ('mr', 549), ('would', 463), ('also', 439), ('ne
entertainment [('said', 789), ('film', 698), ('best', 582), ('music', 413), ('also', 398), ('us', 348)
politics [('said', 2138), ('mr', 1505), ('would', 1051), ('government', 635), ('labour', 587), ('peopl
sport [('said', 926), ('game', 456), ('first', 436), ('would', 396), ('win', 392), ('last', 371), ('wo
tech [('said', 1540), ('people', 922), ('also', 533), ('new', 511), ('technology', 490), ('mr', 485),
```

Частота появления слов, распределенная по темам

РАЗРАБОТКА

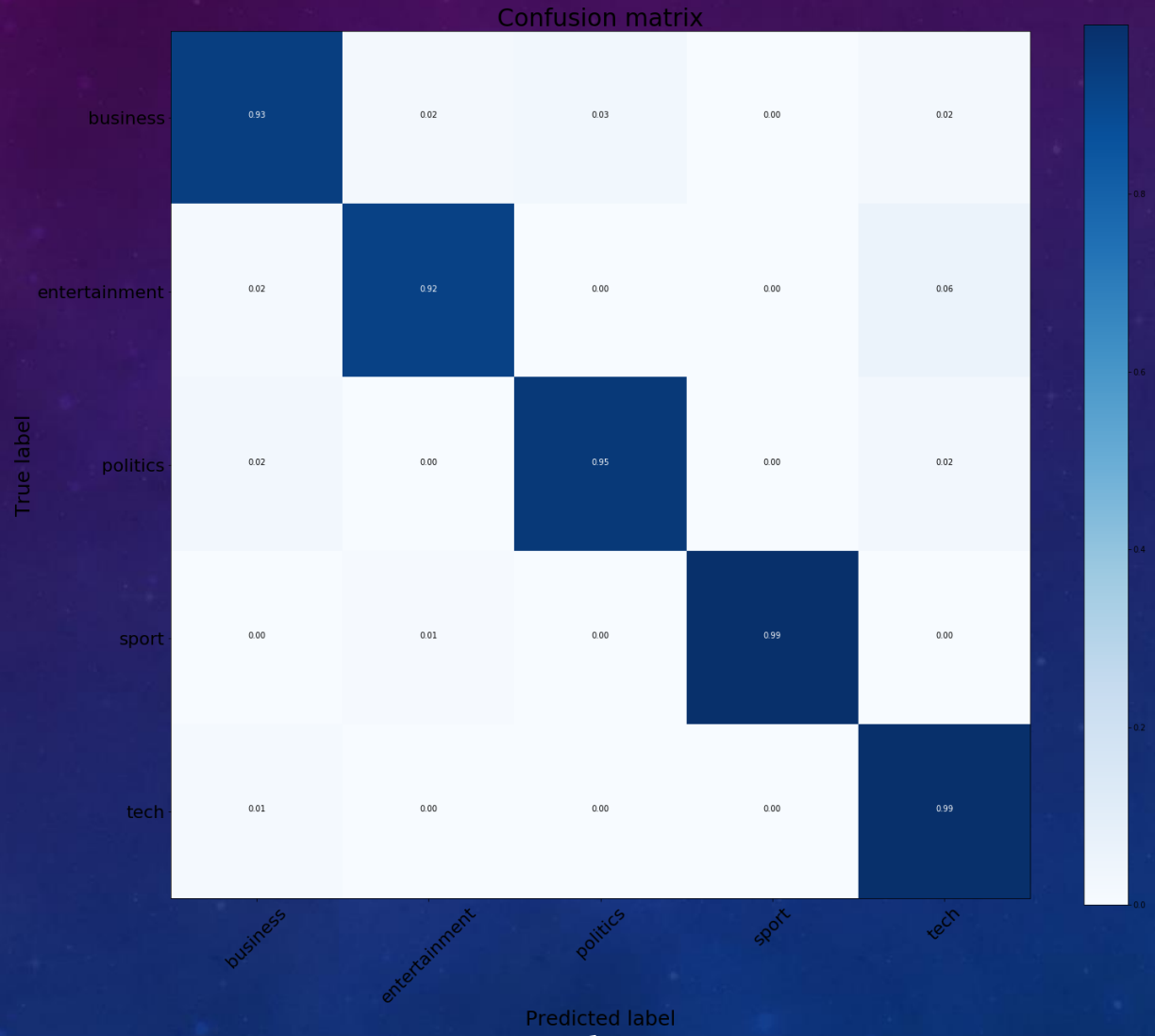
НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР



Частота появления слов, распределенная по темам

РАЗРАБОТКА

ПРОВЕРКА РАБОТЫ АЛГОРИТМА



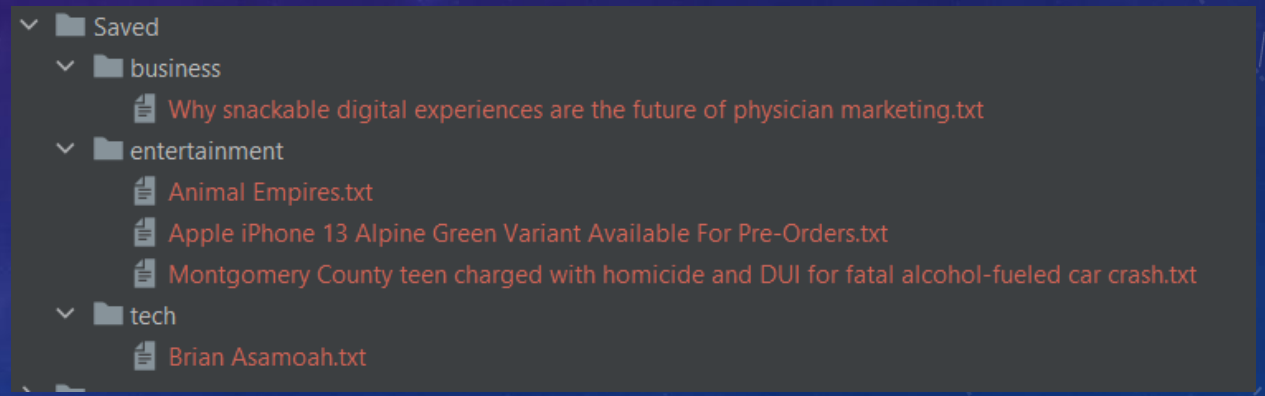
Проверка правильности работы программного комплекса

РАЗРАБОТКА

ПОЛУЧЕННЫЙ РЕЗУЛЬТАТ

```
*******

Краткая аннотация текста:          Apple iPhone 13 Alpine Green Variant Available For Pre-Orders: Best Discounts You Can Avail
Предположительная тема текста:     ENTERTAINMENT

*******

Краткая аннотация текста:          Animal Empires
Предположительная тема текста:     ENTERTAINMENT

*******

Краткая аннотация текста:          Why snackable digital experiences are the future of physician marketing
Предположительная тема текста:     BUSINESS

*******

Краткая аннотация текста:          Brian Asamoah: NFL Draft Prospect Interview
Предположительная тема текста:     TECH

*******
```

Результат категоризации полученной информации



Распределение полученных новостей по категориям в файловой системе компьютера

ВЫВОДЫ

→ В результате проделанной работы было разработано программное обеспечение, базирующееся на байесовском алгоритме. ПО позволяет определять тематику текста на основе данных, полученных во время обучения классификатора.

→ По результатам выполнения НИРС программное обеспечение позволяет производить классификацию текстов на естественном языке по пяти темам. Заложены возможности по расширению библиотек, используемых тем с целью обеспечения охвата более широкого спектра проблем.

→ Реализованный программный комплекс планируется использовать для решения задачи автоматического извлечения тем документов и структурирования данных из файлов на естественном языке. Программное обеспечение позволит обеспечить оптимальную организацию процесса сбора информации и уменьшит временные затраты на поиск информации представляющий интерес.

СПАСИБО ЗА ВНИМАНИЕ!