

**UMCS**

UNIWERSYTET  
SKŁODOWSKIEJ

MARII

CURIE-

W LUBLINIE

Wydział Matematyki, Fizyki i Informatyki

Kierunek: WPISAĆ

Specjalność: WPISAĆ

**Maksim Ryshko**

nr albumu: NR 310030

**Wykorzystanie metod klastrujących podczas analizy  
danych tekstowych**

**Using Clustering Methods in Text Data Analysis**

Praca licencjacka

napisana w Zakładzie JAKIMŚ

pod kierunkiem dr. CZYIMŚ

**Lublin rok WPISAĆ**



# Оглавление

|   |          |
|---|----------|
| <b>Wstęp</b>  | <b>1</b> |
| 0.1. Причина появления проекта . . . . .            | 1        |
| 0.2. Идея проекта . . . . .                         | 1        |
| 0.3. Для кого это . . . . .                         | 1        |
| <b>1. Rozdział 1</b>                                | <b>3</b> |
| 1.1. Klasteryzacja . . . . .                        | 3        |
| 1.1.1. Defenicyja . . . . .                         | 3        |
| 1.1.2. Cele klasteryzacji . . . . .                 | 3        |
| 1.1.3. Metody klasteryzacji . . . . .               | 3        |
| 1.1.4. Etapy klasteryzacji . . . . .                | 4        |
| 1.1.5. Zastosowanie . . . . .                       | 4        |
| 1.1.6. Оценка эффективности кластеризации . . . . . | 4        |
| 1.2. Przetwarzanie języka naturalnego . . . . .     | 5        |
| 1.2.1. Defenicyja . . . . .                         | 5        |
| 1.2.2. Przedprzetwarzanie tekstu . . . . .          | 5        |
| 1.2.3. Stemming . . . . .                           | 5        |
| 1.2.4. Lematyzacja . . . . .                        | 6        |
| 1.3. Векторизация . . . . .                         | 6        |
| 1.3.1. Opis . . . . .                               | 6        |
| 1.3.2. Celi wektoryzacji . . . . .                  | 6        |
| 1.3.3. Metody wektoryzacji . . . . .                | 6        |
| 1.3.4. Применение . . . . .                         | 7        |
| 1.4. снижения размерности . . . . .                 | 7        |
| 1.4.1. Дефениция . . . . .                          | 7        |
| 1.4.2. Методы снижения размерности . . . . .        | 7        |
| 1.4.3. Применение . . . . .                         | 8        |
| <b>2. Rozdział 2</b>                                | <b>9</b> |
| 2.1. Приложение . . . . .                           | 9        |
| 2.1.1. Opis . . . . .                               | 9        |
| 2.1.2. Технологии . . . . .                         | 9        |
| 2.1.3. GUI . . . . .                                | 9        |
| 2.1.4. OOP . . . . .                                | 10       |
| 2.2. Текстовые данные . . . . .                     | 10       |
| 2.2.1. Opis . . . . .                               | 10       |
| 2.2.2. API запросы . . . . .                        | 10       |
| 2.2.3. Обработка полученных данных . . . . .        | 11       |
| 2.2.4. Хранение . . . . .                           | 11       |
| 2.3. Обработка текста . . . . .                     | 11       |
| 2.3.1. Opis . . . . .                               | 11       |
| 2.3.2. Обязательная обработка . . . . .             | 12       |
| 2.3.3. Примитивная обработка . . . . .              | 12       |
| 2.3.4. Удаление stop-words . . . . .                | 12       |

|           |  |           |
|-----------|--|-----------|
| 2.3.5.    | Лематизация . . . . .                      | 12        |
| 2.3.6.    | Анализ . . . . .                           | 13        |
| 2.3.7.    | Итог . . . . .                             | 13        |
| 2.4.      | Wektoryzacja . . . . .                     | 14        |
| 2.4.1.    | Opis . . . . .                             | 14        |
| 2.4.2.    | TF-IDF . . . . .                           | 14        |
| 2.4.3.    | Word2Vec . . . . .                         | 15        |
| 2.4.4.    | Ocena . . . . .                            | 16        |
| 2.5.      | Кластеризация . . . . .                    | 17        |
| 2.5.1.    | Opis . . . . .                             | 17        |
| 2.5.2.    | K-means . . . . .                          | 17        |
| 2.5.3.    | DBSCAN . . . . .                           | 18        |
| 2.5.4.    | Выбор параметров . . . . .                 | 19        |
| 2.6.      | Визуализация . . . . .                     | 19        |
| 2.6.1.    | Opis . . . . .                             | 19        |
| 2.6.2.    | Снижение размерности . . . . .             | 19        |
| 2.6.3.    | T-SNE . . . . .                            | 19        |
| <b>3.</b> | <b>Rozdział 3</b>                          | <b>21</b> |
| 3.1.      | Описание приложения . . . . .              | 21        |
| 3.1.1.    | Рабочее окно . . . . .                     | 21        |
| 3.1.2.    | Меню . . . . .                             | 21        |
| 3.1.3.    | Workspace . . . . .                        | 22        |
| 3.1.4.    | Text . . . . .                             | 22        |
| 3.2.      | Пример использования . . . . .             | 23        |
| 3.2.1.    | Получение текстовых данных . . . . .       | 23        |
| 3.2.2.    | Создание Workspace . . . . .               | 23        |
| 3.2.3.    | Выбор методов и параметров . . . . .       | 23        |
| 3.2.4.    | Текстовая обработка . . . . .              | 23        |
| 3.2.5.    | Векторизация . . . . .                     | 24        |
| 3.2.6.    | Кластеризация . . . . .                    | 24        |
| 3.2.7.    | Интерпритация результатов = Итог . . . . . | 24        |
| <b>4.</b> | <b>Tytuł 4</b>                             | <b>25</b> |
| 4.1.      | Sekcja . . . . .                           | 25        |
|           | <b>Podsumowanie</b>                        | <b>27</b> |





# Wstep

## 0.1. Причина появления проекта

В современном мире количество текстовых данных растет с невероятной скоростью. Текстовые данные становятся неотъемлемой частью многих областей, включая социальные сети, форумы, блоги и другие платформы, где пользователи активно делятся своими мнениями и информацией. Одной из наиболее популярных платформ является YouTube, на которой миллионы пользователей ежедневно оставляют комментарии под видео. Анализ этих текстовых данных может предоставить ценную информацию для маркетинговых исследований, социальных исследований, а также для улучшения взаимодействия с аудиторией.

## 0.2. Идея проекта

Данный проект представляет собой GUI-приложение, предназначенное для анализа текстовых данных. В роли текстовых данных может выступать любые текстовые документы. Проект ориентирован на обработку комментариев с YouTube, что позволяет пользователям легко и быстро получать и анализировать эти данные. После выбора определенных методов обработки комментариев, которые пользователь может настраивать, пользователь может оценить результат работы в качестве итогового 2D графика на котором будут визуальны разделены текстовые данные на группы(кластеры). Для помощи в качественном анализе данных пользователю будут доступны для настройки методы NLP, векторизации и кластеризации.

## 0.3. Для кого это

Таким образом, проект предоставляет мощный инструмент для качественного анализа текстовых данных, который может быть полезен как исследователям, так и специалистам в области маркетинга и социальных медиа.





# Глава 1

## Rozdział 1

### 1.1. Klasterzyacja

#### 1.1.1. Defenicja

Klasteryzacja (ang. clustering) to wielowymiarowa procedura analizy statystycznej, mająca na celu zbieranie danych zawierających informacje (cechy) o próbie obiektów, a następnie porządkowanie tych obiektów w stosunkowo jednorodne grupy. Zakres zastosowania klasteryzacji jest niezwykle szeroki: używa się jej w archeologii, medycynie, psychologii, chemii, biologii, administracji państwowej, filologii, antropologii, marketingu, socjologii, geologii i innych dyscyplinach. [4]

#### 1.1.2. Cele klasteryzacji

Klasteryzacja jest najczęściej wykorzystywana do następujących celów:

1. Zrozumienie danych poprzez identyfikację struktury klastrów. Podział próby na grupy podobnych obiektów pozwala uprościć dalsze przetwarzanie danych i podejmowanie decyzji, stosując dla każdego klastra odpowiednią metodę analizy.
2. Kompresja danych. Jeśli pierwotna próba jest zbyt duża, można ją skompresować, zachowując tylko jednego najbardziej typowego przedstawiciela z każdego klastra. Zastępując wszystkie obiekty w każdym klastrze jednym najbardziej typowym przedstawicielem, można znacznie zmniejszyć objętość danych, zachowując jednocześnie główne cechy i strukturę pierwotnych danych.
3. Wykrywanie nowości (odchyłeń lub anomalii). Klasteryzacja pozwala na wyróżnienie nietypowych obiektów, które nie mogą być przypisane do żadnego z klastrów. Te obiekty mogą być interesujące jako potencjalne anomalie lub odchylenia, wymagające dodatkowych badań, analiz lub uwagi.

#### 1.1.3. Metody klasteryzacji

Nie ma powszechnie przyjętej klasyfikacji metod klasteryzacji, jednak można wyróżnić kilka grup podejść:

1. Algorytmy klasteryzacji oparte na grafach. Klasa ta obejmuje prymitywne algorytmy oparte na budowaniu grafu podobieństwa między obiektami. Obecnie są one praktycznie nie stosowane w praktyce.
2. Algorytmy klasteryzacji probabilistycznej. Algorytmy te przypisują każdemu obiektowi z próby treningowej prawdopodobieństwo przynależności do każdego z klastrów.
3. Hierarchiczne algorytmy klasteryzacji. Algorytmy te porządkują dane, tworząc hierarchię zagnieżdżonych klastrów.

#### 4. Algorytmy oparte na gęstości danych:

- K-średnich (K-means). Iteracyjny algorytm oparty na minimalizacji sumy kwadratowych odchyleń punktów klastrów od ich centrów.
- Rozpowszechnianie podobieństwa (Affinity Propagation). Algorytm, który rozprzestrzenia wiadomości o podobieństwie między parami obiektów w celu wyboru typowych przedstawicieli każdego klastra.
- Przesunięcie średniej (mean shift). Metoda wybierająca centroidy klastrów w obszarach o największej gęstości.
- Klasteryzacja spektralna (spectral clustering). Metoda wykorzystująca wartości własne macierzy odległości do zmniejszenia wymiarowości przed zastosowaniem innych metod klasteryzacji.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Algorytm grupujący punkty w jeden klastery w obszarach o wysokiej gęstości i oznaczający odosobnione punkty jako szum.

#### 1.1.4. Etapy klasteryzacji

Niezależnie od przedmiotu badań, zastosowanie analizy klastrów obejmuje następujące etapy:

1. Wybór próby do klasteryzacji. Zakłada się, że klasyfikować można tylko dane ilościowe, ponieważ wskaźniki liczbowe mogą być ilościowo oceniane.
2. Określenie zestawu zmiennych, na podstawie których obiekty w próbie będą oceniane, tj. przestrzeni cech. Cechy mogą być różnymi charakterystykami obiektów, które mają znaczenie dla badań.
3. Obliczanie wartości wybranej miary podobieństwa (lub różnicy) między obiektami. Celem jest określenie, jak blisko lub różnią się obiekty w próbie pod względem wybranych cech.
4. Zastosowanie metody analizy klastrów do stworzenia grup podobnych obiektów. Zakłada się wykorzystanie jednego z istniejących algorytmów klasteryzacji w taki sposób, aby obiekty w jednym klastrze były maksymalnie podobne do siebie, a obiekty z różnych klastrów maksymalnie się różniły.
5. Sprawdzenie wiarygodności wyników klasteryzacji. Może to być osiągnięte za pomocą różnych metod, takich jak wizualizacja klastrów, analiza stabilności klastrów i ocena jakości podziału.

#### 1.1.5. Zastosowanie

Zastosowanie metody analizy klastrów jest z powodzeniem wykorzystywane w wielu dziedzinach i dyscyplinach. Metoda ta jest również szeroko stosowana w informatyce. Najbardziej udanymi przykładami wykorzystania klasteryzacji w pracy z danymi tekstowymi są:

- klasteryzacja wyników wyszukiwania
- grupowanie użytkowników o podobnych zainteresowaniach, co pomaga w personalizacji treści, rekomendacjach i targetowaniu w reklamie
- grupowanie tekstowych dokumentów według ich tematyki, takich jak artykuły prasowe, publikacje naukowe lub wiadomości w mediach społecznościowych

#### 1.1.6. Оценка эффективности кластеризации

Problem oceny jakości w zadaniu klasteryzacji jest trudny do rozwiązania z co najmniej dwóch powodów. Po pierwsze, zgodnie z twierdzeniem o niemożności Kleinberga [3], nie istnieje optymalny algorytm klasteryzacji. Po drugie, wiele algorytmów klasteryzacji nie jest w stanie samodzielnie określić rzeczywistej liczby klastrów w danych; najczęściej liczba klastrów jest zadawana na wejściu algorytmu i dobierana kilkoma uruchomieniami. Niemniej jednak, wyróżnia się dwie grupy metod oceny jakości klasteryzacji:

1. Zewnętrzne metody porównują wynik klasteryzacji z apriorycznie znanym podziałem na klasy. Przykładem zewnętrznej metody oceny może być indeks Randa. Metoda ta ocenia, jak wiele spośród par elementów, które znajdowały się w jednej klasie, oraz tych, które znajdowały się w różnych klasach, zachowało ten stan po zastosowaniu algorytmu klasteryzacji.
2. Wewnętrzne metody oceniają jakość klasteryzacji, korzystając jedynie z informacji zawartych w samych danych. Przykładem wewnętrznej metody jest metoda Spójności Klastra (Cluster Cohesion). Idea tej metody polega na tym, że im bliżej siebie znajdują się obiekty w klastrach, tym lepsze jest podzielenie. Można również wyróżnić metodę Separacji Klastra (Cluster Separation), która ocenia jakość klasteryzacji na podstawie odległości między obiektami różnych klastrów.

## 1.2. Przetwarzanie języka naturalnego

### 1.2.1. Definicja

Przetwarzanie języka naturalnego (ang. Natural Language Processing, NLP) to dziedzina badań, która łączy metody uczenia maszynowego i lingwistyki matematycznej w celu opracowania algorytmów do analizy, rozumienia i generowania tekstu w językach naturalnych. [5]

### 1.2.2. Przedprzetwarzanie tekstu

Przedprzetwarzanie tekstu jest ważnym etapem NLP, który przekształca tekst w języku naturalnym w format wygodny do rozpoznawania przez algorytmy i dalszej pracy. Przedprzetwarzanie składa się z różnych etapów, które mogą się różnić w zależności od zadania i implementacji. Oto niektóre z najpopularniejszych podstawowych podejść:

- Konwersja wszystkich liter w tekście na małe lub wielkie litery
- Usuwanie cyfr (liczb) lub zamiana na tekstowy ekwiwalent (zwykle używane są wyrażenia regularne)
- Usuwanie znaków białych (whitespaces)
- Tokenizacja (zwykle realizowana na podstawie wyrażeń regularnych)
- Usuwanie słów stop
- Stemming
- Lematyzacja
- Wektoryzacja

Poniżej bardziej szczegółowo opisane są trzy ostatnie metody.

### 1.2.3. Stemming

Liczba poprawnych form wyrazowych, których znaczenia są podobne, ale pisownia różni się sufiksami, przedrostkami, końcówkami itp., jest bardzo duża, co utrudnia tworzenie słowników i dalsze przetwarzanie. Stemming pozwala sprowadzić słowo do jego podstawowej formy. Istota podejścia polega na znalezieniu rdzenia słowa, w tym celu z końca i początku słowa kolejno odcinane są jego części. Reguły odcinania dla stemmera tworzone są z góry i najczęściej stanowią wyrażenia regularne, co sprawia, że podejście to jest pracochłonne, ponieważ przy dodawaniu kolejnego języka potrzebne są nowe badania lingwistyczne. Drugą wadą podejścia jest możliwa utrata informacji podczas odcinania części, na przykład możemy stracić informację o części mowy.

#### 1.2.4. Lematyzacja

Podejście to stanowi alternatywę dla stemmingu. Główna idea polega na sprowadzeniu słowa do formy słownikowej — lematu. Przykład lematyzacji dla języka polskiego:

- dla rzeczowników — mianownik liczby pojedynczej;
- dla przymiotników — mianownik liczby pojedynczej, rodzaj męski;
- dla czasowników, imiesłów, rzeczowników odczasownikowych — czasownik w bezokoliczniku aspektu niedokonanego.

### 1.3. Векторизация

#### 1.3.1. Opis

Większość modeli matematycznych działa w przestrzeniach wektorowych o dużych wymiarach, dlatego konieczne jest przekształcenie tekstu w przestrzeń wektorową. Głównym celem wektoryzacji tekstu jest stworzenie reprezentacji danych tekstowych, która zachowuje informacje semantyczne i syntaktyczne o tekście, a jednocześnie nadaje się do wykorzystania w modelach matematycznych. Pozwala to na efektywną analizę i przetwarzanie danych tekstowych za pomocą algorytmów i metod uczenia maszynowego. [6]

#### 1.3.2. Celi wektoryzacji

Główne celi wektoryzacji tekstu obejmują:

- Reprezentowanie tekstu do analizy przez algorytmy maszynowe.
- Znajdowanie podobieństw między tekstami lub kategoryzowanie tekstów według tematyki.
- Rozwiązywanie zadań klasyfikacji, klasteryzacji lub regresji na podstawie danych tekstowych.
- Tworzenie modeli uczenia maszynowego do automatycznego uogólniania informacji tekstowych.

#### 1.3.3. Metody wektoryzacji

Pierwszym krokiem w wektoryzacji tekstu jest określenie przestrzeni cech, w której tekst będzie reprezentowany jako wektory. Każda cecha może reprezentować słowo, frazę, symbol lub inny element tekstu. Istnieje kilka metod wektoryzacji tekstu, w tym:

- Bag of Words (BoW): Każdy dokument jest reprezentowany jako wektor, gdzie każdy element odpowiada poszczególnemu słowu, a wartościami są częstotliwości występowania słów w dokumencie.
- TF-IDF (Term Frequency-Inverse Document Frequency): Metoda ta uwzględnia nie tylko częstotliwość występowania słów w dokumencie (TF), ale także odwrotną częstotliwość występowania słowa we wszystkich dokumentach korpusu (IDF), co pozwala wyróżnić słowa kluczowe.
- Word Embeddings: Są to metody oparte na trenowaniu sieci neuronowych, które przekształcają słowa w wektory w przestrzeni ciągłej, zachowując ich właściwości semantyczne.

### 1.3.4. Zastosowanie

Wektoryzacja tekstu znajduje zastosowanie w wielu dziedzinach, w tym:

- Analiza настроjów i sentymentów w mediach społecznościowych i recenzjach.
- Klasyfikacja tekstów według tematyki lub kategorii.
- Systemy rekomendacyjne oparte na opisach tekstowych.
- Automatyczne tłumaczenie tekstu i implementacja chatbotów.

## 1.4. снижения размерности

### 1.4.1. Дефиниция

Снижение размерности - это процесс уменьшения количества признаков (измерений) в наборе данных, сохраняя при этом как можно больше информации. Часто используется в машинном обучении где главная цель устранение в наборе признаков избыточных, неинформативных или слабо информативных которые могут понизить эффективность модели, а после такого преобразования она упрощается, и соответственно уменьшается размер набора данных в памяти и ускоряется работа алгоритмов на нем. Так же используется для упрощения анализа и визуализации данных, например путем редукции до низких размерностей таких как 2D или 3D. [7]

### 1.4.2. Методы снижения размерности

Уменьшение размерности может быть осуществлено методами выбора признаков (англ. feature selection) или выделения признаков (англ. feature extraction).

#### Выбор признаков

Метод Выбор признаков оставляют некоторое подмножество исходного набора признаков, избавляясь от признаков избыточных и слабо информативных. Основные преимущества этого класса алгоритмов:

- Уменьшение вероятности переобучения
- Увеличение точности предсказания модели;
- Сокращение времени обучения;
- Увеличивается семантическое понимание модели.

#### Выделение признаков

Другим способом уменьшить размерность входных данных является выделение признаков. Эти методы каким-то образом составляют из уже исходных признаков новые, все также полностью описывающие пространство набора данных, но уменьшая его размерность и теряя в репрезентативности данных, т.к. становится непонятно, за что отвечают новые признаки. Все методы feature extraction можно разделить на линейные и нелинейные.

## Методы выделения признаков

Линейные методы основываются на предположение, что зависимость между независимыми переменными и зависимой переменной линейна. Поэтому они с помощью линейных комбинаций, которые означают, что изменение в одной переменной приводит к пропорциональному изменению в другой. Одним из самых известных методов линейного выделения признаков является PCA . Основной идеей этого метода является поиск гиперплоскости, на которую при ортогональной проекции всех признаков максимизируется дисперсия. Нелинейные методы позволяют моделировать более сложные и нелинейные зависимости между переменными. Алгоритм t-sne стремится сохранить относительные расстояния между точками в исходном пространстве высокой размерности, а затем пытается воссоздать это распределение в низкоразмерном пространстве.

### 1.4.3. Применение

Метод снижения размерности часто и успешно используется в Машинном обучении , Обработка изображений и компьютерное зрение, Биоинформатика и геномика , Финансовый анализ и торговля и многие другие области. Где снижение количества признаков позволяет оптимизировать обработку данных.

## Глава 2

## Rozdział 2

### 2.1. Приложение

#### 2.1.1. Opis

Данный проект направлен на создание инструмента для анализа текстовых данных, включающего предобработку текста, векторизацию и кластеризацию. Основная цель приложения — предоставить пользователю возможность гибкой настройки и использования различных методов обработки данных для решения конкретных задач.

#### 2.1.2. Технологии

Для реализации проекта был выбран язык программирования Python. Он обладает рядом преимуществ: [?]

- Богатый набор библиотек для реализации различных методов, таких как предобработка текста, векторизация и кластеризация.
- Высокая эффективность для анализа данных благодаря библиотекам, таким как NumPy, scikit-learn.
- Сообщество и документация, которые облегчают разработку и поддержку проектов.

#### 2.1.3. GUI

Основная идея приложения заключается в создании удобного инструмента для анализа текстовых данных, где пользователь может самостоятельно выбирать и настраивать методы обработки под конкретные задачи с помощью графического интерфейса. Для разработки графического интерфейса были использованы библиотеки PyQt и Matplotlib:

- PyQt: Обеспечивает создание интуитивно понятных пользовательских интерфейсов.
- Matplotlib: Используется для визуализации данных, что позволяет пользователю легко интерпретировать результаты анализа.

Графический интерфейс позволяет пользователю загружать данные, настраивать параметры обработки текста, векторизации и кластеризации, а также визуализировать результаты.

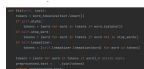


Рис. 2.1: Przykład rys

## 2.1.4. OOP

Главным компонентом моего приложения является класс **Workspace**, который представляет собой контейнер, хранящий параметры для трех основных процессов:

- **Обработка текста:** Включает методы предобработки, такие как токенизация, удаление стоп-слов и лемматизация.
- **Векторизация:** Преобразует текстовые данные в числовые векторы с использованием методов, таких как TF-IDF или Word2Vec.
- **Кластеризация:** Реализует алгоритмы, такие как K-means или DBSCAN, для группировки текстовых данных.

Класс **Workspace** обеспечивает структурированное хранение и управление параметрами каждого этапа, что упрощает процесс настройки и выполнения анализа.

## 2.2. Текстовые данные

### 2.2.1. Opis

Для реализации данного проекта потребуются текстовые данные, которые в данном случае будут комментариями с YouTube. Мы будем использовать YouTube Data API v3 для получения информации о видео и связанных с ним комментариях. Данное API является бесплатным и простым в использовании. Для начала работы необходимо зарегистрироваться и создать проект в Google Cloud, а затем активировать YouTube Data API. После этого будет сгенерирован API-ключ, который мы будем использовать для аутентификации.

### 2.2.2. API запросы

В данном проекте будут использованы два HTTP-запроса:

- "GET <https://www.googleapis.com/youtube/v3/commentThreads>"
- "GET <https://www.googleapis.com/youtube/v3/videos>"

Первый запрос помогает получить комментарии к конкретному видео, а второй — получить информацию о видео. Подробно ознакомиться с полной документацией можно в [8]. Этапы получения комментариев:

1. Пользователь с помощью графического интерфейса (GUI) вводит ссылку на интересующее его видео.
2. Программа извлекает video\_id из предоставленной ссылки.
3. Используя video\_id, выполняется запрос GET <https://www.googleapis.com/youtube/v3/videos> для получения краткой информации о видео. Это позволяет пользователю убедиться, что выбранное видео соответствует его ожиданиям.
4. После подтверждения пользователем, что видео корректное, выполняется запрос GET <https://www.googleapis.com/youtube/v3/commentThreads>, чтобы получить комментарии к этому видео.
5. Полученные комментарии сохраняются в текстовый файл, где каждый комментарий записан в отдельной строке. Название файла соответствует уникальному идентификатору видео (video\_id).



### 2.2.3. Обработка полученных данных

Результаты наших запросов будут предоставлены в формате JSON. Подробную информацию о структуре JSON-ответов можно найти в документации [8]. Для получения информации о видео будут использоваться только следующие поля:

- title: название видеоролика
- channelTitle: имя автора канала
- thumbnails: URL-адрес изображения

При получении комментариев необходимы только следующие поля:

- textDisplay: содержание комментария
- replies: контейнер комментариев-ответов

Поле textDisplay содержит текст комментария, а поле replies является контейнером для комментариев-ответов. Каждый ответ в поле replies.comments также имеет поле textDisplay, которое мы можем отдельно обработать. В конечном итоге мы получим список всех комментариев, где каждый элемент представляет из себя отдельный комментарий.

### 2.2.4. Хранение

После получения все комментарии будут сохраняться в отдельном каталоге comments, который будет создаваться в случае его отсутствия, где каждый файл с комментариями будет иметь название **video\_id.txt**, для возможности индексации. Для упрощения процесса хранения и доступа к данным, а также для обеспечения их совместимости с различными инструментами анализа, принято решение использовать обычный текстовый файл (.txt) для хранения комментариев. Каждый комментарий будет записан в отдельную строку файла, что упрощает структурирование данных и позволяет легко считывать их для анализа. Такой способ хранения данных был выбран по нескольким причинам:

- **Простота реализации:** Текстовые файлы легко создаются, читаются и редактируются с помощью множества программных средств и языков программирования.
- **Совместимость:** Практически все аналитические инструменты и библиотеки могут работать с текстовыми файлами, что облегчает интеграцию данных в различные этапы анализа.
- **Малое влияние на анализ:** Формат хранения не оказывает значительного влияния на сам процесс анализа данных. Основное внимание уделяется содержанию комментариев, а не способу их хранения.

## 2.3. Обработка текста

### 2.3.1. Opis

Следующим процессом является обработка текста. Важно понимать, как выглядят комментарии после их получения и сохранения с YouTube. Реализация обработки будет выполнена с помощью библиотеки NLTK (Natural Language Toolkit) и встроенных функций Python. Ознакомиться с документацией NLTK можно здесь [9]. Обработка текста разделена на несколько этапов:

- Обязательная обработка выполняющаяся независимо от выбранных параметров пользователем.
- Примитивная обработка удаление символ и цифр
- Удаление stop-words

- Лематизация
- Анализ

Все методы кроме обязательной обработки и анализа, являются необязательным, то есть будут выполняться только в случае если пользователь укажет что хочет использовать данные методы.

### 2.3.2. Обязательная обработка

Обязательными этапами обработки текста являются приведение слов к нижнему регистру и процесс токенизации, который устраняет необходимость удаления пробелов. С помощью функции `nlk.word tokenize` каждое слово будет разделено на токены. Эти этапы обязательны, так как они упрощают дальнейшую работу с текстом и гарантируют правильную работу многих последующих функций, которые могут быть чувствительны к регистру.

### 2.3.3. Примитивная обработка

Примитивная обработка подразумевает удаление неалфавитных символов или слов. Этот процесс не является обязательным, однако его использование помогает избавиться от различных языков разметки, которые YouTube использует для визуализации смайликов или инструментов изменения текста. Пример таких комментариев:

```
"Great video! (here should be emoji"
"Check this out: <a href='https://example.com'>link</a>"
```

В данном проекте это реализовано с помощью встроенной функции Python `isalpha()`, которая проверяет символы или слова на наличие неалфавитных символов.

```
1 def remove_non_alpha(tokens):
2     return [word for word in tokens if word.isalpha()]
```

### 2.3.4. Удаление stop-words

Этот этап реализует удаление слов, которые не несут значимой информации. Типичными примерами таких слов являются:

```
"the", "is", "in", "and"
```

Удаление стоп-слов уменьшает размер текста и делает информативные слова более заметными. Для этого используется словарь стоп-слов из библиотеки NLTK, который в случае необходимости можно дополнить своими словами.

```
1 from nltk.corpus import stopwords
2
3 def remove_stop_words(tokens):
4     stop_words = set(stopwords.words('english'))
5     return [word for word in tokens if word not in stop_words]
```

### 2.3.5. Лематизация

Последним необязательным процессом является лемматизация, которая приводит слова к их базовой форме. Это полезно в случаях, когда нам нужно сгруппировать разные формы одного слова (например, "running" "ran" "runs" к "run").

```
1 from nltk.stem import WordNetLemmatizer
2
3 def lemmatize_tokens(tokens):
4     lemmatizer = WordNetLemmatizer()
5     return [lemmatizer.lemmatize(word) for word in tokens]
```

### 2.3.6. Анализ

В конце, независимо от выбранных пользователем этапов, все токены проверяются на непустые значения, чтобы избежать полностью пустых слов и строк. Далее все токены объединяются в одну строку, разделённую пробелами. На этом этапе результат обработки возвращается и анализируется функциями частотного анализа (**freq\_analysis**) и анализа длины (**length\_analysis**). Эти функции будут использоваться для построения графиков, которые помогут пользователю визуально оценить частоту встречаемых слов и длины комментариев для последующего принятия решений.

#### freq analysis

Эта функция использует обработанные строки, создавая единый массив со всеми словами из всех комментариев, и применяет класс **collections.Counter** в Python. **Counter** используется для подсчёта хэшируемых объектов, таких как строки или числа, и предоставляет удобные методы для работы с частотными данными. Таким образом, функция генерирует два массива: массив слов и соответствующий ему массив количества каждого слова в тексте. Эти массивы будут использоваться для построения столбчатой диаграммы (bar chart), показывающей частоту встречаемости слов.

Пример реализации алгоритма в проекте:

```
1 from collections import Counter
2 import matplotlib.pyplot as plt
3
4 def freq_analysis(tokens):
5     data = ' '.join(self.nlp_data)
6     words = nltk.word_tokenize(data)
7     # words = [word.lower() for word in words]
8     # words = [word for word in words if word.isalnum()]
9     word_freq = Counter(words)
10    words_counts_sorted = word_freq.most_common()
11    self.words_sorted = [word_count[0] for word_count in words_counts_sorted]
12    self.counts_sorted = [word_count[1] for word_count in words_counts_sorted]
```

#### length analysis

Эта функция также использует обработанные строки комментариев, определяя количество слов для каждого комментария. В конечном итоге генерируется массив количеств слов, который будет использован для построения гистограммы (histogram), показывающей распределение длины комментариев.

Пример реализации алгоритма в проекте:

```
1 from collections import Counter
2 import matplotlib.pyplot as plt
3
4 def freq_analysis(tokens):
5     word_counts = [len(nltk.word_tokenize(el)) for el in self.nlp_data]
6     self.max_size = max(word_counts) + 1
7     self.lengths = [0] * (self.max_size + 1)
8     for count in word_counts:
9         self.lengths[count] += 1
10    endTime = time.time()
```

### 2.3.7. Итог

Все этапы обработки текста, от обязательных до необязательных, играют важную роль в подготовке данных для дальнейшего анализа. Их правильное выполнение гарантирует структурированный и удобный вид для обработки. В визуализация графиков анализирующих содержания комментариев позволит пользователям лучше понимать структуру и содержание комментариев, а также принимать обоснованные решения на основе представленных данных.

## 2.4. Wektoryzacja

### 2.4.1. Opis

На данном этапе текст готов к следующему этапу обработки. Для последующего анализа необходимо преобразовать комментарии в векторную форму (векторизировать). Будт использовано два метода: TF-IDF и Word2Vec. Эти методы позволят представить текстовые данные в виде числовых векторов, что облегчит дальнейший анализ и обработку. Векторизация будет реализована с использованием библиотек **scikit-learn** для метода TF-IDF и **gensim** для метода Word2Vec.

### 2.4.2. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) — это статистическая мера, используемая для оценки важности термина в документе относительно коллекции документов (корпуса). Она широко применяется в задачах обработки естественного языка и информационного поиска. В библиотеке Scikit-learn существует удобный инструмент для расчета TF-IDF: класс `TfidfVectorizer`.

#### Параметры

- `max_df`: Число или доля. Термины, которые встречаются в большем числе документов, чем указано, игнорируются. Полезно для удаления общеупотребительных слов.
- `min_df`: Число или доля. Термины, которые встречаются в меньшем числе документов, чем указано, игнорируются. Полезно для удаления редких слов.
- `gram_range`: Тупл (`min_n`, `max_n`). Определяет диапазон для извлечения n-грамм.
- `stop_words`: Список слов или строка, определяющая слова, которые следует игнорировать.

#### Пояснение алгоритма

- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- TF-IDF

#### Пример использования

```
from sklearn.feature_extraction.text import TfidfVectorizer

def vectorize_tfidf(corpus):
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(corpus)
    return tfidf_matrix
```

Пример преобразования текстовых данных:

- До: ["This is a great video "I love this content"]
- После: [[0.5, 0.5, 0, 0.5, 0, 0.5], [0, 0, 0.7, 0, 0.7, 0]]

Для работы с TF-IDF необходимо иметь корпус текстов, который можно представить в виде списка строк, где каждая строка — это отдельный документ.

### 2.4.3. Word2Vec

Word2Vec — это алгоритм, разработанный для обучения векторных представлений слов (эмбеддингов), которые захватывают семантические отношения между словами. В библиотеке Gensim предоставляется удобный интерфейс для работы с Word2Vec.

#### Важные параметры

- `sentences`: Корпус текстов, представленный в виде списка списков слов.
- `vector_size`: Размерность векторного представления слов.
- `window`: Максимальное расстояние между текущим и предсказанным словом в предложении.
- `min_count`: Минимальная частота слова в корпусе для его учета.
- `workers`: Количество потоков для параллельного обучения.

Более глобальные настройки

- SG (Skip-Gram) и CBOW (Continuous Bag of Words): В Word2Vec есть два основных архитектурных подхода — Skip-Gram и CBOW. Параметр `sg` в Gensim позволяет выбрать между ними:
  - `sg=0` использует CBOW (по умолчанию).
  - `sg=1` использует Skip-Gram.
- Negative Sampling: В Word2Vec используется негативное сэмплирование для повышения эффективности обучения. Параметр `negative` задает количество "негативных" примеров для каждого положительного примера.
- Learning Rate: Параметр `alpha` задает начальную скорость обучения, а `min_alpha` — минимальную скорость обучения, которая уменьшается до этого значения по мере обучения.

#### Описание алгоритма

1. Инициализация весов: Веса нейронной сети инициализируются случайными значениями. Это веса между входным слоем и скрытым слоем, а также между скрытым слоем и выходным слоем.
2. Прямой проход (Forward Pass): Для каждого слова в предложении создается входной вектор. В случае Skip-Gram этот входной вектор используется для предсказания контекстных слов. В случае CBOW контекстные слова используются для предсказания текущего слова.
  - Skip-Gram:
    - Входное слово кодируется в виде одnorазового вектора.
    - Одноразовый вектор умножается на веса между входным и скрытым слоями для получения скрытого представления.
    - Скрытое представление умножается на веса между скрытым и выходным слоями для предсказания контекстных слов.
  - CBOW:
    - Контекстные слова кодируются в виде одnorазовых векторов.
    - Одноразовые векторы суммируются и умножаются на веса между входным и скрытым слоями для получения скрытого представления.
    - Скрытое представление умножается на веса между скрытым и выходным слоями для предсказания текущего слова.

3. Обратное распространение (Backpropagation): Ошибка предсказания рассчитывается путем сравнения предсказанного слова с фактическим. Затем эта ошибка распространяется обратно через сеть для обновления весов. Используется метод стохастического градиентного спуска (SGD) и, в некоторых случаях, негативное сэмплирование для ускорения обучения.
  - Обновление весов: Веса обновляются в направлении, противоположном градиенту ошибки. Это помогает минимизировать ошибку предсказания.
4. Повторение процесса: Этот процесс повторяется для всех слов в предложении и для всех предложений в корпусе. Модель проходит через корпус несколько раз, как определено параметром epochs.

### Пример использования

```
from gensim.models import Word2Vec

def vectorize_word2vec(tokens):
    model = Word2Vec(sentences=tokens, vector_size=100, window=5, min_count=1, workers=4)
    word_vectors = model.wv
    return word_vectors
```

Пример преобразования текстовых данных:

- До: ["This is a great video "I love this content"]
- После: Векторы слов: "this": [...], "is": [...], "a": [...], "great": [...], "video": [...], "I": [...], "love": [...], "content": [...]

### 2.4.4. Осена

После векторизации текста с использованием методов TF-IDF и Word2Vec важно провести оценку их эффективности и понять, насколько хорошо они справляются с задачей представления текстовых данных. В данной секции будут рассмотрены критерии оценки и результаты применения каждого из методов на наших данных.

### Критерии

Для оценки работы методов TF-IDF и Word2Vec используются следующие критерии:

- **Семантическая значимость:** Насколько хорошо метод улавливает семантические связи между словами и контекст их использования.
- **Скорость вычислений:** Время, необходимое для векторизации текстов.
- **Простота интерпретации:** Насколько легко интерпретировать результаты векторизации и использовать их в дальнейших анализах.
- **Применимость к задачам машинного обучения:** Насколько хорошо векторы подходят для задач классификации, кластеризации и других методов машинного обучения.

Оценка TF-IDF

Оценка Word2Vec

## 2.5. Кластеризация

### 2.5.1. Opis

Следующим этапом обработки данных является кластеризация, которая использует векторизированные комментарии и разбивает их на кластеры. Реализовано это будет с помощью двух методов K-means и DBSCAN используя библиотеку sklearn.

### 2.5.2. K-means

Алгоритм K-means разбивает набор данных на непересекающиеся кластеры, каждый из которых описывается средним значением (центроидом) образцов в кластере. K-means стремится выбрать центроиды, минимизируя инерцию, или критерий суммы квадратов внутри кластеров.

#### Описание алгоритма

Алгоритм K-means можно понять через три основных шага. Первый шаг заключается в выборе начальных центроидов. Обычно используется простой метод выбора случайных образцов из набора данных. После инициализации K-means выполняется последовательное выполнение двух других шагов. Первый шаг заключается в присвоении каждого образца ближайшему центроиду. Второй шаг создает новые центроиды, вычисляя среднее значение всех образцов, принадлежащих каждому предыдущему центроиду. Разница между старыми и новыми центроидами вычисляется, и алгоритм повторяет эти два последних шага до тех пор, пока эта разница не станет менее определенного порога. Другими словами, алгоритм повторяется, пока центроиды не перестанут существенно перемещаться.

Как и многие алгоритмы, K-means подвержен проблеме локальных минимумов, что сильно зависит от начальной инициализации центроидов. Обычно алгоритм запускается несколько раз с различными инициализациями центроидов.

#### Параметры

Для настройки алгоритма K-means используются следующие параметры:

- `n_clusters`: Количество кластеров (K), на которые нужно разделить данные.
- `init`: Метод инициализации центроидов.
- `max_iter`: Максимальное количество итераций для одного запуска K-means.
- `tol`: Порог для остановки алгоритма. Если разница между старыми и новыми центроидами меньше этого значения, алгоритм останавливается.
- `n_init`: Количество запусков алгоритма с различными инициализациями. Результат с наименьшей инерцией будет выбран как окончательный.
- `random_state`: Начальное значение генератора случайных чисел для воспроизводимости результатов.
- `algorithm`: Алгоритм для вычисления K-means. Возможные значения включают 'auto', 'full', и 'elkan'.

### Пример использования

```
kmeans = KMeans(n_clusters=2, init='k-means++', max_iter=300, n_init=10, random_state=0)

# Обучение модели
kmeans.fit(X)

# Предсказание кластеров для данных
y_kmeans = kmeans.predict(X)
```

Пример преобразования данных: были точки, а теперь число которое показывает в каком кластере.

### 2.5.3. DBSCAN

Алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) рассматривает кластеры как области высокой плотности, отделенные от областей низкой плотности. Это позволяет DBSCAN обнаруживать кластеры любой формы, в отличие от K-means, который предполагает, что кластеры имеют выпуклую форму. Основной компонент DBSCAN - это понятие ядерных образцов, которые представляют собой образцы в областях высокой плотности.

#### Параметры

Параметры алгоритма DBSCAN:

- `eps`: Максимальное расстояние между двумя образцами, при котором один считается соседом другого. Это параметр радиуса для определения плотности области.
- `min_samples`: Минимальное число образцов (включая сам образец), необходимое для того, чтобы область считалась плотной, т.е. чтобы образец стал ядерным.
- `metric`: Метрика для вычисления расстояний между точками (по умолчанию 'euclidean'). Это может быть любая метрика, поддерживаемая функцией `scipy.spatial.distance.pdist`.
- `algorithm`: Алгоритм, используемый для вычисления ближайших соседей ('auto', 'ball\_tree', 'kd\_tree', 'brute').
- `leaf_size`: Размер листа, передаваемый алгоритмам 'ball\_tree', 'kd\_tree'. Это может повлиять на скорость построения и запроса ближайших соседей.
- `p`: Степень Минковского расстояния, когда используется метрика 'minkowski' (по умолчанию `p=2`, что соответствует евклидову расстоянию).

#### Описание алгоритма

Алгоритм DBSCAN реализуется следующим образом:

1. Определяются ядерные образцы: Образцы, для которых количество соседей в пределах `eps` не меньше `min_samples`.
2. Строятся кластеры: Для каждого ядерного образца строится кластер путем нахождения всех его соседей, которые также являются ядерными образцами, и их собственных соседей и так далее. Каждый кластер также включает не-ядерные образцы, которые являются соседями ядерных образцов в кластере.
3. Выделяются выбросы: Все образцы, которые не являются ядерными и не находятся на достаточном расстоянии от ядерных образцов, считаются выбросами.

Алгоритм DBSCAN детерминирован и всегда генерирует одинаковые кластеры при одинаковых входных данных в одном и том же порядке. Однако результаты могут отличаться при предоставлении данных в другом порядке. Это происходит из-за того, что порядок обработки данных влияет на то, каким образом назначаются образцы кластерам.



### Пример использования

```
X = np.array([[0, 1], [1, 1], [2, 1], [3, 1], [4, 1],
              [0, 0], [1, 0], [2, 0], [3, 0], [4, 0]])

# Создание модели t-SNE
tsne = TSNE(n_components=2, perplexity=5, n_iter=300, random_state=0)

# Преобразование данных
X_embedded = tsne.fit_transform(X)
```

Пример преобразования данных: были точки, а теперь число которое показывает в каком кластере.

### 2.5.4. Выбор параметров

## 2.6. Визуализация

### 2.6.1. Opis

Заключительный процесс анализа это визуализация данных. Итоговый результат представляет из себя **2D Plot**, где каждая точка будет представлять некоторые текстовые данные (комментарии) с определённым цветом, соответствующим цвету кластера, которому она принадлежит, или же цвету отсутствия кластера как в случае DBSCAN. Так же на графике будут визуализироваться центры каждого кластера в виде красного крестика. Для удобства навигации и проверки результатов при наведении на точку(комментарий) или крустик(центр кластера) будет высвечиваться информация с содержанием элемента, которая репрезентуется с помощью библиотеки **mplcursor**. В случае точки это содержание обработанного комментария. В случае крестика репрезентуется центр кластера, которая с помощью метода обратного векторизации будет пытаться представить данную точку как слово.

### 2.6.2. Снижение размерности

Для возможности визуализации векторизованных данных в качестве 2D точек необходимо воспользоваться методом снижения размерности (T-SNE). Данный метод является нелейным методом выделения признаков. Изначально трудно определить за что отвечают признаки, то выбрать какие-то наиболее подходящие признаки невозможно. Нелейный метод поможет нам сохранить неленные связи.

### 2.6.3. T-SNE

#### Параметры

- `n_components`: Количество измерений, в которое нужно проецировать данные (обычно 2 или 3 для визуализации).
- `perplexity`: Параметр, связанный с количеством ближайших соседей. Он влияет на баланс между локальной и глобальной структурой данных.
- `early_exaggeration`: Параметр, который контролирует расстояния между точками в начальной фазе оптимизации.
- `learning_rate`: Скорость обучения для градиентного спуска.
- `n_iter`: Количество итераций для оптимизации.
- `metric`: Метрика для вычисления расстояний между точками (по умолчанию 'euclidean').
- `init`: Метод инициализации ('random' или 'pca').

- `random_state`: Начальное значение генератора случайных чисел для воспроизводимости результатов.

### Описание алгоритма

Алгоритм t-SNE выполняется следующим образом:

1. Преобразование расстояний в вероятности: Для каждой точки рассчитываются вероятности Пирсона, которые определяют вероятность того, что соседняя точка будет выбрана в качестве соседа.
2. Определение целевых вероятностей: На низкоразмерной проекции рассчитываются вероятности  $Q$ , аналогичные вероятностям Пирсона, но в другой размерности.
3. Минимизация расхождения (Kullback-Leibler divergence): Градиентный спуск используется для минимизации расхождения между распределениями  $P$  и  $Q$ , что приводит к сохранению близости точек.

### Пример использования

```
X = np.array([[1, 2], [2, 2], [2, 3],  
              [8, 7], [8, 8], [25, 80]])
```

```
# Создание модели DBSCAN  
dbscan = DBSCAN(eps=3, min_samples=2)
```

```
# Обучение модели  
dbscan.fit(X)
```

```
# Предсказание кластеров для данных  
labels = dbscan.labels_
```

Пример преобразования данных: [тут фотку](#)

## Глава 3

# Rozdział 3

### 3.1. Описание приложения

#### 3.1.1. Рабочее окно

Графическое приложение представляет собой окно с меню-баром, содержащим два поля ("File" и "Comments"), и основной зоной, где располагается таб-меню. Каждый таб представляет собой отдельное рабочее пространство (workspace). Рассмотрим эти элементы подробнее.

#### 3.1.2. Меню

Меню-бар позволяет быстро навигировать по рабочему окну в любой момент. В меню-баре реализованы следующие функции:

- Открытие диалогового окна для создания нового рабочего пространства ("File> "New Project").
- Открытие диалогового окна для получения комментариев с YouTube ("Comments> "YouTube").

Рассмотрим более подробно использование этих функций:

#### Создание workspace

Для создания нового рабочего пространства необходимо указать имя рабочей области (Workspace name) и файл-источник (Source file) с текстовыми данными. Вот как выглядит пустое окно **CreateWorkspaceWindowEmpty**. Пользователю необходимо ввести любое имя для рабочего пространства, а также указать путь к текстовому файлу с данными. Это можно сделать вручную или выбрать файл с помощью диалогового окна `FileDialog` из библиотеки `PyQt`, что позволяет найти необходимый файл на компьютере. Пример правильно заполненного окна **CreateWorkspaceWindowFill**.

#### Получение комментариев

Для получения комментариев используется диалоговое окно, в котором необходимо указать ссылку на интересующее видео, как показано на скриншоте **CommentsWindowStep1**. Если видео найдено, то в поле Video Info появится информация о нем, позволяющая пользователю убедиться, что это именно то видео, комментарии которого необходимо получить. Далее пользователю остается только нажать на кнопку Download comments, которая сохранит все комментарии.

### 3.1.3. Workspace

После создания хотя бы одного рабочего пространства появляется меню навигации по вкладкам, где каждая вкладка соответствует отдельному рабочему пространству. По этим вкладкам можно свободно переключаться и закрывать их. Содержимое каждой вкладки включает:

- Поле для графиков.
- Кнопки Previous и Next для переключения графиков в поле для графиков.
- Кнопку Text для открытия диалогового окна управления обработкой текста.
- Кнопку Vectorization для открытия диалогового окна управления векторизацией.
- Кнопку Cluster для открытия диалогового окна управления кластеризацией.
- Кнопку Apply для применения всех выбранных параметров и открытия окна с результатами кластеризации.

#### Поле для графиков

Поле для графиков в каждый момент времени может отображать один из трех предложенных графиков:

- График частот слов (freq). Описан ранее в
- График длин текста(length). Описан ранее в
- График ближайших соседей (k-nearest). Описан ранее в

С помощью кнопок Previous и Next графики можно переключать соответственно вперед и назад.

### 3.1.4. Text

Кнопка Text открывает диалоговое окно, в котором пользователь может указать параметры для обработки текста. Вот как выглядит окно по умолчанию **TextProcessingWindowStandard**. По умолчанию все параметры установлены на значение true, то есть используются. Пользователь может выбрать один из предложенных опций обработки текста:

- Alpha - примитивная обработка описанная ранее
- Stop-words - удаление стоп слов описанная ранее
- Lemmatize - лемматизация описанная ранее

#### Vectorization

Кнопка Vectorization открывает диалоговое окно, в котором пользователь может выбрать один из предложенных методов векторизации и указать параметры для выбранного метода. Вот как выглядит окно по умолчанию **VectorizationWindowStandard**. По умолчанию выбран метод TF-IDF с параметрами по умолчанию. Далее описан список доступных методов и параметров для каждого метода:

- Метод TF-IDF. Доступные параметры для настройки:
  - max\_df: Диапазон допустимых значений ()
  - min\_df: Диапазон допустимых значений ()
- Методы Word2vec. Доступные параметры для настройки:

- vector\_size: Диапазон допустимых значений ()
- window: Диапазон допустимых значений ()
- min\_count: Диапазон допустимых значений ()
- sg: Диапазон допустимых значений ()

## Cluster

Кнопка Cluster открывает диалоговое окно, в котором пользователь может выбрать один из предложенных методов кластеризации и указать параметры для выбранного метода. Вот как выглядит окно по умолчанию **ClusterWindowStandard**. По умолчанию выбран метод K-means с параметрами по умолчанию. Далее описан список доступных методов и параметров для каждого метода:

- Метод K-means. Доступные параметры для настройки:
  - n\_cluster: Диапазон допустимых значений ()
  - max\_iter: Диапазон допустимых значений ()
  - n\_iter: Диапазон допустимых значений ()
- Методы DBSCAN. Доступные параметры для настройки:
  - min\_samples: Диапазон допустимых значений ()
  - eps: Диапазон допустимых значений ()

## Result Window

Кнопка Apply открывает новое окно с результатами проделанной работы. Данное окно включает поле с графиком (2D plot) и панель инструментов (toolbar), где пользователь может воспользоваться готовыми функциями, предложенными библиотекой Matplotlib, для навигации по графику и других операций. Окно сохраняется, и пользователь может анализировать результаты, закрывать окно или сворачивать его и продолжать работу. Для упрощения навигации по всем окнам с результатами (ResultWindows) они будут иметь уникальные названия, чтобы можно было легко идентифицировать нужное окно.

## 3.2. Пример использования

### 3.2.1. Получение текстовых данных

### 3.2.2. Создание Workspace

### 3.2.3. Выбор методов и параметров

Возможных комбинаций методов и параметров может быть большое количество, и достаточно сложно предугадать итоговый результат работы. Подробно рассмотрим как каждый элемент влияет на тестовых данных:

### 3.2.4. Текстовая обработка

Далее будет рассмотрено использование всех трех методов, то есть активен при обработке будет только один. Оценивать эффективность этих методов можно с помощью графиков (freq, length). Сравнивая графики исходных данных и графики после конкретного метода обработки.

## Alpha

Исходные графики, графики после. Визуально видны изменения в длине общей длины комментариев, все они стали немного меньше, более это заметно на графике freq где мы можем увидеть теперь более встречаемые слова.

## **stop-words**

Исходные графики, графики после. Визуально графики не должны были сильно измениться ведь шум от alpha не дает это сделать. Однако в комбинации с alpha обработкой можно заметить что слова которые оставались только после alpha обработки исчезли так как ("the "in "a") считаются наиболее встречающимися словами в английском языке. Однако stop-words включает в себя эти слова.

## **lemmatize**

При использовании только лемматизации сложно заметить какое-то весомое влияние этой функции. Более заметно оно станет только при использовании всех трех методов. Когда особо встречаемые смогут объединиться так можно заметить на примере (human , humans) объединились в одну группу.

### **3.2.5. Векторизация**

При дальнейшем рассмотрении работы этих методов, мы будем рассматривать их отдельно. Рассмотрим как работает каждый из этих методов на необработанных текстовых данных и обработанных, а так же рассмотрим влияние изменений параметров. Для анализа наших результатов нам потребуются полностью выполнить весь процесс анализа данных, чтобы визуализировать наши векторизованные данные, так как необходимо выбрать метод кластеризации будет использоваться метод K-means с 1 кластером.

## **TF-IDF**

При использовании на необработанных данных При использовании на обработанных данных Изменяем параметры min , max Замечаем типичные результаты с центром и то что рядом стоящие точки никак не связаны.

## **Word2Vec**

При использовании на необработанных данных При использовании на обработанных данных Изменяем параметры vector\_size , window, min\_count, sg Замечаем типичные результаты ( при увеличении вектора как будто все растягивается и становится дальше).

### **3.2.6. Кластеризация**

Далее рассмотрим методы кластеризации, рассмотрим их отдельно. Рассмотрим их на обработанных данных с зафиксированными параметрами векторизации word2vec.

## **k-means**

Рассмотрим как работает увеличение количество кластеров. И остальных

## **DBSCAN**

Рассмотрим как работает увеличение eps, min\_pts и попробуем подобрать значения основываясь на графике k-means

Заметим что сложно подобрать эти значения.

### **3.2.7. Интерпритация результатов = Итог**

## Глава 4

## Tytuł 4

Tekst.

### 4.1. Sekcja

Na Rysunku 4.1.



Рис. 4.1: Podpis (źródło: <http://jakis.adres>)





# Podsumowanie

Podsumujac analiza danych tekstowych jest bardzo chujna, jest bardzo przydatna dla analizy malych komentarzej.



# Литература

- [1] Autor, *Tytuł*, Wydawnictwo, Rok wydania.
- [2] Tytuł strony, *Tytuł artykułu*, <http://adres.strony> (dostęp:20.10.2023).
- [3] Impossibility Theorem for Clustering *Theorem for Clustering*, <https://www.cs.cornell.edu/home/kleinber/nips15.pdf>
- [4] A. K. Jain, R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc., 1988.
- [5] Daniel Jurafsky, James H. Martin, *Speech and Language Processing*, <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>, Stanford University., 2024.
- [6] C. C. Aggarwal, C. Zhai (Eds.), *Mining Text Data*, Springer, 2012.
- [7] A. Y. Ng, M. I. Jordan, Y. Weiss, *On Spectral Clustering: Analysis and an Algorithm*, Advances in Neural Information Processing Systems, 14, 2001.
- [8] YouTube Data API *YouTube Data API*, <https://developers.google.com/youtube/v3/docs?hl=pl>
- [9] Steven Bird, Ewan Klein, Edward Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.
- [10] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12, pp. 2825-2830, 2011.
- [11] Radim Řehůřek, Petr Sojka, *Software Framework for Topic Modelling with Large Corpora*, In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, 2010.
- [12] G. Salton, A. Wong, C. S. Yang, *A Vector Space Model for Automatic Indexing*, Communications of the ACM, 18(11), pp. 613-620, 1975.
- [13] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, arXiv preprint arXiv:1301.3781, 2013.
- [14] J. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, No. 14, pp. 281-297), 1967.
- [15] M. Ester, H. P. Kriegel, J. Sander, X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In KDD (Vol. 96, No. 34, pp. 226-231), 1996.
- [16] L. van der Maaten, G. Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research, 9(Nov), pp. 2579-2605, 2008.
- [17] Mark Summerfield, *Rapid GUI Programming with Python and Qt*, Prentice Hall, 2007.
- [18] John D. Hunter, *Matplotlib: A 2D Graphics Environment*, Computing in Science And Engineering, 9(3), pp. 90-95, 2007.
- [19] Matplotlib Development Team, *mplcursors*, <https://mplcursors.readthedocs.io/>, (доступ: 20.10.2023).



# Листинги



## Список таблиц





# Список иллюстраций

|  |    |
|--|----|
| 2.1. Przykład rys . . . . .  | 9  |
| 4.1. Podpis (źródło: <a href="http://jakis.adres">http://jakis.adres</a> ) . . . . . | 25 |