In [ ]:

```
import requests
from bs4 import BeautifulSoup
import lxml
import csv
import re
```

In [ ]:

```
def remove_tags(input_string):
    clean_string = re.sub(r'<.*?>', '', input_string)
    return clean_string
def remove_extra_spaces(input_string):
    return ' '.join(input_string.split())
```

In [ ]:

```
file = open('Content4.csv', mode='a', encoding='utf-8', newline='')
writer = csv.writer(file)
writer.writerow(['Article_name', 'Year', 'Content'])

url='https://www.jmir.org/'
for j in range(2004,2024):
    print(url+str(j))
    links = url+str(j)
    response = requests.get(url=links, stream=True)
    for i in range(20):
        # имя статьи
        soup4 = BeautifulSoup(response.text, 'lxml')
        soup4 = soup4.find_all('p', class_='h4 full-width-card-info-title')
        z =(soup4)[i]
        z = z.text # Здесь имя твое...
        #================
        # Берем HXML так-же из спискака
        soup3 = BeautifulSoup(response.text, 'lxml')
        soup3 = soup3.find_all('div', class_='full-width-card-info-group-buttons')
        name = soup3[i]
        a = str(name).rsplit('data-v-3802195d="" href=')[2]
        a = a.split('target',1)[0].strip().replace('"','')
        # Делаем запрос
        response2 = requests.get(a)
        soup2 = BeautifulSoup(response2.text, 'lxml')
        b = str(soup2)
        b = remove_tags(b)
        b = remove_extra_spaces(b)
        writer.writerow([z,j,b])
        print(f'Скаченная статья {j} года № {i}')
file.close()
```

In [1]:

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df = pd.read_csv('Content4.csv')
df.head()
```

Out[2]:

	Article_name	Year	Content
0	Using Claims Data to Examine Patients Using Pr...	2004	JMIR J Med Internet Res Journal of Medical Int...
1	Online Consumer Surveys as a Methodology for A...	2004	JMIR J Med Internet Res Journal of Medical Int...
2	A Multimedia Interactive Education System for ...	2004	JMIR J Med Internet Res Journal of Medical Int...
3	DietPal: A Web-Based Dietary Menu-Generating a...	2004	JMIR J Med Internet Res Journal of Medical Int...
4	Sex Differences in Youth-Reported Depressive S...	2004	JMIR J Med Internet Res Journal of Medical Int...

In [80]:

```
# list(df['Article_name'])
```

In [83]:

```
# Загрузка и предобработка данных

# corpus = df["Content"].tolist()
corpus = df[["Content",'Article_name']]
```

In [84]:

```

# Вычисление TF-IDF векторов
vectorizer = TfidfVectorizer(stop_words="english")
X = vectorizer.fit_transform(corpus)
```

In [105]:

```
corpus = df['Article_name'].str.lower()
corpus.head()
```

Out[105]:

```
0    using claims data to examine patients using pr...
1    online consumer surveys as a methodology for a...
2    a multimedia interactive education system for ...
3    dietpal: a web-based dietary menu-generating a...
4    sex differences in youth-reported depressive s...
Name: Article_name, dtype: object
```

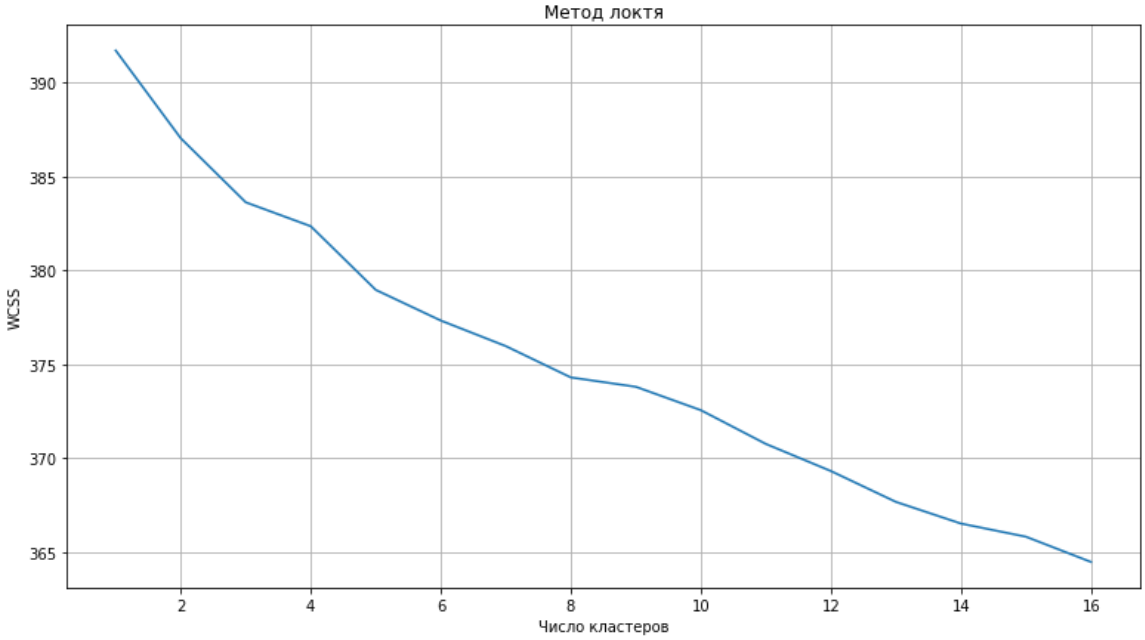
In [123]:

```
corpus = df['Article_name'].str.lower()
# Вычисление TF-IDF векторов
vectorizer = TfidfVectorizer(stop_words= m, analyzer='word', ngram_range=(1,1))
X = vectorizer.fit_transform(corpus)
# Задаем максимальное количество кластеров
max_clusters = 16
# Создаем список для значений WCSS
wcss = []

# Вычисляем WCSS для каждого числа кластеров от 1 до max_clusters
for i in range(1, max_clusters + 1):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

# Строим график зависимости WCSS от числа кластеров
plt.figure(figsize=(13, 7))
plt.plot(range(1, max_clusters + 1), wcss)

plt.title('Метод локтя')
plt.xlabel('Число кластеров')
plt.ylabel('WCSS')
plt.grid(True)
plt.show()
```



In [21]:

```
# Обучение KMeans модели
k = 8  # количество кластеров
kmeans = KMeans(n_clusters=k, random_state=42).fit(X)
# Вывод результатов кластеризации
labels = kmeans.labels_
df["cluster"] = labels
df.head(7)
```

Out[21]:

	Article_name	Year	Content	слова	cluster
0	Using Claims Data to Examine Patients Using Pr...	2004	JMIR J Med Internet Res Journal of Medical Int...	[using, claims, data, to, examine, patients, u...	1
1	Online Consumer Surveys as a Methodology for A...	2004	JMIR J Med Internet Res Journal of Medical Int...	[online, consumer, surveys, as, a, methodology...	0
2	A Multimedia Interactive Education System for ...	2004	JMIR J Med Internet Res Journal of Medical Int...	[a, multimedia, interactive, education, system...	4
3	DietPal: A Web-Based Dietary Menu-Generating a...	2004	JMIR J Med Internet Res Journal of Medical Int...	[dietpal:, a, web-based, dietary, menu-generat...	7
4	Sex Differences in Youth-Reported Depressive S...	2004	JMIR J Med Internet Res Journal of Medical Int...	[sex, differences, in, youth-reported, depress...	4
5	Can Clinical Trials Requiring Frequent Partici...	2004	JMIR J Med Internet Res Journal of Medical Int...	[can, clinical, trials, requiring, frequent, p...	5
6	Online Pediatric Information Seeking Among Mot...	2004	JMIR J Med Internet Res Journal of Medical Int...	[online, pediatric, information, seeking, amon...	4

In [30]:

```
1 df['Article_name'][df['cluster'] == 7].head(10).values.tolist()
```

Out[30]:

```
['DietPal: A Web-Based Dietary Menu-Generating and Management System',
 'Providing a Web-based Online Medical Record with Electronic Communicatio
n Capabilities to Patients With Congestive Heart Failure: Randomized Tria
l',
 'Usage and Longitudinal Effectiveness of a Web-Based Self-Help Cognitive
Behavioral Therapy Program for Panic Disorder',
 'Defining Participant Exposure Measures in Web-Based Health Behavior Chan
ge Programs',
 'Rates and Determinants of Repeated Participation in a Web-Based Behavior
Change Program for Healthy Body Weight and Healthy Lifestyle',
 'Effectiveness of a Web-Based Self-Help Intervention for Symptoms of Depr
ession, Anxiety, and Stress: Randomized Controlled Trial',
 'Randomized Controlled Trial of an Internet-Based Versus Face-to-Face Dys
pnea Self-Management Program for Patients With Chronic Obstructive Pulmona
ry Disease: Pilot Study',
 'Therapist-Assisted, Internet-Based Treatment for Panic Disorder: Can Gen
eral Practitioners Achieve Comparable Patient Outcomes to Psychologists?',
 'Integrating an eHealth Program for Pregnant Women in Midwifery Care: A F
easibility Study Among Midwives and Program Users',
 'Comparison of Trial Participants and Open Access Users of a Web-Based Ph
ysical Activity Intervention Regarding Adherence, Attrition, and Repeated
Participation']
```