GROUP

WEBSITE CATEGORIZATION

CAPSTONE PROJECT PRESENTATION

OUTLINE

I. INTRODUCTION

II. DATA SCIENCE METHODOLOGY

1. DATA CRAWLING
2. DATA PREPROCESSING
3. DATA VISUALIZATION
4. MODEL TRAINING

III. RESULT

1. FINAL RESULT
2. EVALUATE
3. FUTURE IMPROVEMENT

I. ABOUT OUR PROJECT

CATEGORIZATION CAN BE DEFINED AS “THE PROCESS BY WHICH INDIVIDUALS GROUP OBJECTS OR EVENTS INTO CATEGORIES”.

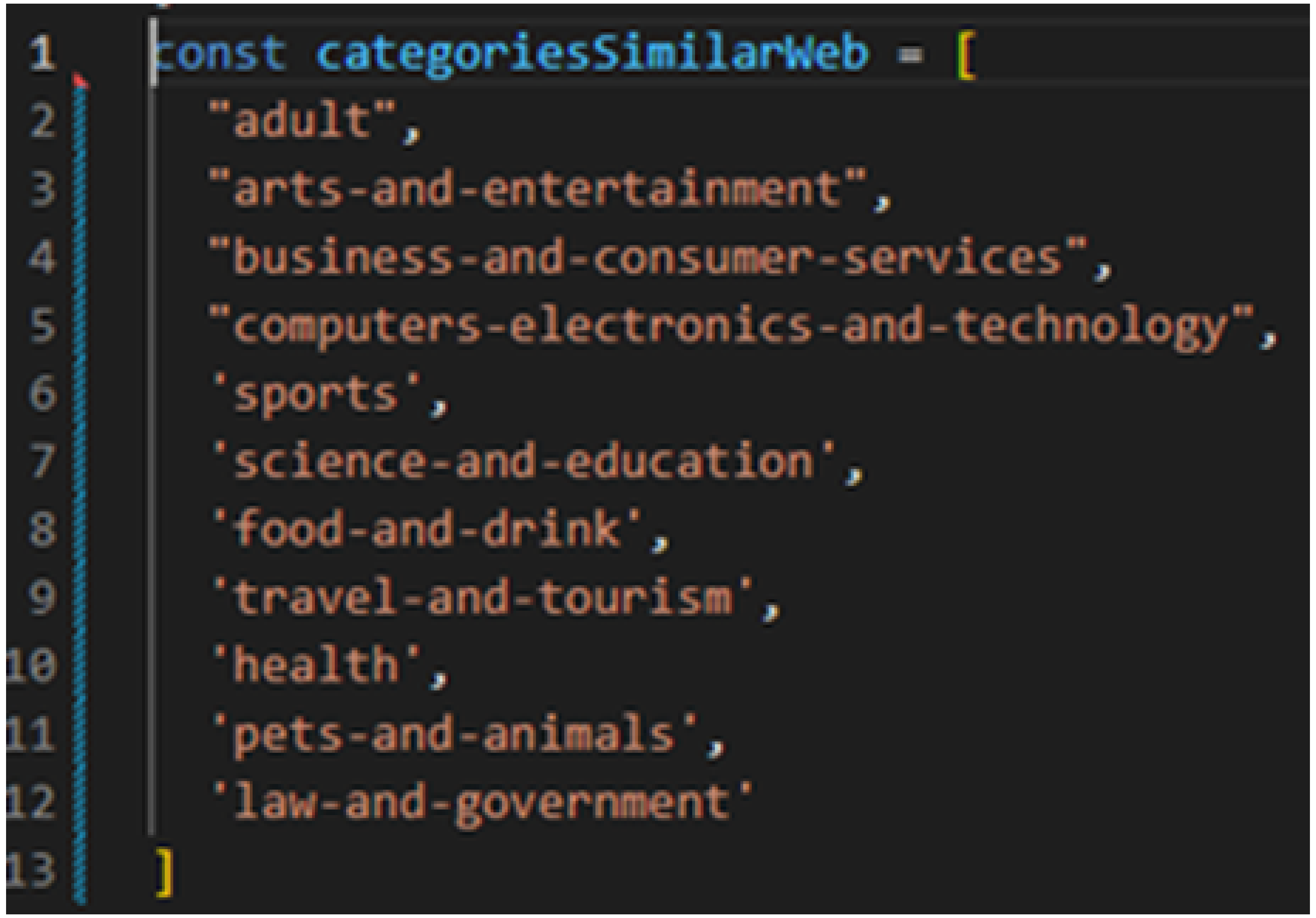
WEBSITE CATEGORIZATION:

- **INPUT** INFORMATION IS THE CONTENT OF WEBSITES IN TEXT FORM.
- **OUTPUT** LABEL FOR EACH CONTENT

II. DATA SCIENCE METHODOLOGY

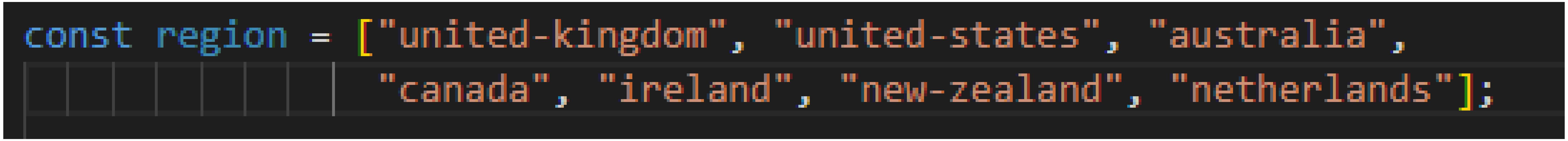
1. DATA CRAWLING

- Source list of websites: https://www.similarweb.com
- **11 categories**
- Average **50–60 websites** per topic



```
const categoriesSimilarWeb = [
  "adult",
  "arts-and-entertainment",
  "business-and-consumer-services",
  "computers-electronics-and-technology",
  'sports',
  'science-and-education',
  'food-and-drink',
  'travel-and-tourism',
  'health',
  'pets-and-animals',
  'law-and-government'
]
```

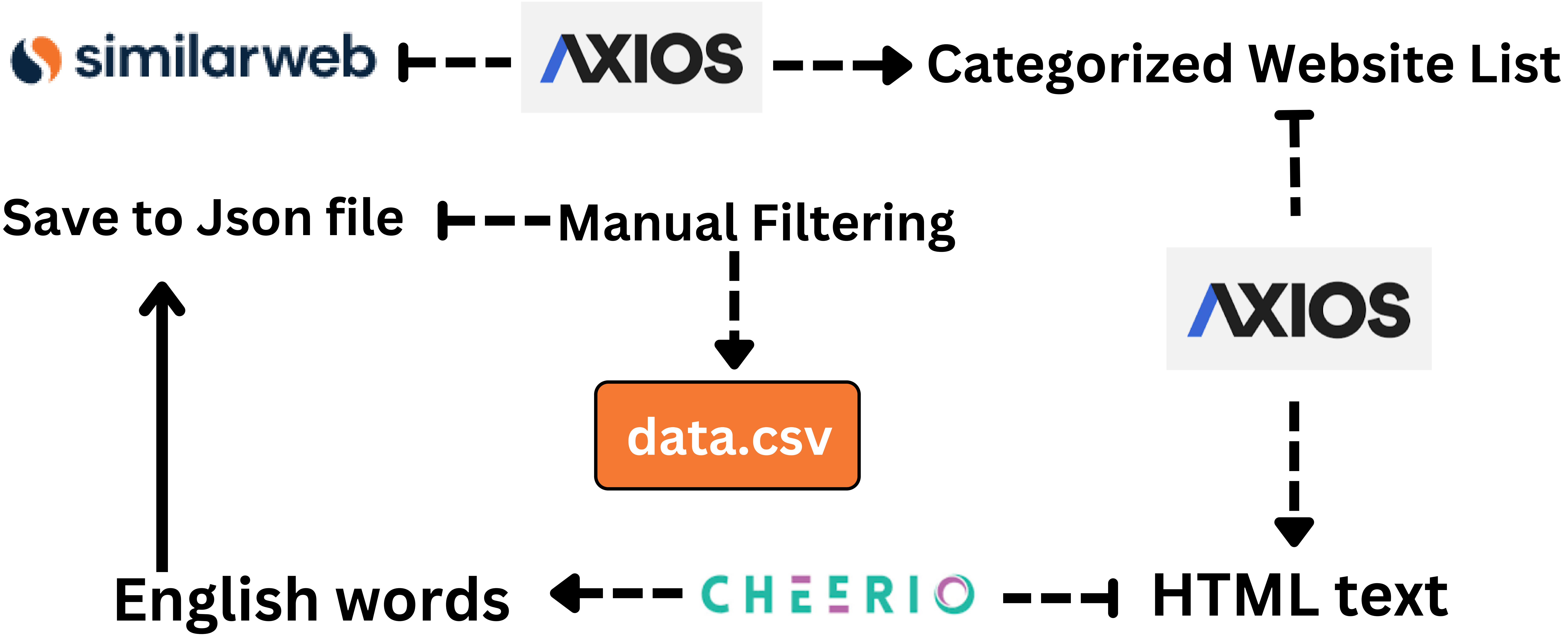
We only select websites from nations where English is widely spoken:



```
const region = ["united-kingdom", "united-states", "australia",
                "canada", "ireland", "new-zealand", "netherlands"];
```

II. DATA SCIENCE METHODOLOGY

1. DATA CRAWLING



II. DATA SCIENCE METHODOLOGY

1. DATA CRAWLING

Note:

- Only use **Javascript library** to crawl website for convenience
- Only crawl **landing page** or **homepage** of each website
- We **don't handle Javascript** implementation and only focus on **SSR website**, which has plenty of HTML text (> 1000 English words)
- Remove noise: Manual filtering to **remove unreadable and non-English** content.

2. DATA PROCESSING

*LOWERING ALL THE TEXT

*REMOVE ALL HYPERLINK , URL IN ALL DOCUMENTS

*REMOVE ALL PUNCTUATION SUCH AS “;”, “?” ,“!” AND OTHER UNNECESSARY CHARACTER OR SYMBOL

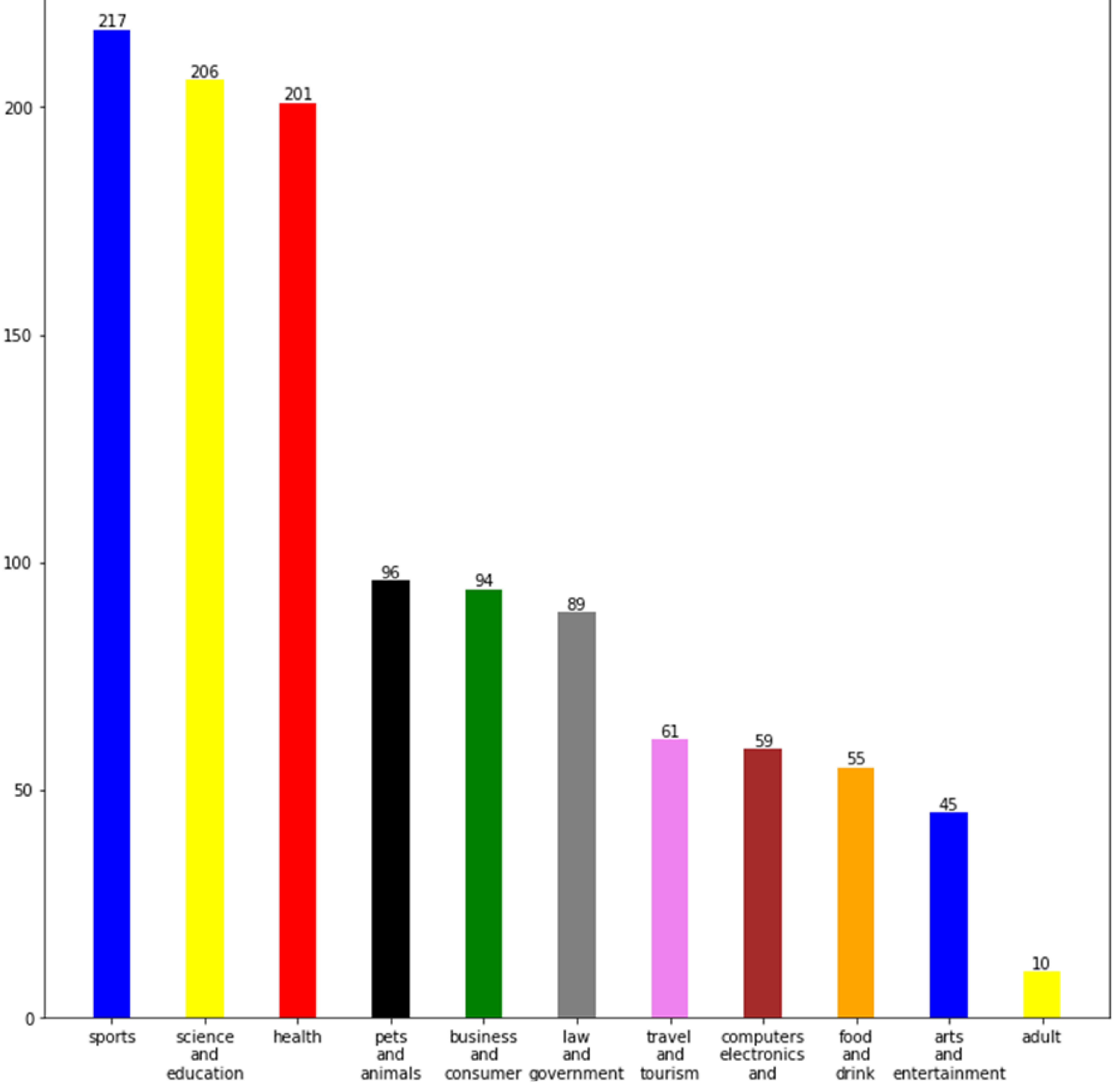
*REMOVE ALL HTML TAG

*REMOVE ALL STOP WORD SUCH AS A , AN ,THE ,....

*USING SNOWBALLSTEMMER TO STEM THE ALL THE WORD

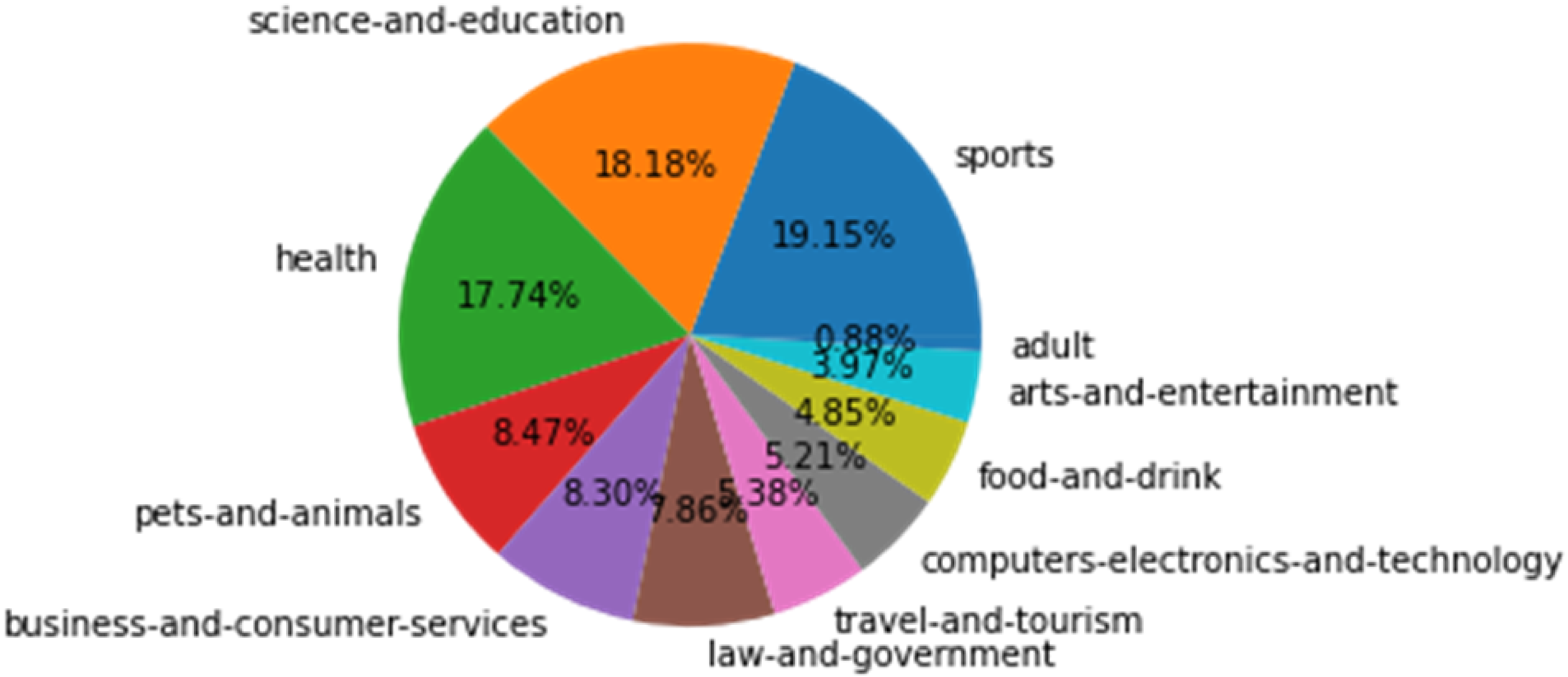
3. DATA VISUALIZATION

The bar chart illustrates the amount of crawled websites.

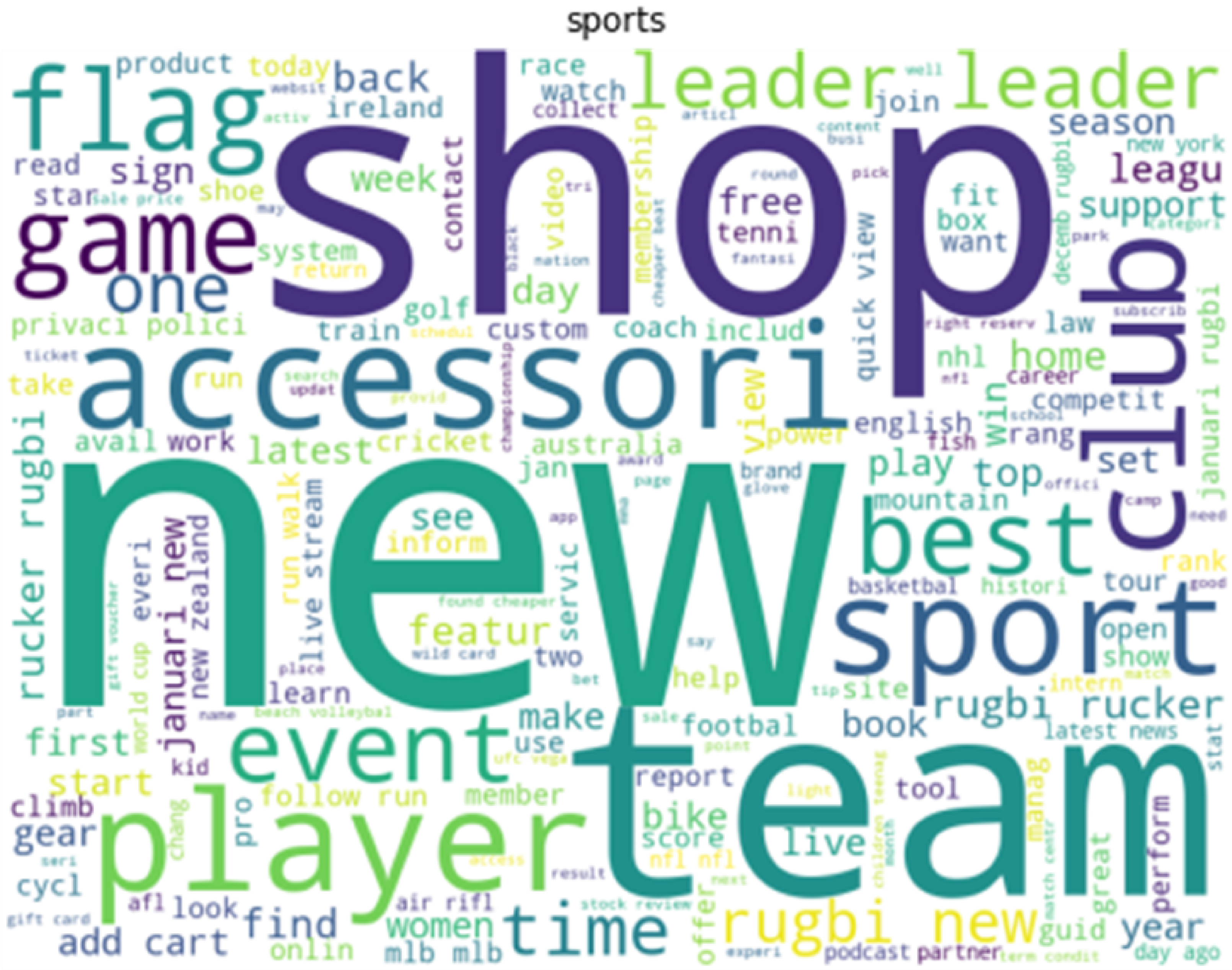


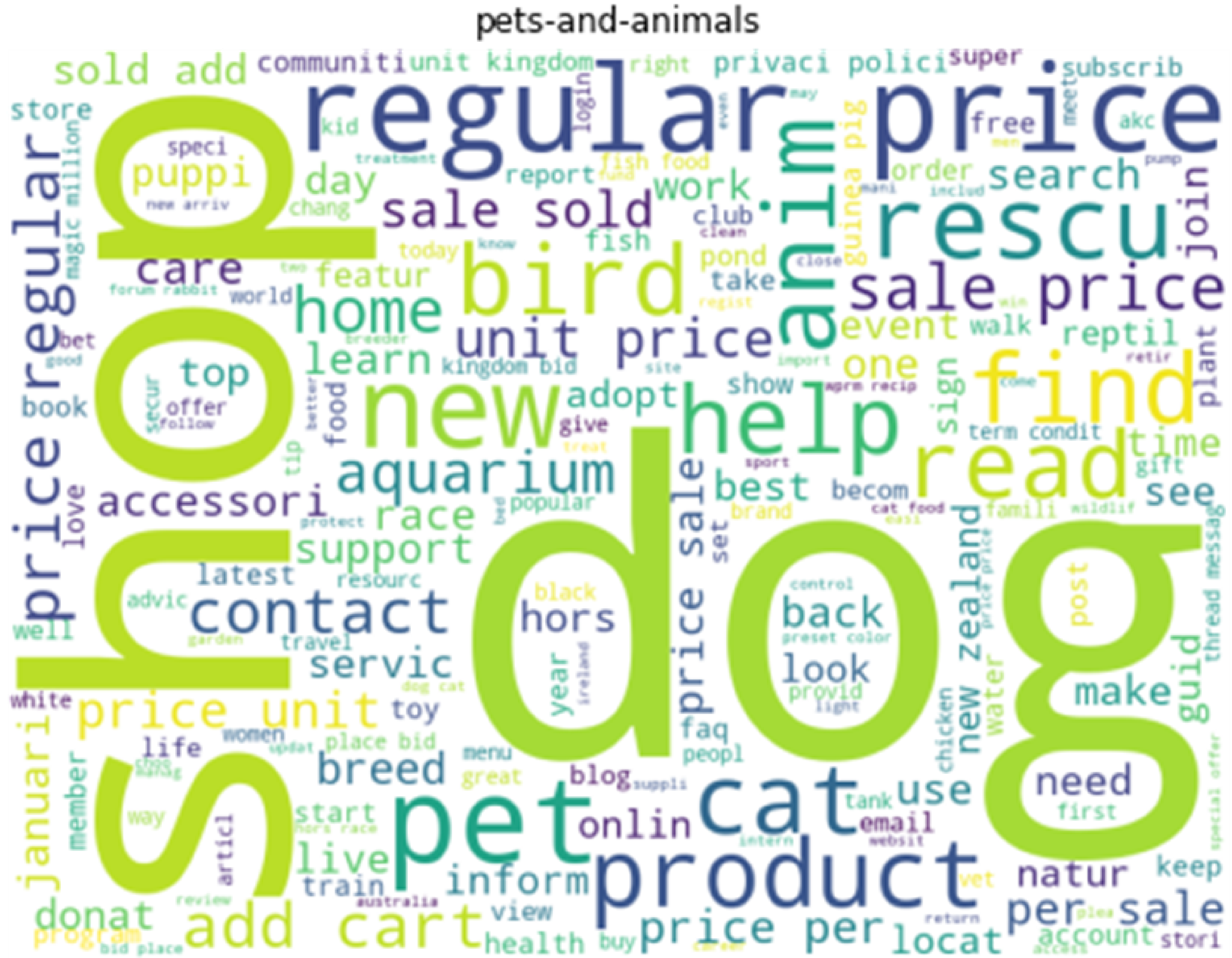
3. DATA VISUALIZATION

The pie chart representing the percentage of the number of websites in each topic

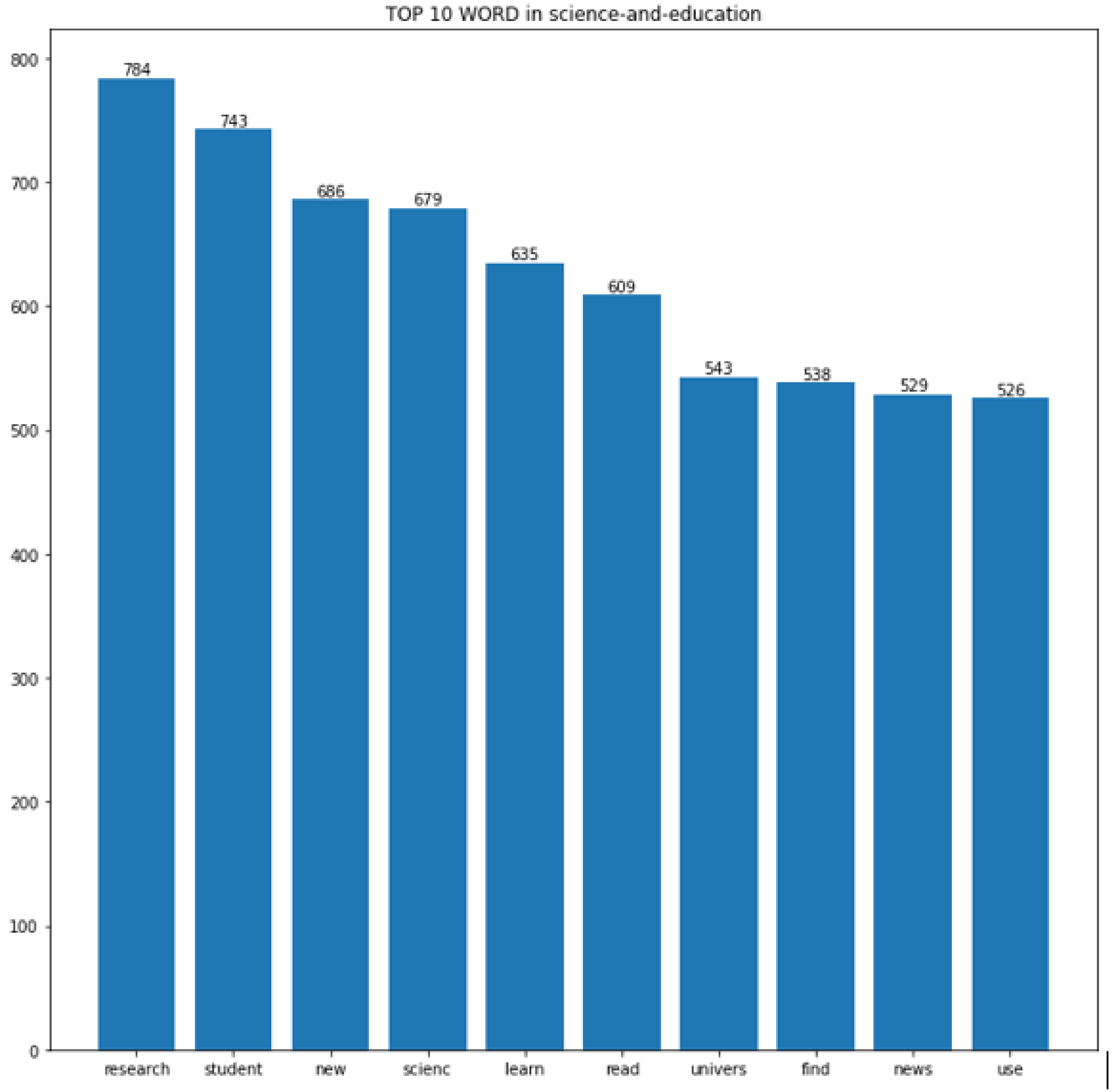


3. DATA VISUALIZATION


sports


pets-and-animals

3. DATA VISUALIZATION



4. FEATURE EXTRACTION

TF(TERM FREQUENCY):

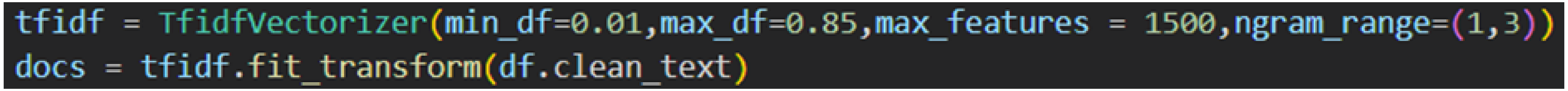
$$\text{TF} = \frac{\text{Number of times word appear in a document}}{\text{Total number of word in that document}}$$

IDF(INVERSE DOCUMENT FREQUENCY)

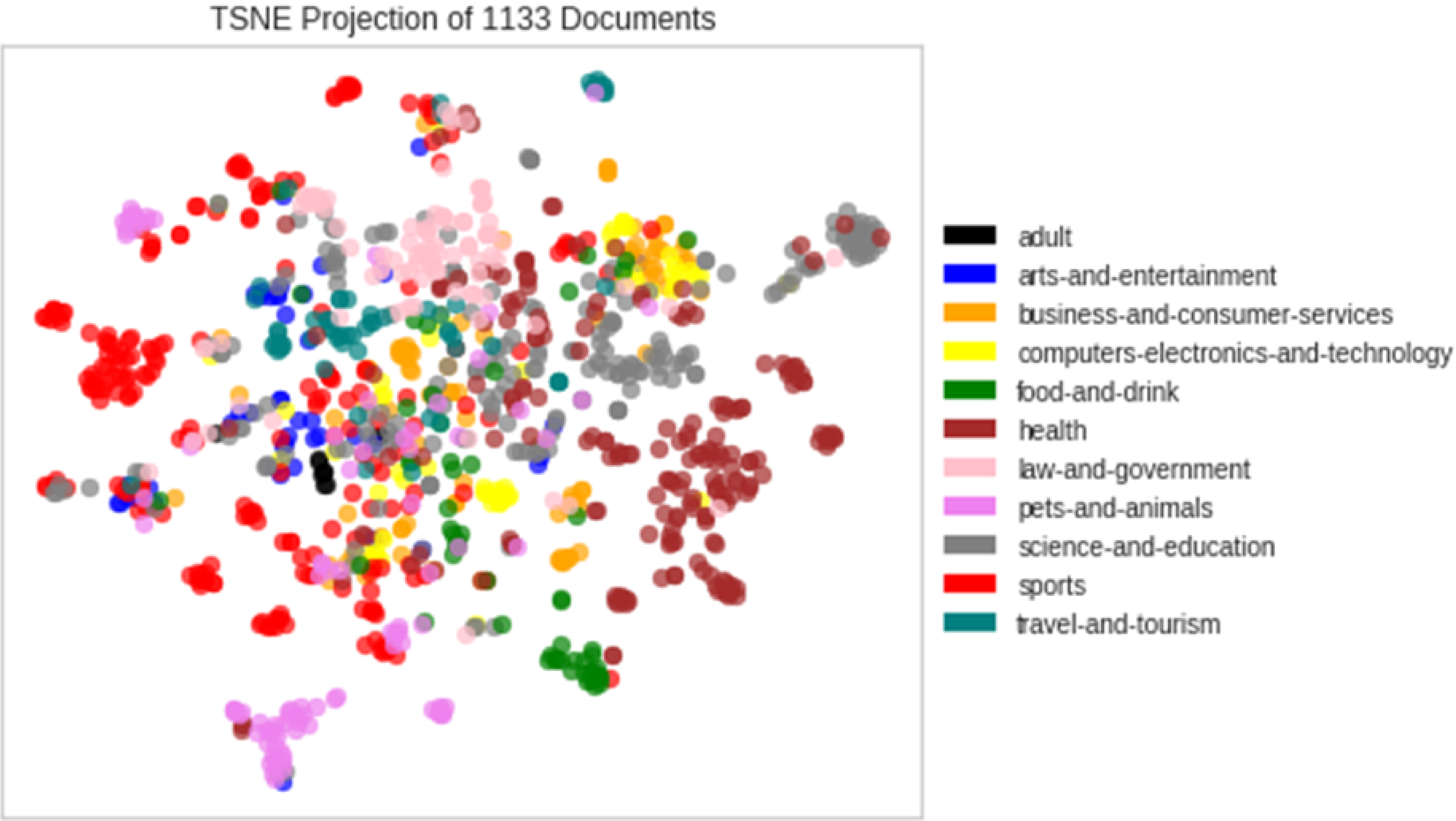
$$\text{IDF} = \log(\frac{\text{Total number of documents}}{\text{Document that contain word W}})$$

=> TF-IDF SCORE= TF*IDF

TF-IDF VECTORIZER:



```
tfidf = TfidfVectorizer(min_df=0.01,max_df=0.85,max_features = 1500,ngram_range=(1,3))
docs = tfidf.fit_transform(df.clean_text)
```


TSNE Projection of 1133 Documents
adult
arts-and-entertainment
business-and-consumer-services
computers-electronics-and-technology
food-and-drink
health
law-and-government
pets-and-animals
science-and-education
sports
travel-and-tourism

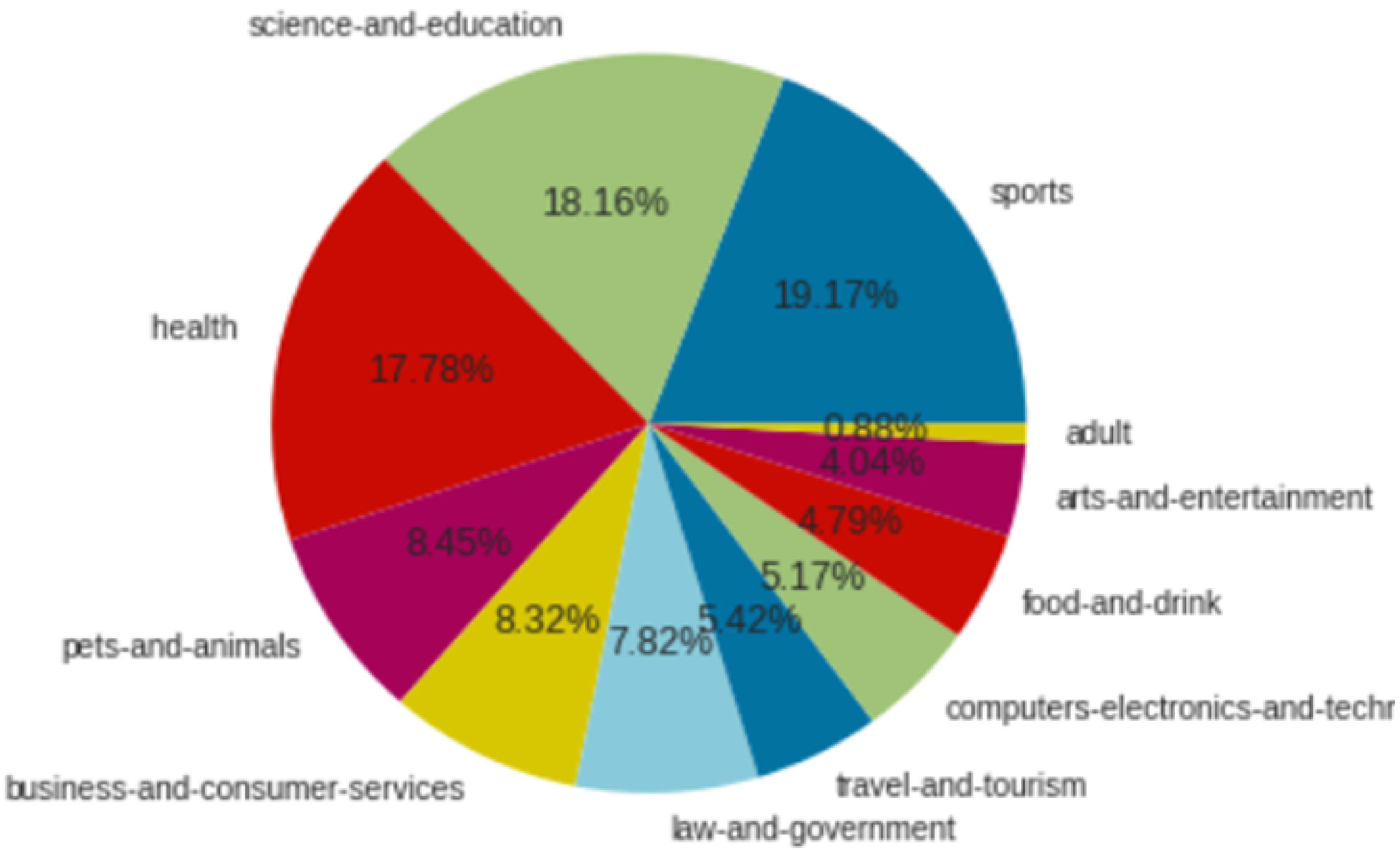
4. MODEL TRAINING

Divide the data set into 2 subsets:

- Train set: 70%
- Test set: 30 %

4. MODEL TRAINING

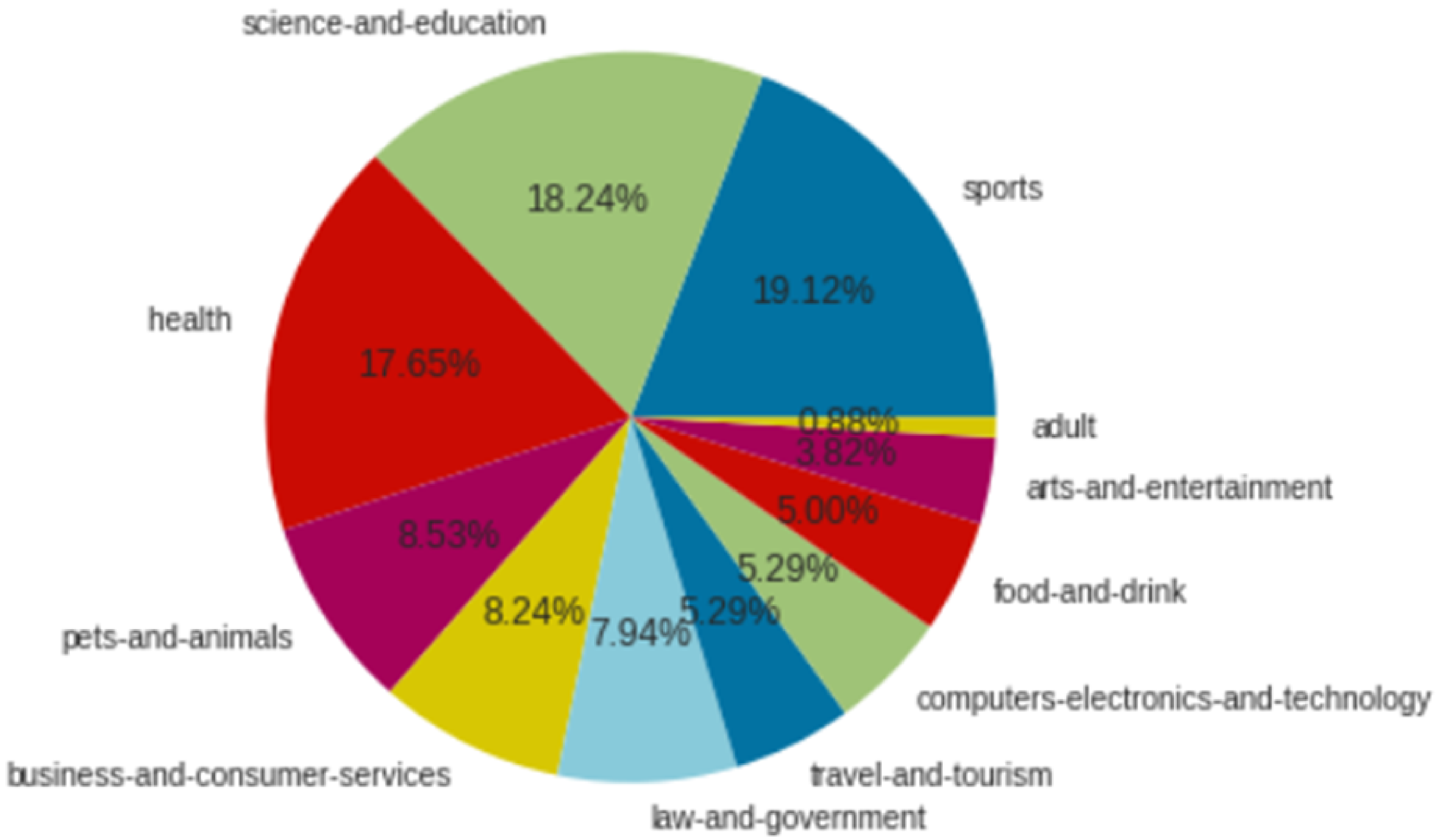
The percentage of websites crawled by topic in training set and test set:



Training set(793 rows)

4. MODEL TRAINING

The percentage of websites crawled by topic in training set and test set:



Test set(340 Rows)

4. MODEL TRAINING

NAIVE BAYES:

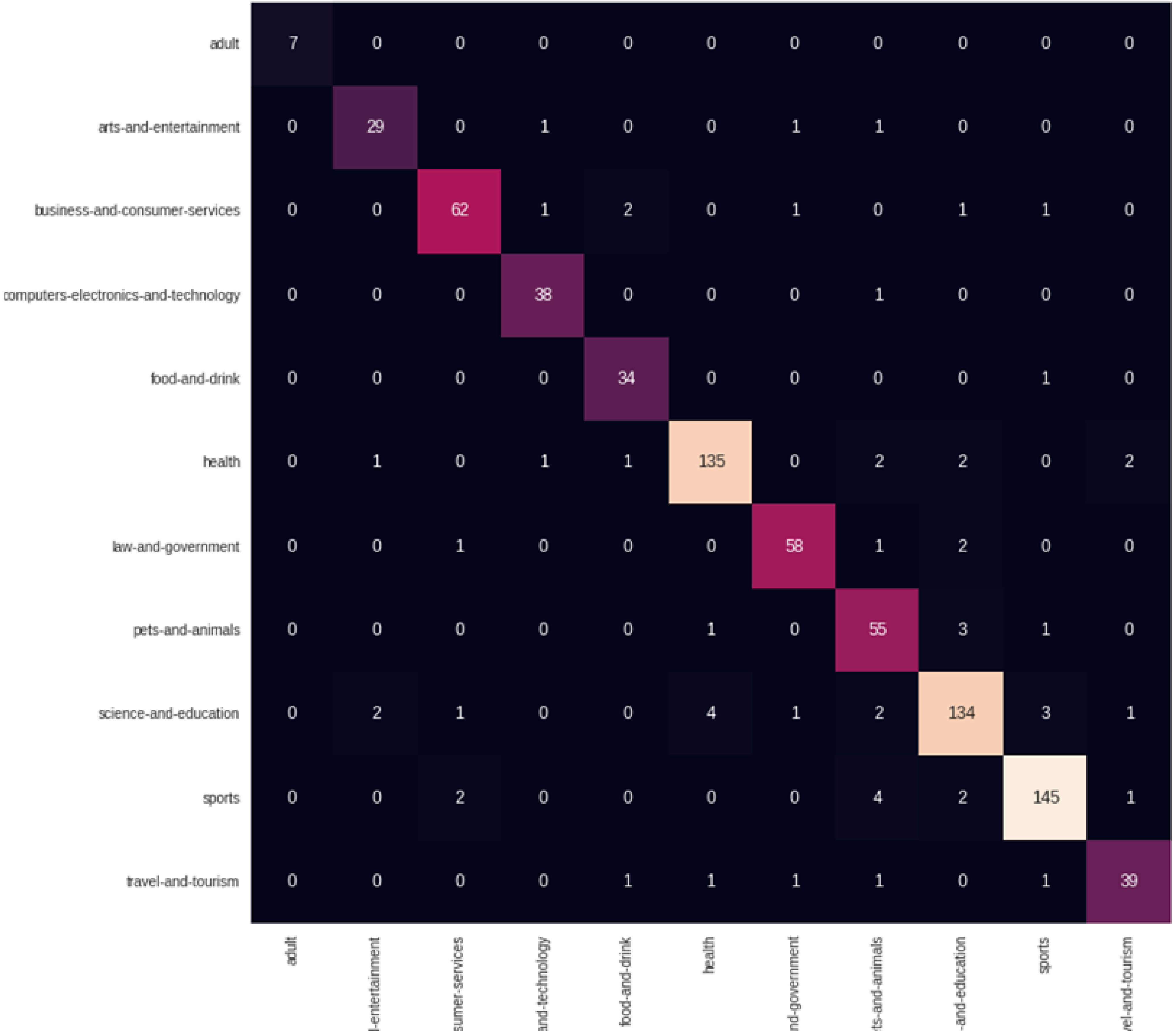
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

THE ATTRIBUTES ARE CONDITIONALLY INDEPENDENT GIVEN CLASSIFICATION

$$c_{MAP} = \underset{c_i \in C}{arg\ max}\, P(c_i).\prod_{j=1}^{n} P\left(z_j|c_i\right)$$

II. RESULT

1. FINAL RESULT



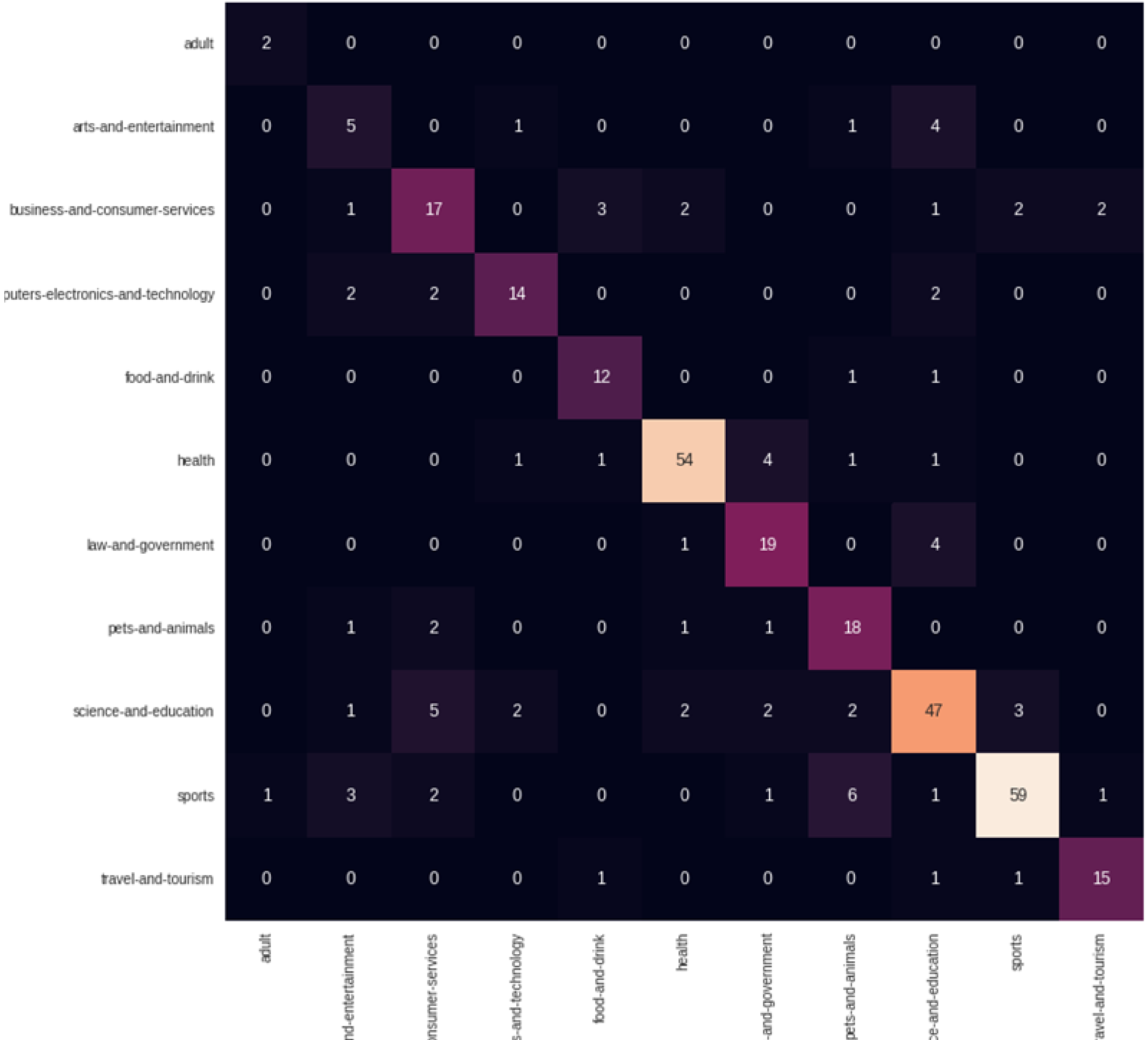
	adult	l-entertainment	sumer-services	and-technology	food-and-drink	health	nd-government	ts-and-animals	-and-education	sports	vel-and-tourism
adult	7	0	0	0	0	0	0	0	0	0	0
arts-and-entertainment	0	29	0	1	0	0	1	1	0	0	0
business-and-consumer-services	0	0	62	1	2	0	1	0	1	1	0
computers-electronics-and-technology	0	0	0	38	0	0	0	1	0	0	0
food-and-drink	0	0	0	0	34	0	0	0	0	1	0
health	0	1	0	1	1	135	0	2	2	0	2
law-and-government	0	0	1	0	0	0	58	1	2	0	0
pets-and-animals	0	0	0	0	0	1	0	55	3	1	0
science-and-education	0	2	1	0	0	4	1	2	134	3	1
sports	0	0	2	0	0	0	0	4	2	145	1
travel-and-tourism	0	0	0	0	1	1	1	1	0	1	39

***TRAINING:**

-ACCURACY:93%

II. RESULT

1. FINAL RESULT



	adult	nd-entertainment	onsumer-services	s-and-technology	food-and-drink	health	-and-government	pets-and-animals	ce-and-education	sports	avel-and-tourism
adult	2	0	0	0	0	0	0	0	0	0	0
arts-and-entertainment	0	5	0	1	0	0	0	1	4	0	0
business-and-consumer-services	0	1	17	0	3	2	0	0	1	2	2
puters-electronics-and-technology	0	2	2	14	0	0	0	0	2	0	0
food-and-drink	0	0	0	0	12	0	0	1	1	0	0
health	0	0	0	1	1	54	4	1	1	0	0
law-and-government	0	0	0	0	0	1	19	0	4	0	0
pets-and-animals	0	1	2	0	0	1	1	18	0	0	0
science-and-education	0	1	5	2	0	2	2	2	47	3	0
sports	1	3	2	0	0	0	1	6	1	59	1
travel-and-tourism	0	0	0	0	1	0	0	0	1	1	15

***TEST**

ACCURACY:77%

II. RESULT

2. DIFFICULTY

*CRAWL WEB

*MODEL ACCURACY

II. RESULT

3. FUTURE IMPROVEMENT

*FIND AND CRAWL MORE WEB TO INCREASE DATASET

*HANDLE MULTI-LABLE PROBLEM

*TRY ANOTHER MODEL

THANKS FOR LISTENING
OUR PRESENTATION