Data analysis & exploration

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from sklearn.decomposition import LatentDirichletAllocation
```

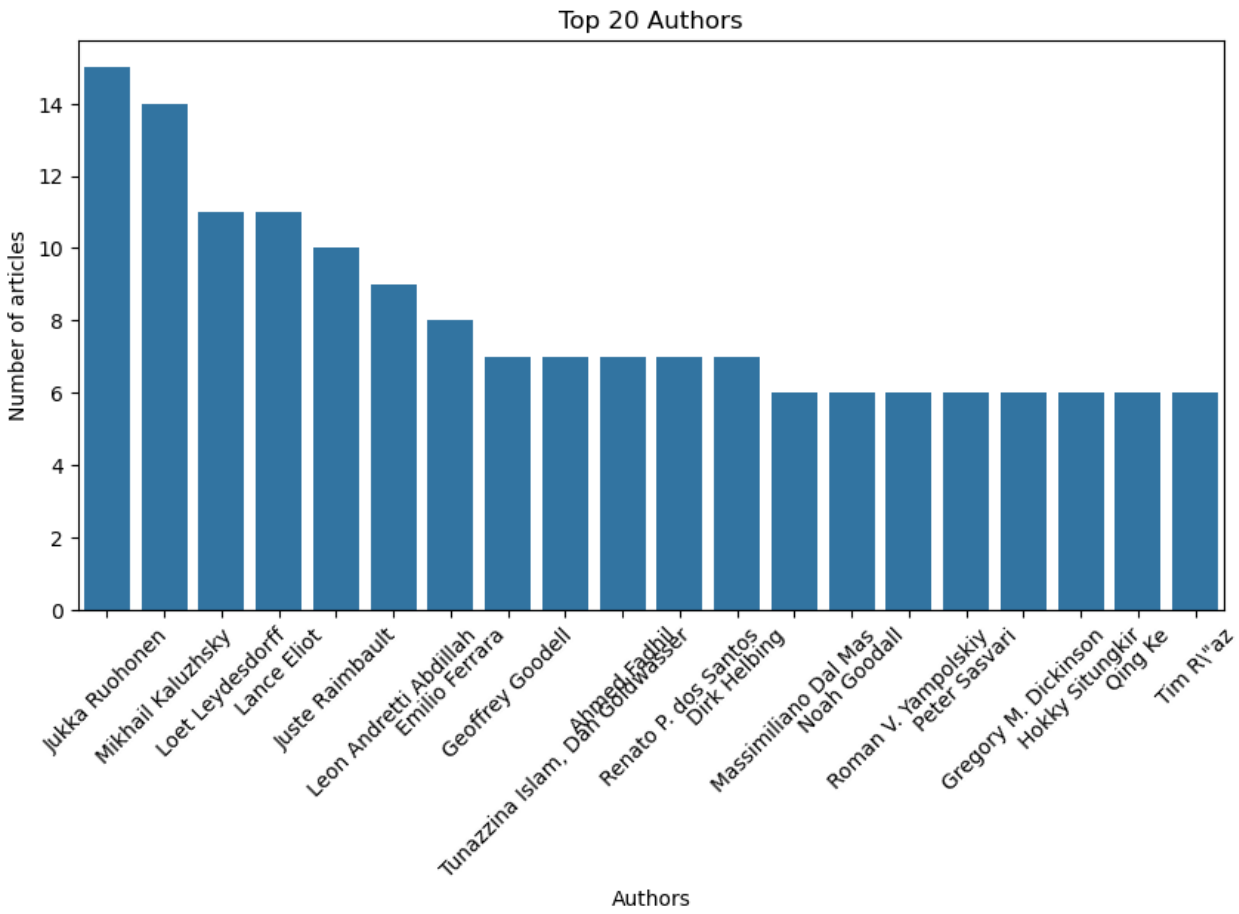
In [5]:
```
df = pd.read_csv('arxiv_cs_CY_articles.csv')
print(df.head())
print(df.info())
print(df.describe(include='all'))
```

```
         id           submitter  \
0  0704.1158   Bernardo Huberman
1  0704.1675     Kristina Lerman
2  0704.1676     Kristina Lerman
3  0704.3316        Ciro Cattuto
4  0704.3647  Catherine Marshall

                                             authors  \
0                  Fang Wu and Bernardo A. Huberman
1           Anon Plangprasopchok and Kristina Lerman
2  Kristina Lerman, Anon Plangprasopchok and Chio...
3  Ciro Cattuto, Andrea Baldassarri, Vito D. P. S...
4  Catherine C. Marshall, Frank McCown, and Micha...

                                               title  \
0                  Novelty and Collective Attention
1  Exploiting Social Annotation for Automatic Res...
2       Personalizing Image Search Results on Flickr
3  Vocabulary growth in collaborative tagging sys...
4  Evaluating Personal Archiving Strategies for I...

                                            comments journal-ref  \
0                                                NaN         NaN
1  6 pages, submitted to AAAI07 workshop on Infor...         NaN
2  12 pages, submitted to AAAI07 workshop on Inte...         NaN
3                                6 pages, 7 figures         NaN
4  6 pages, 2 tables, to be published in the Proc...         NaN

                       doi report-no  \
0  10.1073/pnas.0704916104       NaN
1                      NaN       NaN
2                      NaN       NaN
3                      NaN       NaN
4                      NaN       NaN

                                    categories license  \
0                    cs.CY cs.IR physics.soc-ph     NaN
1                              cs.AI cs.CY cs.DL     NaN
2                  cs.IR cs.AI cs.CY cs.DL cs.HC     NaN
```

```
3  cs.IR cond-mat.stat-mech cs.CY physics.data-an     NaN
4                              cs.DL cs.CY cs.HC     NaN

                                             abstract  \
0    The subject of collective attention is centr...
1    Information integration applications, such a...
2    The social media site Flickr allows users to...
3    We analyze a large-scale snapshot of del.ici...
4    Internet-based personal digital belongings p...

                                             versions update_date  \
0  [{'version': 'v1', 'created': 'Mon, 9 Apr 2007...  2009-11-13
1  [{'version': 'v1', 'created': 'Thu, 12 Apr 200...  2016-09-08
2  [{'version': 'v1', 'created': 'Thu, 12 Apr 200...  2007-05-23
3  [{'version': 'v1', 'created': 'Wed, 25 Apr 200...  2007-05-23
4  [{'version': 'v1', 'created': 'Fri, 27 Apr 200...  2007-05-23

                                      authors_parsed
0  [['Wu', 'Fang', ''], ['Huberman', 'Bernardo A....
1  [['Plangprasopchok', 'Anon', ''], ['Lerman', '...
2  [['Lerman', 'Kristina', ''], ['Plangprasopchok...
3  [['Cattuto', 'Ciro', ''], ['Baldassarri', 'And...
4  [['Marshall', 'Catherine C.', ''], ['McCown', ...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19876 entries, 0 to 19875
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              19876 non-null  object
 1   submitter       19876 non-null  object
 2   authors         19876 non-null  object
 3   title           19876 non-null  object
 4   comments        12309 non-null  object
 5   journal-ref     3606 non-null   object
 6   doi             4360 non-null   object
 7   report-no       550 non-null    object
 8   categories      19876 non-null  object
 9   license         19618 non-null  object
 10  abstract        19876 non-null  object
 11  versions        19876 non-null  object
 12  update_date     19876 non-null  object
 13  authors_parsed  19876 non-null  object
dtypes: object(14)
memory usage: 2.1+ MB
None
               id submitter         authors  \
count       19876     19876           19876
unique      19876     12292           18824
top     0704.1158     EPTCS  Jukka Ruohonen
freq            1        55              15

                                                 title  \
count                                            19876
unique                                           19872
```

```
top     Improving International Climate Policy via Mut...
freq                                                    2

                                                   comments  \
count                                                 12309
unique                                                 9756
top     ISBN# 978-0-646-95337-3 Presented at the Austr...
freq                                                     82

                                                journal-ref  \
count                                                  3606
unique                                                 3556
top     2022 ACM Conference on Fairness, Accountabilit...
freq                                                      4

                           doi       report-no categories  \
count                     4360             550      19876
unique                    4352             532       2287
top     10.1145/3448139.3448188  ISSN 1947 5500      cs.CY
freq                         2               7       5176

                                                  license  \
count                                               19618
unique                                                  9
top     http://arxiv.org/licenses/nonexclusive-distrib...
freq                                                10924

                                                abstract  \
count                                              19876
unique                                             19874
top       Archival information systems in government a...
freq                                                   2

                                                versions update_date  \
count                                              19876       19876
unique                                             19873        3201
top     [{'version': 'v1', 'created': 'Wed, 5 Oct 2022...  2007-05-23
freq                                                   2         178

                   authors_parsed
count                       19876
unique                      18678
top     [['Ruohonen', 'Jukka', '']]
freq                           15
```

In [8]:
```
author_counts = df['authors'].value_counts().head(0)

plt.figure(figsize=(10, 5))
sns.barplot(x=author_counts.index, y=author_counts.values)
plt.xticks(rotation=45)
plt.title('Top 20 Authors')
plt.xlabel('Authors')
plt.ylabel('Number of articles')
plt.show()
```



In [13]:
```
top_categories = df['categories'].value_counts().nlargest(10)
top_categories_df = top_categories.reset_index()
top_categories_df.columns = ['Category', 'Number of articles']
print(top_categories_df)
```

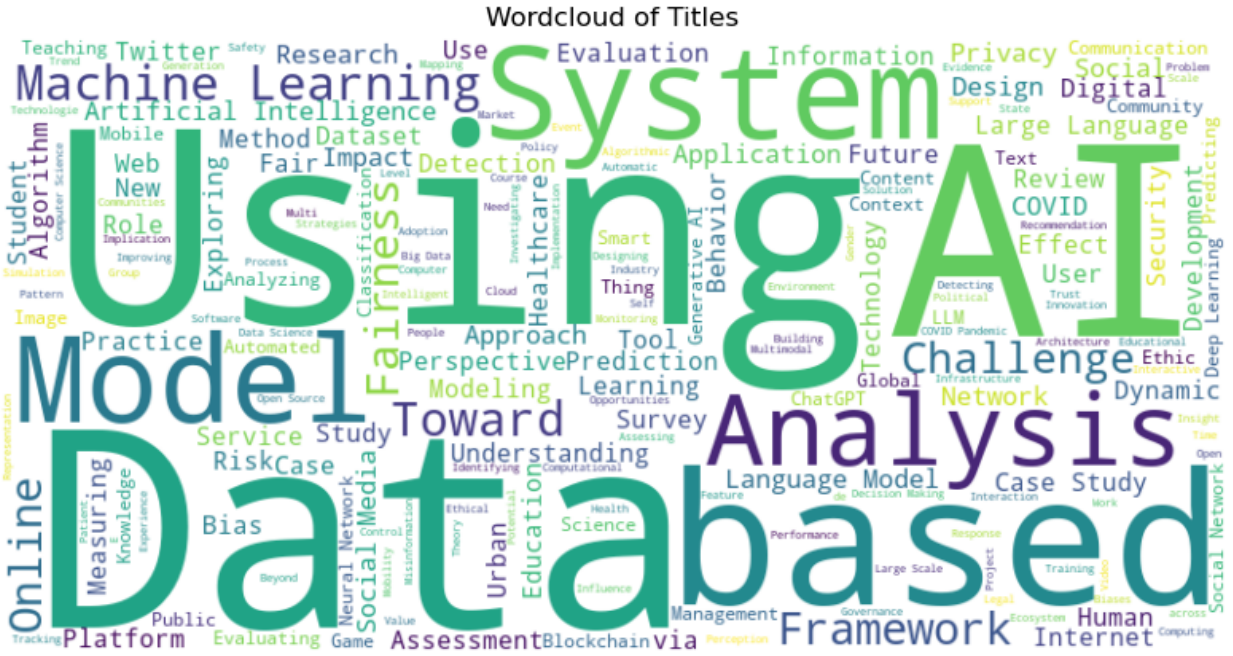
```
            Category  Number of articles
0              cs.CY                5176
1        cs.CY cs.AI                 791
2        cs.SI cs.CY                 535
3        cs.LG cs.CY                 522
4        cs.HC cs.CY                 515
5        cs.CR cs.CY                 498
6        cs.CY cs.HC                 439
7        cs.CL cs.CY                 371
8        cs.CY cs.SI                 335
9  cs.LG cs.AI cs.CY                 321
```

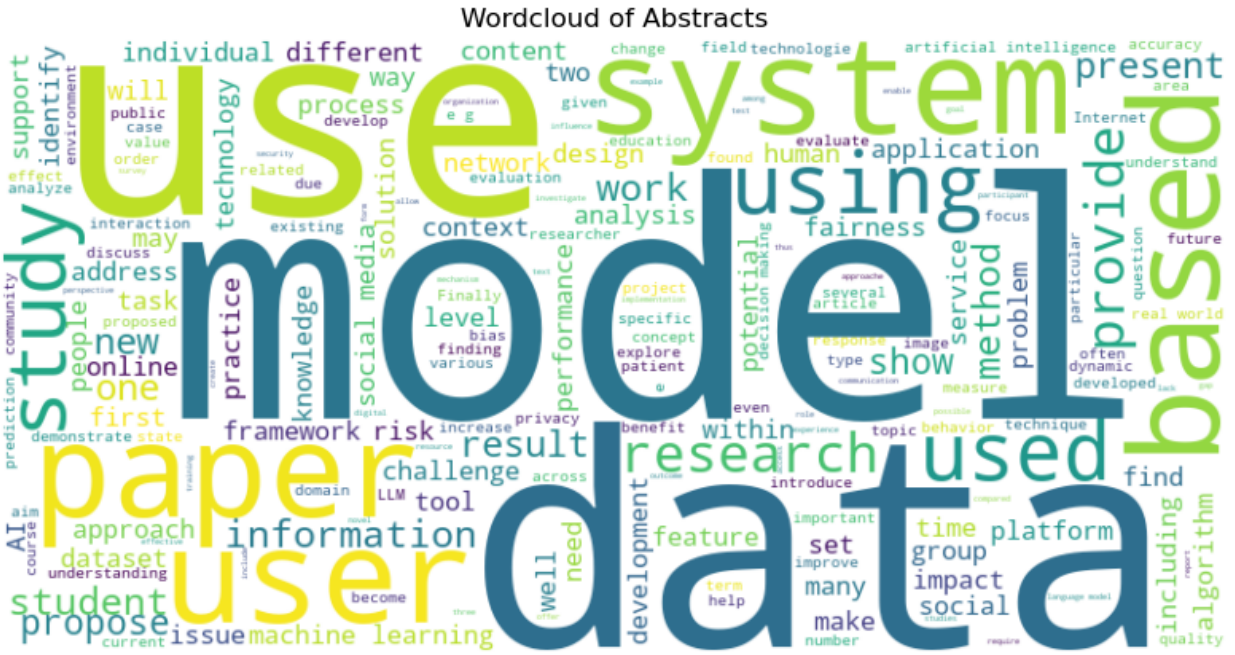
Creating wordclouds for title and description

In [18]:
```
def generate_wordcloud(df, column, title):
    text = ' '.join(df[column].dropna())
    wordcloud = WordCloud(width=800, height=400, background_color='white'

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
```

```
    plt.title(f'Wordcloud of {title}')
    plt.show()
```

In [19]:
```
generate_wordcloud(df, 'title', 'Titles')
generate_wordcloud(df, 'abstract', 'Abstracts')
```



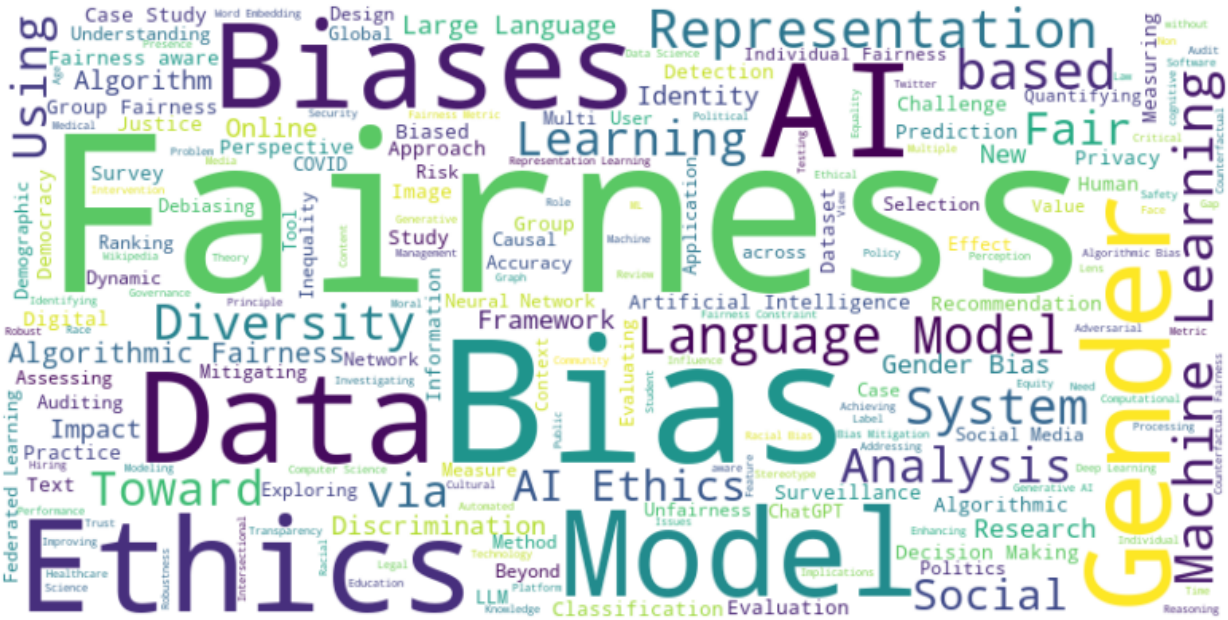


Create wordclouds by filtering only articles with some words about social issues, as the general wordcloud included a lot of words about computer science

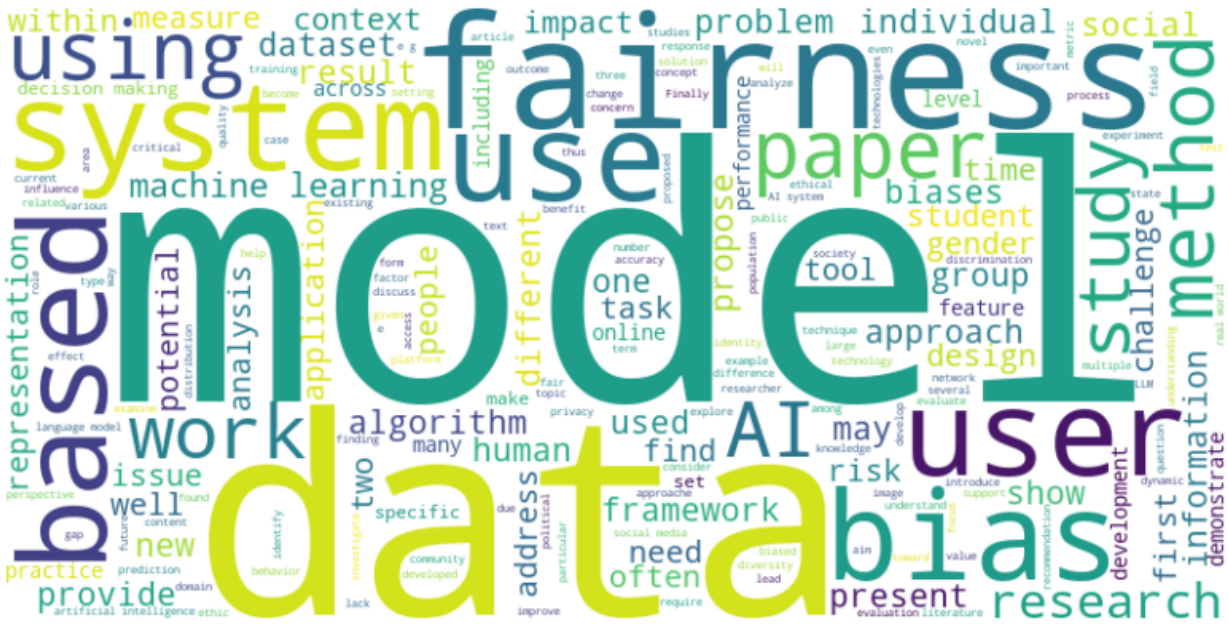
In [21]:
```
social_keywords = ["justice", "equality", "bias", "fairness", "discrimina
                   "racism", "critical theory", "social justice", "inequa
                   "feminism", "gender", "diversity", "human rights", "cu
                   "intersectionality", "activism", "identity", "democrac
                   "oppression", "ethics", "surveillance", "capitalism",
```

```
def filter_social_issues(df, keywords, parameter):
    return df[df[parameter].str.contains('|'.join(keywords), case=False,


social_df_title = filter_social_issues(df, social_keywords, 'title')
social_df_abstract = filter_social_issues(df, social_keywords, 'abstract'
generate_wordcloud(social_df_title, 'title', 'Titles')
generate_wordcloud(social_df_abstract, 'abstract', 'Abstracts')
```

Wordcloud of Titles



Wordcloud of Abstracts



In [29]:
```
from sklearn.feature_extraction.text import CountVectorizer

def most_common_words(df, column, n=20):
    vectorizer = CountVectorizer(stop_words='english')
    word_matrix = vectorizer.fit_transform(df[column].dropna())
```

```
    word_counts = pd.DataFrame(word_matrix.sum(axis=0), columns=vectorize
    word_counts.columns = ['counts']
    return word_counts.sort_values(by='counts', ascending=False).head(n)

most_common_titles = most_common_words(df, 'title')
print('Most common words on titles\n')
print(most_common_titles)
print('\n')

most_common_abstracts = most_common_words(df, 'abstract')
print('Most common words on abstracts\n')
print(most_common_abstracts)
```

Most common words on titles

	counts
learning	1994
data	1961
ai	1646
based	1418
using	1360
social	1284
analysis	1018
study	904
models	885
fairness	822
machine	788
systems	754
online	718
language	715
model	664
covid	660
19	631
case	625
privacy	622
approach	597

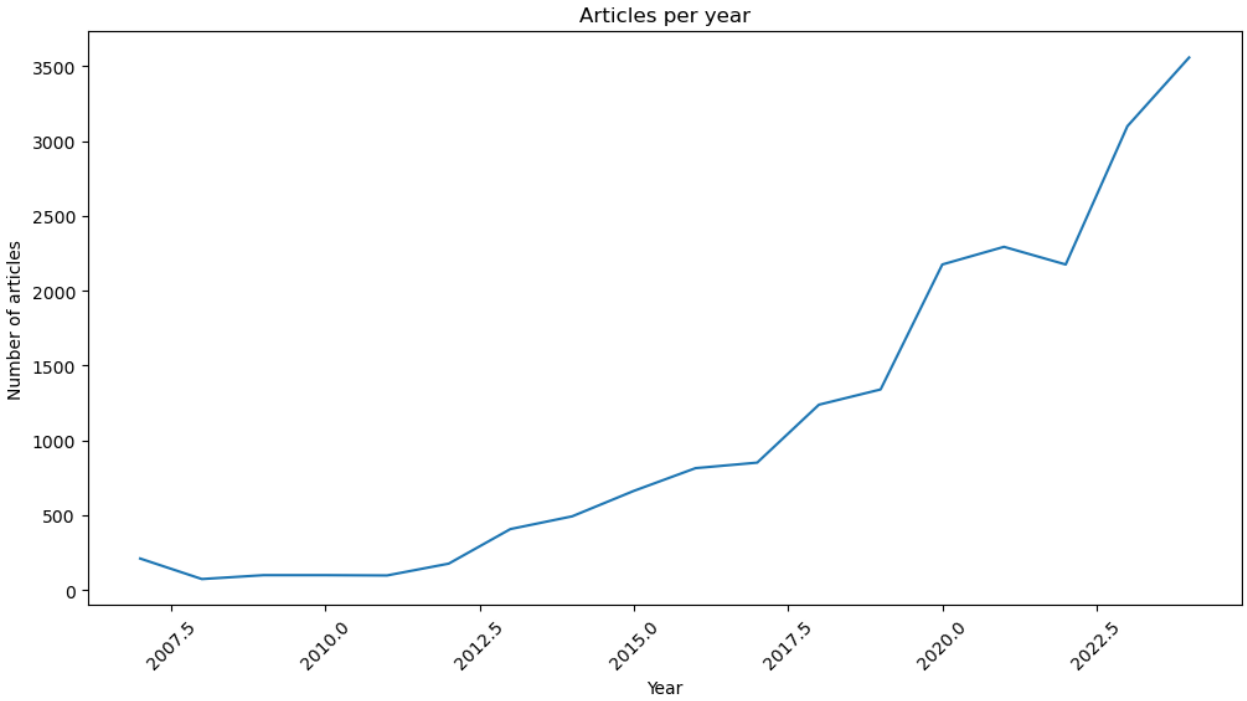
Most common words on abstracts

	counts
data	21221
ai	10723
learning	10532
based	10090
paper	9599
research	9046
social	8994
model	8799
study	8223
models	8084
information	7941
using	7406
systems	7319
use	7230
results	6065
users	6037
different	5967
analysis	5916
work	5747
used	5725

Analyze articles by year

In [23]:
```
df['update_date'] = pd.to_datetime(df['update_date'])
df['year'] = df['update_date'].dt.year
```

```
articles_per_year = df['year'].value_counts().sort_index()

plt.figure(figsize=(12, 6))
sns.lineplot(x=articles_per_year.index, y=articles_per_year.values)
plt.title('Articles per year')
plt.xlabel('Year')
plt.ylabel('Number of articles')
plt.xticks(rotation=45)
plt.show()
```



Use an LDA model to try to create the main topics based on the articles abstracts

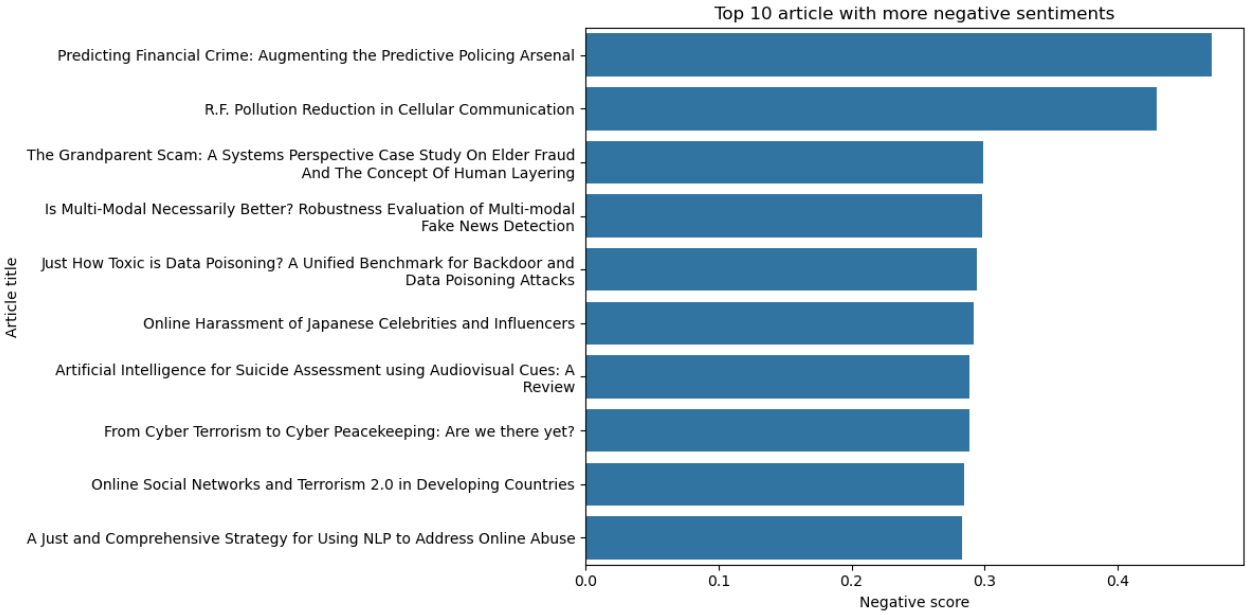
In [33]:
```
vectorizer = CountVectorizer(stop_words='english', max_features=1000)
abstract_matrix = vectorizer.fit_transform(df['abstract'].dropna())

lda_model = LatentDirichletAllocation(n_components=10, random_state=42)
lda_model.fit(abstract_matrix)

for index, topic in enumerate(lda_model.components_):
    print(f'Topic {index+1}:')
    print([vectorizer.get_feature_names_out()[i] for i in topic.argsort()
```

```
Topic 1:
['cyber', 'paper', 'user', 'internet', 'information', 'users', 'digital',
'data', 'security', 'privacy']
Topic 2:
['accuracy', 'based', 'bias', 'performance', 'machine', 'data', 'learnin
g', 'model', 'fairness', 'models']
Topic 3:
['work', 'results', 'impact', 'social', 'people', 'pandemic', 'study', '1
9', 'covid', 'health']
Topic 4:
['knowledge', 'based', 'programming', 'course', 'study', 'educational', 's
tudent', 'education', 'learning', 'students']
Topic 5:
['cloud', 'challenges', 'systems', 'paper', 'energy', 'smart', 'computin
g', 'technologies', 'technology', 'research']
Topic 6:
['approach', 'game', 'decision', 'systems', 'blockchain', 'process', 'pape
r', 'information', 'model', 'based']
Topic 7:
['paper', 'models', 'research', 'llms', 'ethical', 'artificial', 'intellig
ence', 'systems', 'human', 'ai']
Topic 8:
['paper', 'tools', 'knowledge', 'machine', 'open', 'learning', 'analysis',
'science', 'research', 'data']
Topic 9:
['political', 'twitter', 'news', 'information', 'language', 'users', 'onli
ne', 'content', 'media', 'social']
Topic 10:
['information', 'urban', 'web', 'using', 'users', 'user', 'based', 'networ
k', 'time', 'data']
```

Do an analysis sentiment to try to guess if the articles have a more positive, negative, neutral, critical approach...

In [38]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer # ty

analyzer = SentimentIntensityAnalyzer()

def get_sentiment(text):
    if pd.isnull(text):
        return {"neg": 0, "neu": 0, "pos": 0, "compound": 0}
    return analyzer.polarity_scores(text)

df['sentiment'] = df['abstract'].apply(lambda x: get_sentiment(x))
df[['neg', 'neu', 'pos', 'compound']] = df['sentiment'].apply(pd.Series)
df_neg_sorted = df.sort_values(by='neg', ascending=False)

top_negative_articles = df_neg_sorted[['id', 'title', 'abstract', 'neg']]

plt.figure(figsize=(12, 6))
sns.barplot(x=top_negative_articles['neg'], y=top_negative_articles['titl
plt.title('Top 10 article with more negative sentiments')
plt.xlabel('Negative score')
```

```
plt.ylabel('Article title')
plt.tight_layout()
plt.show()

print(top_negative_articles)
```



	id	title	abstract	neg
3411	1704.07826	Predicting Financial Crime: Augmenting the Pre...	Financial crime is a rampant but hidden thre...	0.471
497	1204.1789	R.F. Pollution Reduction in Cellular Communica...	Erroneous submission in violation of copyrig...	0.429
18314	2405.11789	The Grandparent Scam: A Systems Perspective Ca...	In April 2024, an 81-year-old Ohio man was c...	0.299
12512	2206.08788	Is Multi-Modal Necessarily Better? Robustness ...	The proliferation of fake news and its serio...	0.298
7879	2006.12557	Just How Toxic is Data Poisoning? A Unified Be...	Data poisoning and backdoor attacks manipula...	0.294
13202	2210.07599	Online Harassment of Japanese Celebrities and ...	Famous people, such as celebrities and influ...	0.292
11642	2201.09130	Artificial Intelligence for Suicide Assessment...	Death by suicide is the seventh leading deat...	0.289
8768	2010.07041	From Cyber Terrorism to Cyber Peacekeeping: Ar...	In Cyberspace nowadays, there is a burst of ...	0.289
1508	1410.0531	Online Social Networks and Terrorism 2.0 in De...	The advancement in technology has brought a ...	0.285
5962	1906.01738	A Just and Comprehensive Strategy for Using NL...	Online abusive behavior affects millions and...	0.283