Algoritmo de Agrupamento

Bibliotecas

In [1]:
```python
import numpy as np
import pandas as pd
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
import re, json
import warnings
import umap
warnings.filterwarnings('ignore')

from bs4 import BeautifulSoup
from sklearn.metrics import silhouette_score
from sklearn.decomposition import TruncatedSVD
from collections import Counter
from wordcloud import WordCloud,STOPWORDS
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer, ENGLISH_STOP_WORDS
from sklearn.cluster import KMeans
from sklearn.decomposition import TruncatedSVD
from prettytable import PrettyTable
```

Carregamento do Dataset

In [2]:
```python
df = pd.read_csv('uci-news-aggregator.csv')
```

In [3]:
```python
df.head()
```

Out[3]:

	ID	TITLE	URL	PUBLISHER	CATEGORY	STORY	HOSTNAME	TIMESTAMP
0	1	Fed official says weak data caused by weather,...	http://www.latimes.com/business/money/la-fi-mo...	Los Angeles Times	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.latimes.com	1394470370698
1	2	Fed's Charles Plosser sees high bar for change...	http://www.livemint.com/Politics/H2EvwJSK2VE6O...	Livemint	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.livemint.com	1394470371207
2	3	US open: Stocks fall after Fed official hints ...	http://www.ifamagazine.com/news/us-open-stocks...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371550
3	4	Fed risks falling 'behind the curve', Charles ...	http://www.ifamagazine.com/news/fed-risks-fall...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371793
4	5	Fed's Plosser: Nasty Weather Has Curbed Job Gr...	http://www.moneynews.com/Economy/federal-reser...	Moneynews	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.moneynews.com	1394470372027

In [4]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 422419 entries, 0 to 422418
Data columns (total 8 columns):
 #   Column     Non-Null Count   Dtype 
---  ------     --------------   ----- 
 0   ID         422419 non-null  int64 
 1   TITLE      422419 non-null  object
 2   URL        422419 non-null  object
 3   PUBLISHER  422417 non-null  object
 4   CATEGORY   422419 non-null  object
 5   STORY      422419 non-null  object
 6   HOSTNAME   422419 non-null  object
 7   TIMESTAMP  422419 non-null  int64 
dtypes: int64(2), object(6)
memory usage: 25.8+ MB
```

Pré-processamento

In [5]:
```python
df = df[['TITLE']]
```

In [6]:
```python
stopword_list=stopwords.words('english')
STOP_WORDS = ENGLISH_STOP_WORDS.union(stopword_list).union(STOPWORDS)
```

In [7]:
```python
def clean_text(text):
    soup = BeautifulSoup(text)
    rgx = r'[^A-Za-z0-9\s\.\']'
    return re.sub(rgx, '', soup.get_text()).replace('.', ' ')

def strip_stopwords(text):
    tokens = text.split()
    tokens = [word for word in tokens if word not in STOP_WORDS]
    return ' '.join(tokens)

def lemmatize_word(word):
    lemmatizer = WordNetLemmatizer()
    try:
        pos = json.loads(vb.part_of_speech(word))[0]['text'].split(' ')[-1][0]
        lemma = lemmatizer.lemmatize(word, pos=pos)
    except:
        lemma = lemmatizer.lemmatize(word)
    return lemma

def lemmatize_sentence(text):
    return ' '.join([lemmatize_word(word) for word in text.split()])

def preprocess_text(text):
    text = text.lower()
    cleaned_text = clean_text(text)
    cleaned_text = strip_stopwords(cleaned_text)
    cleaned_text = lemmatize_sentence(cleaned_text)
    return cleaned_text
```

In [8]:
```python
df['processed_text'] = df['TITLE'].apply(preprocess_text)
```

In [9]:
```python
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(df['processed_text'])
```

Treinamento (Kmeans)

In [10]:
```python
kmeans = KMeans(n_clusters=5, random_state=42)
kmeans.fit(X)
df['kmeans_cluster'] = kmeans.labels_
```

Visualização

Lista de notícias em cada cluster

In [11]:
```python
max_title_length = 60

for i, group in df.groupby('kmeans_cluster'):
    table = PrettyTable()
    table.field_names = ["Cluster " + str(i+1)]
    for title in group['TITLE'].head(10):
        truncated_title = title[:max_title_length].strip() + '...' if len(title) > max_title_length else title.strip()
        table.add_row([truncated_title])
    print(table)
```

```
+----------------------------------------------------------------+
|                           Cluster 1                            |
+----------------------------------------------------------------+
| Fed official says weak data caused by weather, should not sl... |
| Fed's Charles Plosser sees high bar for change in pace of ta... |
|    Fed risks falling 'behind the curve', Charles Plosser says   |
|        Fed's Plosser: Nasty Weather Has Curbed Job Growth       |
|        Plosser: Fed May Have to Accelerate Tapering Pace        |
|             Fed's Plosser: Taper pace may be too slow           |
| Fed's Plosser expects US unemployment to fall to 6.2% by the... |
| US jobs growth last month hit by weather:Fed President Charl... |
|   ECB unlikely to end sterilisation of SMP purchases - traders  |
|   ECB unlikely to end sterilization of SMP purchases: traders   |
+----------------------------------------------------------------+
+----------------------------------------------------------------+
|                           Cluster 2                            |
+----------------------------------------------------------------+
|  Apple iPhone Air designer concept shows the iPhone 6 we all... |
| Apple iPad Air Specs, Rumors, and Update: Air 2 Release Date... |
| IPhone 6 concept video: Is this what we should expect from A... |
| Here's A Concept Video For What Apple Could Do With The iPho... |
|  Meet the iPhone Air: An incredibly believable Apple concept... |
| Apple Release Round Up: iPhone 6 and iPad Air 2 Fall Release... |
|          Google Nexus 10 2: Is HTC Or Samsung Behind It?        |
| Spritz Speed Reading Technology for Samsung Claims It Can He... |
|  Tech that makes you read faster hits Samsung Gear 2 and S5,... |
| App that makes you read faster hits Samsung Gear 2 and S5, s... |
+----------------------------------------------------------------+
+----------------------------------------------------------------+
|                           Cluster 3                            |
+----------------------------------------------------------------+
|             Kim Kardashian steals Kylie Jenner's bikini         |
| Kim Kardashian showed some big-time cleavage while wearing h... |
| Kim Kardashian Wears Kylie Jenner's Bikini With Sexy Results... |
|          In Pictures: Top 15 Kim Kardashian sexy selfies        |
| Kim Kardashian dons 16-year-old sister Kylie's bikini for ey... |
|  Kim Kardashian Selfie Alert! Kim Steals Kylie Jenner's Itsy... |
|             Kim Kardashian squeezes into sister's bikini        |
|       Kim Kardashian's wardrobe malfunction: Spanx on display!  |
| Kim Kardashian shows off cleavage in sister Kylie's sexy cut... |
|       Kim Kardashian vs. Kylie Jenner: Battle of the Bikinis!   |
+----------------------------------------------------------------+
+----------------------------------------------------------------+
|                           Cluster 4                            |
+----------------------------------------------------------------+
|          Eurozone banks' sovereign exposure hits new high       |
|                   3 Predictions for the New Week                |
|        Mt. Gox files for US bankruptcy amid new hacker claims   |
|        New York Metro-North worker struck and killed by train   |
| Herbalife Comments on New York Times Report Exposing Pershin... |
|  JetBlue airplanes at their gates at John F. Kennedy Airport... |
|   As Transit Debate Continues, New Report Suggests High Demand  |
|              EU delays talks on new Russian pipeline            |
|             Bromund: Europe must wake up to new danger          |
|         World has new top banana as Chiquita, Fyffes merge      |
+----------------------------------------------------------------+
+----------------------------------------------------------------+
|                           Cluster 5                            |
+----------------------------------------------------------------+
| US open: Stocks fall after Fed official hints at accelerated... |
| ECB FOCUS-Stronger euro drowns out ECB's message to keep rat... |
|          TECH STOCKS: EBay And Icahn Keep Trading Punches       |
| eBay's John Donahoe talks Icahn, conflicts, and $100 stock p... |
| Stock market live blog: S&P 500 retreats from record after d... |
| Five years after stock meltdown, most Cleveland-area compani... |
|         Weak China exports weigh on stocks, hit commodities     |
|              US stocks dip on weak Asian data, Ukraine          |
|                Stocks fall on weak China export data            |
|              Global Growth Worries May Pressure Stocks          |
+----------------------------------------------------------------+
```

Palavras mais frequentes

In [12]:
```python
# criando uma lista de dicionários com os dados da tabela
data = []

for i in range(5):
    cluster_text = ' '.join(df[df['kmeans_cluster'] == i]['processed_text'])
    words = Counter(cluster_text.split()).most_common(15)
    top_words = [word[0] for word in words]
    data.append({'Cluster': f'Cluster {i+1}', 'Palavras mais frequentes': top_words})

# criando a tabela
table = PrettyTable()
table.field_names = [item['Cluster'] for item in data]
for i in range(15):
    table.add_row([item['Palavras mais frequentes'][i] for item in data])

print(table)
```

```
+-----------+-----------+--------------+-----------+--------------+
| Cluster 1 | Cluster 2 |  Cluster 3   | Cluster 4 |  Cluster 5   |
+-----------+-----------+--------------+-----------+--------------+
|    say    |   apple   |     kim      |    new    |    stock     |
|   google  |  samsung  |  kardashian  |    york   |     rate     |
|   video   |   galaxy  |    kanye     |   video   |    market    |
|    2014   |     s5    |     west     |  trailer  |   mortgage   |
|   report  |     4     |   wedding    |   album   |     rise     |
|     2     |   iphone  | kardashian's |  release  |    higher    |
|     1     |     v     |    west's    |   apple   |     fall     |
|   price   |     2     |    vogue     |   google  |     bank     |
|    day    |    beat   |    photo     |   season  |     gain     |
|    star   |  apple's  |    cover     |   report  | unemployment |
| microsoft |     3     |    north     |  feature  |    lower     |
|  million  |   google  |    paris     |    say    |   earnings   |
|  facebook |    tab    |   married    |   watch   |     data     |
|   review  |    note   |    jenner    |   movie   |     drop     |
|   watch   |   price   |     rob      |    star   |     fed      |
+-----------+-----------+--------------+-----------+--------------+
```

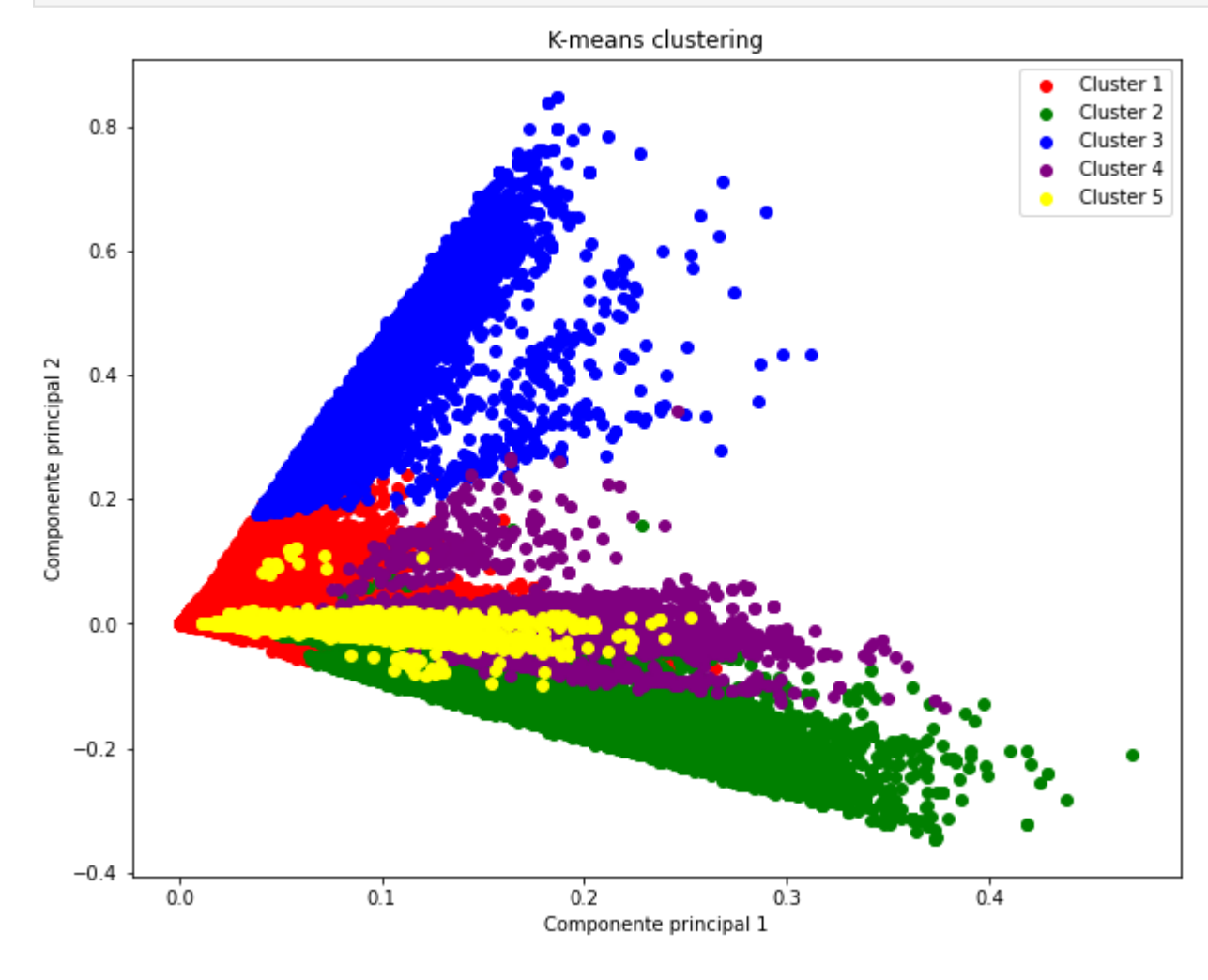
Demonstração Gráfica

In [13]:
```python
# Reduzindo a dimensionalidade com SVD
svd = TruncatedSVD(n_components=2, random_state=42)
X_2d = svd.fit_transform(X)
```

In [14]:
```python
# Cor de cada cluster
colors = ['red', 'green', 'blue', 'purple', 'yellow']
plt.figure(figsize=(10, 8))

# Plotando os pontos de cada cluster com uma cor específica
for i in range(len(colors)):
    plt.scatter(X_2d[kmeans.labels_ == i, 0], X_2d[kmeans.labels_ == i, 1], c=colors[i], label=f'Cluster {i+1}')

plt.title('K-means clustering')
plt.xlabel('Componente principal 1')
plt.ylabel('Componente principal 2')
plt.legend()
plt.show()
```

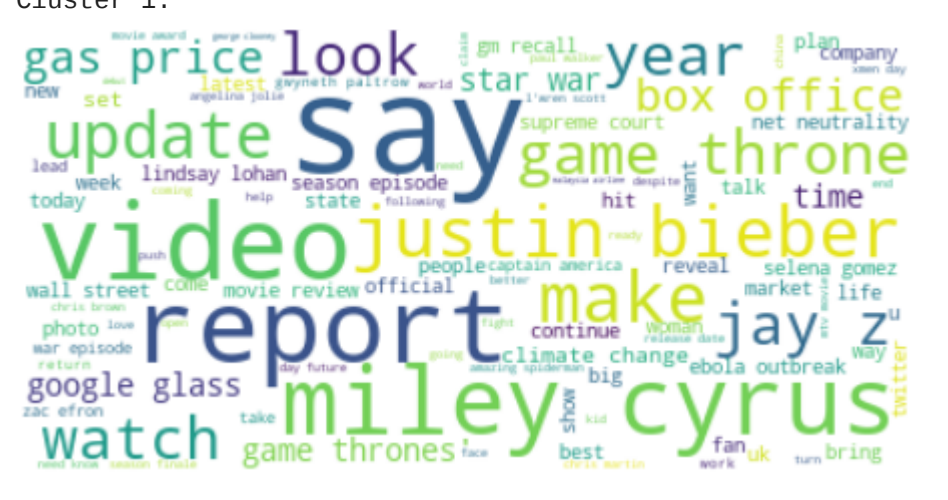


Nuvem de palavras

In [15]:
```python
# Função para gerar a nuvem de palavras
def generate_wordcloud(text):
    wordcloud = WordCloud(max_font_size=60, max_words=100, background_color="white").generate(text)
    plt.figure(figsize=(8, 6))
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.show()
    print("\n")

# Gerando a nuvem de palavras para cada cluster do K-means
for i in range(5):
    cluster_text = ' '.join(df[df['kmeans_cluster'] == i]['processed_text'])
    print(f'Cluster {i+1}:')
    generate_wordcloud(cluster_text)
```

Cluster 1:



Cluster 2:

Cluster 3:

Cluster 4:

Cluster 5:

Métricas

In [16]:
```python
kmeans_inertial = kmeans.inertia_
#silhouette = silhouette_score(X, kmeans.labels_)

# Criando a tabela
table = PrettyTable()
table.field_names = ["Métrica", "Valor"]
table.add_row(["Silhouette score", "-"])
table.add_row(["Inércia", round(kmeans_inertial, 2)])

# Imprimindo a tabela
print(table)
```

```
+------------------+----------+
|     Métrica      |  Valor   |
+------------------+----------+
| Silhouette score |    -     |
|     Inércia      | 417913.0 |
+------------------+----------+
```

In [23]:
```python
inertias = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    inertias.append(kmeans.inertia_)

plt.plot(range(1, 11), inertias, marker='o')
plt.title('K-means Inertia')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()
```