In [ ]:
```
### Analiza - minim 10 fraze
# Fisier txt - fraze despre criptomonede
```

In [1]:
```
# Importarea bibliotecilor necesare
import nltk # Natural Language Toolkit - analiza si procesarea textului
from nltk.corpus import stopwords # stopwords ofera o lista de cuvinte comune de sfarsit
from nltk.tokenize import word_tokenize # word_tokenize este folosit pt a imparti textul
from nltk.probability import FreqDist # FreqDist - distributii de frecventa a cuvintelor
from nltk.sentiment import SentimentIntensityAnalyzer # analiza de sentiment asupra text
import pandas as pd # manipularea si procesarea datelor
import matplotlib.pyplot as plt # vizualizarea datelor - crearea de grafice (bar chart,
from collections import Counter # numararea cuvintelor frecvente
import re # procesarea si extragerea textului
```

In [19]:
```
fraze=pd.read_table(r"C:\Materiale facultate\MASTER CEC\Data Mining\fraze.txt", delimite
fraze
```

Out[19]:

	sentences
0	Cryptocurrencies have taken the world by storm...
1	These digital assets have become a hot topic, ...
2	In this article, we will delve into the fascin...
3	The cryptocurrency revolution began with the c...
4	Bitcoin introduced the concept of a decentrali...
5	This technology allows for secure and transpar...
6	At the heart of cryptocurrencies is blockchain...
7	A blockchain is a distributed ledger that reco...
8	These transactions are grouped into "blocks" a...
9	This decentralized and tamper-resistant ledger...
10	Cryptocurrencies operate on a peer-to-peer net...
11	This process, often referred to as mining, inv...
12	The future of cryptocurrencies is both promisi...
13	As technology continues to evolve, cryptocurre...
14	Innovations in blockchain technology and the g...
15	Cryptocurrencies are digital or virtual curren...
16	Bitcoin, created in 2009, was the first decent...
17	The growing popularity of cryptocurrencies has...
18	Cryptocurrencies enable peer-to-peer transacti...
19	Ethereum, a prominent cryptocurrency, introduc...
20	Despite their innovative potential, cryptocurr...

In [5]:
```
pip install spacy
```

```
Collecting spacy
  Downloading spacy-3.7.2-cp310-cp310-win_amd64.whl (12.1 MB)
     ---------------------------------------- 12.1/12.1 MB 6.1 MB/s eta 0:00:00
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from spacy) (5.2.1)
Collecting preshed<3.1.0,>=3.0.2
  Downloading preshed-3.0.9-cp310-cp310-win_amd64.whl (122 kB)
```

```
     ------------------------------------- 122.2/122.2 kB 3.5 MB/s eta 0:00:00
Collecting murmurhash<1.1.0,>=0.28.0
  Downloading murmurhash-1.0.10-cp310-cp310-win_amd64.whl (25 kB)
Collecting wasabi<1.2.0,>=0.9.1
  Downloading wasabi-1.1.2-py3-none-any.whl (27 kB)
Collecting weasel<0.4.0,>=0.1.0
  Downloading weasel-0.3.4-py3-none-any.whl (50 kB)
     ---------------------------------------- 50.1/50.1 kB 2.5 MB/s eta 0:00:00
Requirement already satisfied: jinja2 in c:\users\oana\anaconda3\lib\site-packages (from
spacy) (3.1.2)
Collecting langcodes<4.0.0,>=3.2.0
  Downloading langcodes-3.3.0-py3-none-any.whl (181 kB)
     ------------------------------------- 181.6/181.6 kB 5.4 MB/s eta 0:00:00
Collecting srsly<3.0.0,>=2.4.3
  Downloading srsly-2.4.8-cp310-cp310-win_amd64.whl (481 kB)
     ------------------------------------- 481.9/481.9 kB 4.3 MB/s eta 0:00:00
Collecting thinc<8.3.0,>=8.1.8
  Downloading thinc-8.2.2-cp310-cp310-win_amd64.whl (1.5 MB)
     ---------------------------------------- 1.5/1.5 MB 5.9 MB/s eta 0:00:00
Collecting spacy-legacy<3.1.0,>=3.0.11
  Downloading spacy_legacy-3.0.12-py2.py3-none-any.whl (29 kB)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy) (2.28.1)
Collecting cymem<2.1.0,>=2.0.2
  Downloading cymem-2.0.8-cp310-cp310-win_amd64.whl (39 kB)
Collecting typer<0.10.0,>=0.3.0
  Downloading typer-0.9.0-py3-none-any.whl (45 kB)
     ---------------------------------------- 45.9/45.9 kB 2.2 MB/s eta 0:00:00
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy) (4.64.1)
Collecting spacy-loggers<2.0.0,>=1.0.0
  Downloading spacy_loggers-1.0.5-py3-none-any.whl (22 kB)
Requirement already satisfied: numpy>=1.19.0 in c:\users\oana\anaconda3\lib\site-package
s (from spacy) (1.23.5)
Requirement already satisfied: setuptools in c:\users\oana\anaconda3\lib\site-packages
(from spacy) (65.6.3)
Requirement already satisfied: packaging>=20.0 in c:\users\oana\anaconda3\lib\site-packa
ges (from spacy) (22.0)
Collecting pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4
  Downloading pydantic-2.5.3-py3-none-any.whl (381 kB)
     ------------------------------------- 381.9/381.9 kB 6.0 MB/s eta 0:00:00
Collecting catalogue<2.1.0,>=2.0.6
  Downloading catalogue-2.0.10-py3-none-any.whl (17 kB)
Collecting pydantic-core==2.14.6
  Downloading pydantic_core-2.14.6-cp310-none-win_amd64.whl (1.9 MB)
     ---------------------------------------- 1.9/1.9 MB 2.5 MB/s eta 0:00:00
Collecting typing-extensions>=4.6.1
  Downloading typing_extensions-4.9.0-py3-none-any.whl (32 kB)
Collecting annotated-types>=0.4.0
  Downloading annotated_types-0.6.0-py3-none-any.whl (12 kB)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\oana\anaconda3\lib\site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy) (2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\oana\anaconda3\lib\site
-packages (from requests<3.0.0,>=2.13.0->spacy) (1.26.14)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\oana\anaconda3\lib\s
ite-packages (from requests<3.0.0,>=2.13.0->spacy) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\oana\anaconda3\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy) (3.4)
Collecting confection<1.0.0,>=0.0.1
  Downloading confection-0.1.4-py3-none-any.whl (35 kB)
Collecting blis<0.8.0,>=0.7.8
  Downloading blis-0.7.11-cp310-cp310-win_amd64.whl (6.6 MB)
     ---------------------------------------- 6.6/6.6 MB 5.6 MB/s eta 0:00:00
Requirement already satisfied: colorama in c:\users\oana\anaconda3\lib\site-packages (fr
om tqdm<5.0.0,>=4.38.0->spacy) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\oana\anaconda3\lib\site-p
```

```
ackages (from typer<0.10.0,>=0.3.0->spacy) (8.0.4)
Collecting cloudpathlib<0.17.0,>=0.7.0
  Downloading cloudpathlib-0.16.0-py3-none-any.whl (45 kB)
     ---------------------------------------- 45.0/45.0 kB 552.1 kB/s eta 0:00:00
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\oana\anaconda3\lib\site-packa
ges (from jinja2->spacy) (2.1.1)
Installing collected packages: cymem, wasabi, typing-extensions, spacy-loggers, spacy-le
gacy, murmurhash, langcodes, catalogue, blis, annotated-types, typer, srsly, pydantic-co
re, preshed, cloudpathlib, pydantic, confection, weasel, thinc, spacy
  Attempting uninstall: typing-extensions
    Found existing installation: typing_extensions 4.4.0
    Uninstalling typing_extensions-4.4.0:
      Successfully uninstalled typing_extensions-4.4.0
Successfully installed annotated-types-0.6.0 blis-0.7.11 catalogue-2.0.10 cloudpathlib-
0.16.0 confection-0.1.4 cymem-2.0.8 langcodes-3.3.0 murmurhash-1.0.10 preshed-3.0.9 pyda
ntic-2.5.3 pydantic-core-2.14.6 spacy-3.7.2 spacy-legacy-3.0.12 spacy-loggers-1.0.5 srsl
y-2.4.8 thinc-8.2.2 typer-0.9.0 typing-extensions-4.9.0 wasabi-1.1.2 weasel-0.3.4
Note: you may need to restart the kernel to use updated packages.
```

In [26]:

```python
# Etapa de preprocesare a textului

# Tokenizare, Lemmatizare, Eliminare stopwords, a semnelor de punctuatie si a numerelor
import spacy
import nltk
import pandas as pd
from nltk.corpus import stopwords
from string import punctuation

# # Incarcam English Language Model in SpaCy
nlp = spacy.load("en_core_web_sm")


# Descarcam NLTK stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

def preprocess_text(text):
    # Tokenizare si Lemmatizare folosind SpaCy
    doc = nlp(text)
    tokens = [token.lemma_ for token in doc]

    # Eliminarea stopwords, a semnelor de punctuatie si a numerelor
    tokens = [word for word in tokens if word.lower() not in stop_words and not word.isd


    processed_text = ' '.join(tokens)
    return processed_text
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Oana\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [23]:

```
!pip install spacy
!python -m spacy download en_core_web_sm
```

```
Requirement already satisfied: spacy in c:\users\oana\anaconda3\lib\site-packages (3.7.
2)
Requirement already satisfied: setuptools in c:\users\oana\anaconda3\lib\site-packages
(from spacy) (65.6.3)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\oana\anaconda3\lib\site-
packages (from spacy) (1.1.2)
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy) (8.2.2)
Requirement already satisfied: numpy>=1.19.0 in c:\users\oana\anaconda3\lib\site-package
s (from spacy) (1.23.5)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\oana\anaconda3\li
```

```
b\site-packages (from spacy) (3.0.12)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in c:\users\oana\ana
conda3\lib\site-packages (from spacy) (2.5.3)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\users\oana\anaconda3\lib\site-
packages (from spacy) (0.9.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\oana\anaconda3\lib\site
-packages (from spacy) (3.0.9)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\oana\anaconda3\lib
\site-packages (from spacy) (1.0.10)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy) (2.4.8)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in c:\users\oana\anaconda3\lib\site-
packages (from spacy) (0.3.4)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy) (3.3.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy) (2.0.8)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy) (2.28.1)
Requirement already satisfied: packaging>=20.0 in c:\users\oana\anaconda3\lib\site-packa
ges (from spacy) (22.0)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy) (2.0.10)
Requirement already satisfied: jinja2 in c:\users\oana\anaconda3\lib\site-packages (from
spacy) (3.1.2)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from spacy) (5.2.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\oana\anaconda3\li
b\site-packages (from spacy) (1.0.5)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy) (4.64.1)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (4.9.0)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\oana\anaconda3\lib\sit
e-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.6.0)
Requirement already satisfied: pydantic-core==2.14.6 in c:\users\oana\anaconda3\lib\site
-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (2.14.6)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\oana\anaconda3\lib\s
ite-packages (from requests<3.0.0,>=2.13.0->spacy) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\oana\anaconda3\lib\site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy) (2022.12.7)
Requirement already satisfied: idna<4,>=2.5 in c:\users\oana\anaconda3\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy) (3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\oana\anaconda3\lib\site
-packages (from requests<3.0.0,>=2.13.0->spacy) (1.26.14)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\oana\anaconda3\lib\site-pa
ckages (from thinc<8.3.0,>=8.1.8->spacy) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from thinc<8.3.0,>=8.1.8->spacy) (0.1.4)
Requirement already satisfied: colorama in c:\users\oana\anaconda3\lib\site-packages (fr
om tqdm<5.0.0,>=4.38.0->spacy) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\oana\anaconda3\lib\site-p
ackages (from typer<0.10.0,>=0.3.0->spacy) (8.0.4)
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in c:\users\oana\anaconda3\li
b\site-packages (from weasel<0.4.0,>=0.1.0->spacy) (0.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\oana\anaconda3\lib\site-packa
ges (from jinja2->spacy) (2.1.1)
Collecting en-core-web-sm==3.7.1
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm
-3.7.1/en_core_web_sm-3.7.1-py3-none-any.whl (12.8 MB)
     ---------------------------------------- 12.8/12.8 MB 5.9 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.8.0,>=3.7.2 in c:\users\oana\anaconda3\lib\site-p
ackages (from en-core-web-sm==3.7.1) (3.7.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\oana\anaconda3\lib\si
```

```
te-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.0.10)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in c:\users\oana\ana
conda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.5.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\oana\anaconda3\lib
\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.0.10)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\users\oana\anaconda3\lib\site-
packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.9.0)
Requirement already satisfied: setuptools in c:\users\oana\anaconda3\lib\site-packages
(from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (65.6.3)
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (8.2.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\oana\anaconda3\li
b\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (3.0.12)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (5.2.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (4.64.1)
Requirement already satisfied: jinja2 in c:\users\oana\anaconda3\lib\site-packages (from
spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (3.1.2)
Requirement already satisfied: numpy>=1.19.0 in c:\users\oana\anaconda3\lib\site-package
s (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.23.5)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\oana\anaconda3\li
b\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.0.5)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\oana\anaconda3\lib\site-
packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.1.2)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in c:\users\oana\anaconda3\lib\site-
packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.3.4)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (3.3.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\oana\anaconda3\lib\site-p
ackages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.0.8)
Requirement already satisfied: packaging>=20.0 in c:\users\oana\anaconda3\lib\site-packa
ges (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (22.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\oana\anaconda3\lib\site
-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (3.0.9)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\oana\anaconda3\lib\si
te-packages (from spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.28.1)
Requirement already satisfied: pydantic-core==2.14.6 in c:\users\oana\anaconda3\lib\site
-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.8.0,>=3.7.2->en-core-web-s
m==3.7.1) (2.14.6)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.8.0,>=3.7.2->en-core-we
b-sm==3.7.1) (4.9.0)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\oana\anaconda3\lib\sit
e-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.8.0,>=3.7.2->en-core-web-
sm==3.7.1) (0.6.0)
Requirement already satisfied: idna<4,>=2.5 in c:\users\oana\anaconda3\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\oana\anaconda3\lib\site
-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.
26.14)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\oana\anaconda3\lib\s
ite-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1)
(2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\oana\anaconda3\lib\site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2022.
12.7)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\oana\anaconda3\lib\s
ite-packages (from thinc<8.3.0,>=8.1.8->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.
1.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\oana\anaconda3\lib\site-pa
ckages (from thinc<8.3.0,>=8.1.8->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.7.11)
Requirement already satisfied: colorama in c:\users\oana\anaconda3\lib\site-packages (fr
om tqdm<5.0.0,>=4.38.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\oana\anaconda3\lib\site-p
ackages (from typer<0.10.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (8.0.4)
```

```
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in c:\users\oana\anaconda3\lib\site-packages (from weasel<0.4.0,>=0.1.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\oana\anaconda3\lib\site-packages (from jinja2->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.1.1)
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-3.7.1
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [28]:
```
# Aplicarea functiei de preprocesare pe data frame-ul fraze
fraze["sentences_preprocess"]=fraze["sentences"].apply(preprocess_text)
fraze
```

Out[28]:

	sentences	sentences_preprocess
0	Cryptocurrencies have taken the world by storm...	cryptocurrencie take world storm transform way...
1	These digital assets have become a hot topic, ...	digital asset become hot topic capture attenti...
2	In this article, we will delve into the fascin...	article delve fascinating world cryptocurrenci...
3	The cryptocurrency revolution began with the c...	cryptocurrency revolution begin creation Bitco...
4	Bitcoin introduced the concept of a decentrali...	Bitcoin introduce concept decentralized digita...
5	This technology allows for secure and transpar...	technology allow secure transparent transactio...
6	At the heart of cryptocurrencies is blockchain...	heart cryptocurrencie blockchain technology
7	A blockchain is a distributed ledger that reco...	blockchain distribute ledger record transactio...
8	These transactions are grouped into "blocks" a...	transaction group block link together chronolo...
9	This decentralized and tamper-resistant ledger...	decentralized tamper resistant ledger ensure i...
10	Cryptocurrencies operate on a peer-to-peer net...	cryptocurrencie operate peer peer network part...
11	This process, often referred to as mining, inv...	process often refer mining involve solve compl...
12	The future of cryptocurrencies is both promisi...	future cryptocurrencie promising uncertain
13	As technology continues to evolve, cryptocurre...	technology continue evolve cryptocurrencie lik...
14	Innovations in blockchain technology and the g...	innovation blockchain technology growth crypto...
15	Cryptocurrencies are digital or virtual curren...	cryptocurrencie digital virtual currency use c...
16	Bitcoin, created in 2009, was the first decent...	Bitcoin create first decentralized cryptocurre...
17	The growing popularity of cryptocurrencies has...	grow popularity cryptocurrencie spark debate p...
18	Cryptocurrencies enable peer-to-peer transacti...	cryptocurrencie enable peer peer transaction w...
19	Ethereum, a prominent cryptocurrency, introduc...	Ethereum prominent cryptocurrency introduce sm...
20	Despite their innovative potential, cryptocurr...	despite innovative potential cryptocurrencie a...

In [32]:
```
# Exploratory Data Analysis - EDA

# Structura datelor
fraze.head() # primele 5 inregistrari
```

Out[32]:

	sentences	sentences_preprocess
0	Cryptocurrencies have taken the world by storm...	cryptocurrencie take world storm transform way...
1	These digital assets have become a hot topic, ...	digital asset become hot topic capture attenti...
2	In this article, we will delve into the fascin...	article delve fascinating world cryptocurrenci...
3	The cryptocurrency revolution began with the c...	cryptocurrency revolution begin creation Bitco...
4	Bitcoin introduced the concept of a decentrali...	Bitcoin introduce concept decentralized digita...

In [31]:
```
# Dimensiunea datelor
fraze.shape # 21 de randuri si 2 coloane
```

Out[31]:
```
(21, 2)
```

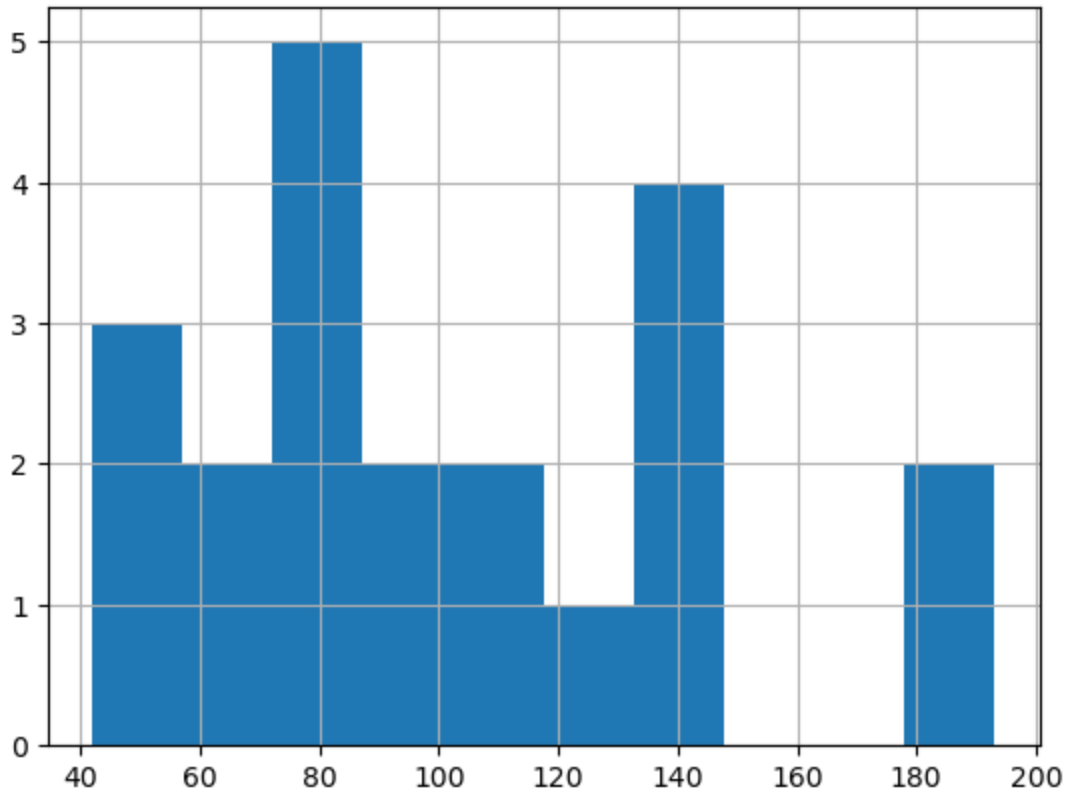
In [41]:
```
# Dimensiunea frazelor
fraze_length=fraze['sentences'].str.len()
average_fraze_length = fraze['sentences'].str.len().mean()
print('Lungimile frazelor inainte de prepreprocesare:\n',fraze_length)
print(f'Lungimea medie a frazelor inainte de preprocesare: {average_fraze_length:.2f} ca

print("_"*50)
fraze_length_preprocesat=fraze['sentences_preprocess'].str.len()
average_fraze_length_preprocesat = fraze['sentences_preprocess'].str.len().mean()
print('Lungimile frazelor dupa prepreprocesare:\n',fraze_length_preprocesat)
print(f'Lungimea medie a frazelor dupa preprocesare: {average_fraze_length_preprocesat:.
```

```
Lungimile frazelor inainte de prepreprocesare:
 0     103
1     131
2     165
3     147
4     124
5     110
6      59
7      98
8      89
9      98
10    113
11    255
12     64
13    257
14    108
15    167
16    114
17    186
18    153
19    166
20    171
Name: sentences, dtype: int64
Lungimea medie a frazelor inainte de preprocesare: 137.05 caractere
__________________________________________________
Lungimile frazelor dupa prepreprocesare:
 0      66
1      96
2      99
3     106
4     103
5      83
6      43
7      71
8      57
9      76
10     87
11    193
12     42
13    187
14     81
15    133
16     75
17    133
18    124
19    136
20    145
```

```
Name: sentences_preprocess, dtype: int64
Lungimea medie a frazelor dupa preprocesare: 101.71 caractere
```

In [42]:
```
# Distributia datelor de tip text
import matplotlib.pyplot as plt
fraze["sentences_preprocess"].str.len().hist()
plt.show() # cele mai multe fraze au 80 sau 140 de caractere
```



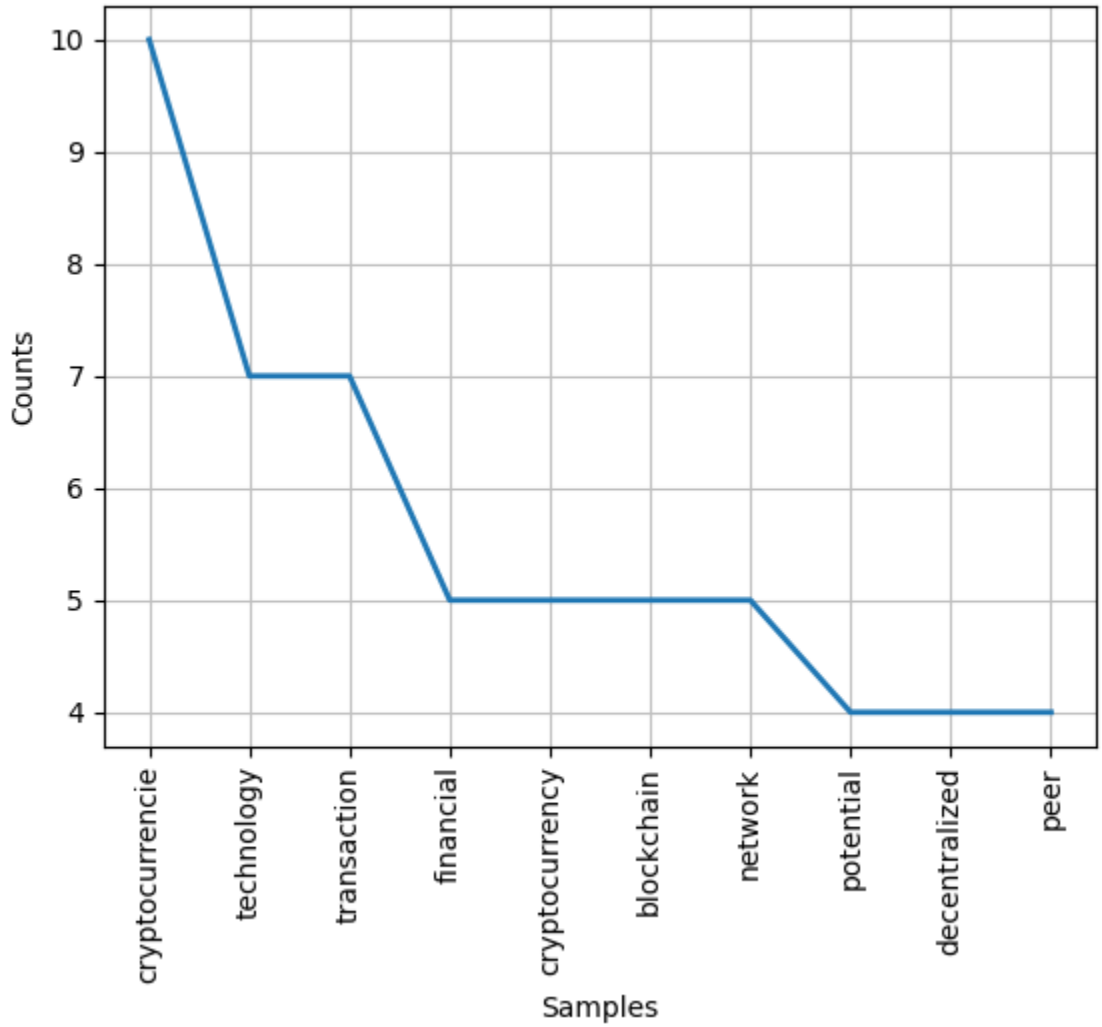
In [47]:
```
# Grafic pt reprezentarea frecventei de aparitie a cuvintelor
from nltk.probability import FreqDist
from nltk.tokenize import word_tokenize

text = ' '.join(fraze['sentences_preprocess'].tolist())

tokens = word_tokenize(text.lower())

tokens = [token for token in tokens if token.isalpha()]
freqdist = FreqDist(tokens)
print(freqdist.most_common(20))
freqdist.plot(10)
plt.show()
```

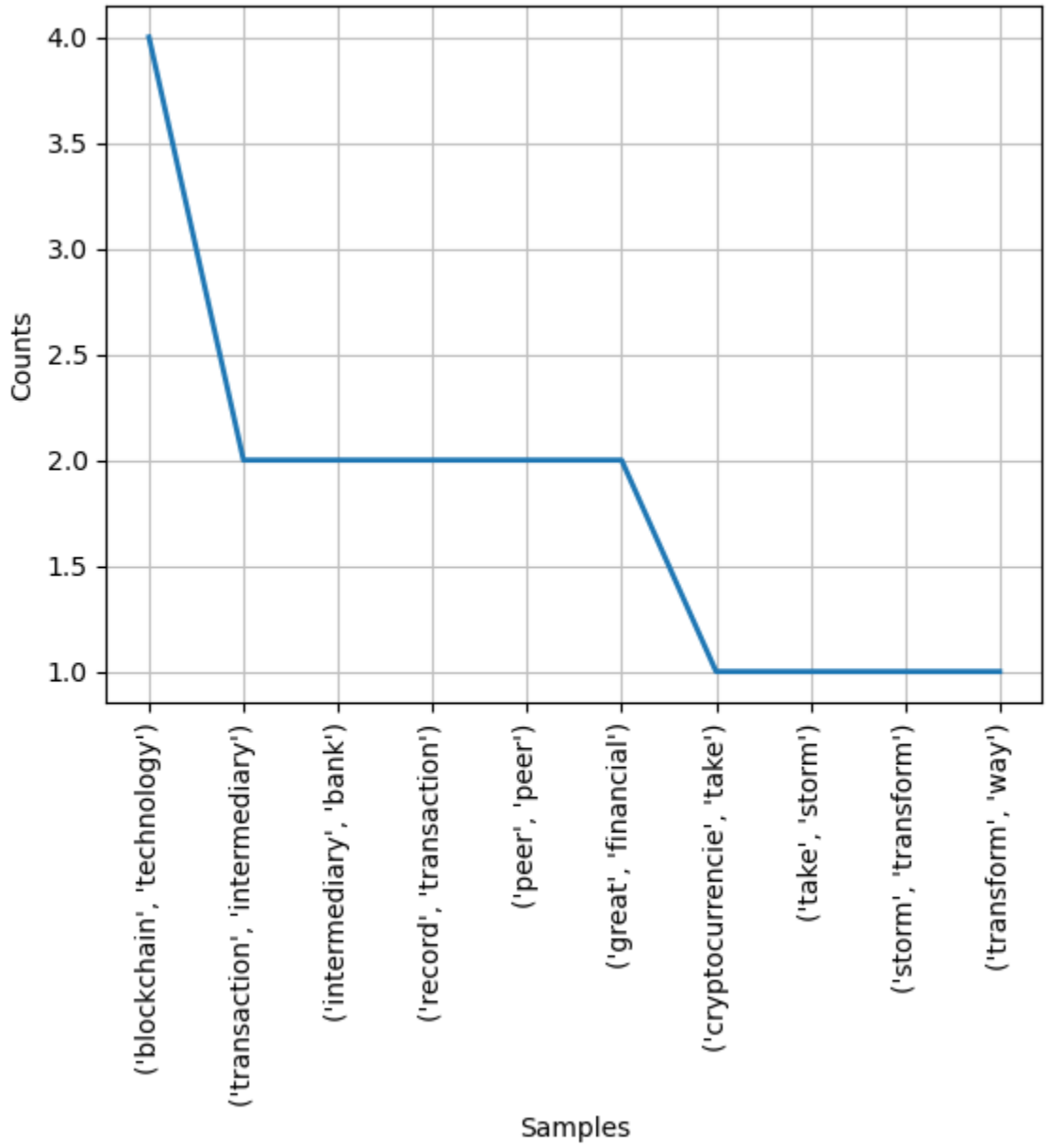
```
[('cryptocurrencie', 10), ('technology', 7), ('transaction', 7), ('financial', 5), ('cry
ptocurrency', 5), ('blockchain', 5), ('network', 5), ('potential', 4), ('decentralized',
4), ('peer', 4), ('digital', 3), ('use', 3), ('bitcoin', 3), ('secure', 3), ('become',
2), ('future', 2), ('group', 2), ('introduce', 2), ('currency', 2), ('base', 2)]
```



In [48]:
```
# Analiza N-gramelor
# BIGRAME
from nltk.util import ngrams
text = ' '.join(fraze['sentences_preprocess'].tolist())
tokens = word_tokenize(text.lower())
tokens = [token for token in tokens if token.isalpha()]
bigrams = ngrams(tokens, 2)
freqdist1 = FreqDist(bigrams)
print(freqdist1.most_common(10))

freqdist1.plot(10)
plt.show()
```

```
[(('blockchain', 'technology'), 4), (('transaction', 'intermediary'), 2), (('intermediar
y', 'bank'), 2), (('record', 'transaction'), 2), (('peer', 'peer'), 2), (('great', 'fina
ncial'), 2), (('cryptocurrencie', 'take'), 1), (('take', 'storm'), 1), (('storm', 'trans
form'), 1), (('transform', 'way'), 1)]
```

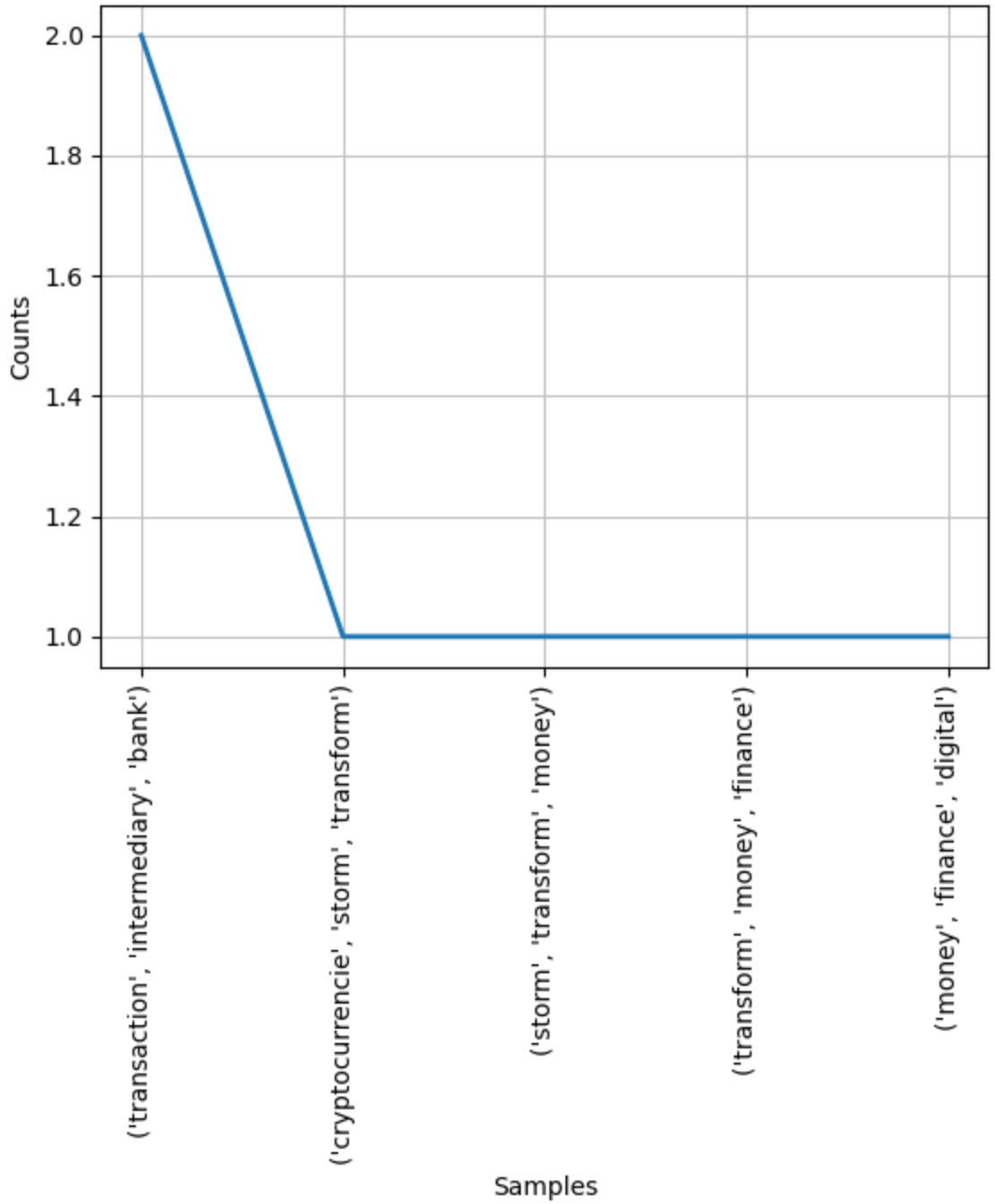


In [51]:
```
# TRIGRAME
text = ' '.join(fraze['sentences_preprocess'].tolist())
tokens = word_tokenize(text.lower())
tokens = [token for token in tokens if token.isalpha()]

trigrams = ngrams(tokens, 3)
freqdist2 = FreqDist(trigrams)
print(freqdist2.most_common(10))

freqdist2.plot(5)
plt.show()
```

```
[(('transaction', 'intermediary', 'bank'), 2), (('cryptocurrencie', 'storm', 'transfor
m'), 1), (('storm', 'transform', 'money'), 1), (('transform', 'money', 'finance'), 1),
(('money', 'finance', 'digital'), 1), (('finance', 'digital', 'asset'), 1), (('digital',
'asset', 'become'), 1), (('asset', 'become', 'hot'), 1), (('become', 'hot', 'topic'),
1), (('hot', 'topic', 'capture'), 1)]
```

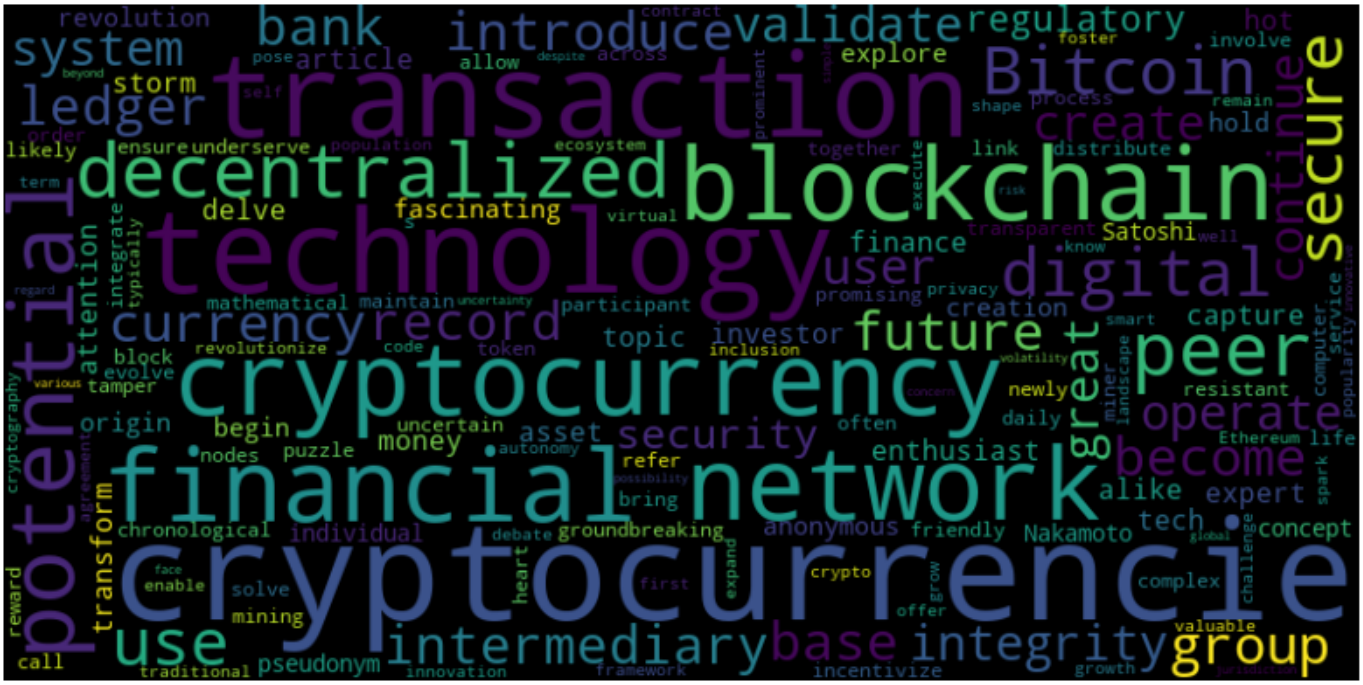


In [50]:
```
words_to_remove = ['without', 'like', 'think', 'need', 'world', 'way', 'take']
fraze['sentences_preprocess'] = fraze['sentences_preprocess'].apply(lambda x: ' '.join([
```

In [52]:
```
# Word Cloud
from wordcloud import WordCloud
import matplotlib.pyplot as plt

text = ' '.join(fraze['sentences_preprocess'])
# Word Cloud pentru cuvintele cele mai frecvente
wordcloud = WordCloud(width=800, height=400, background_color='black').generate(text)
# Afișare Word Cloud
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud fraze criptomonede')
plt.show()
```

WordCloud fraze criptomonede



In [82]:
```
# Vectorizarea textului
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(fraze['sentences_preprocess'])
```

In [65]:
```
#ignore wornings
import warnings

# Settings the warnings to be ignored
warnings.filterwarnings('ignore')

# This warning won't display due to the disabled warnings
warnings.warn('Error: A warning just appeared')
```

In [68]:
```
# K-Means ++
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
# Definim un interval pentru valorile K
text_data = fraze.shape[0]
K_values = range(2, min(text_data, 11))


best_K = 2
best_silhouette_score = -1

for K in K_values:
    kmeans = KMeans(n_clusters=K, init='k-means++', random_state=0)
    kmeans.fit(X)


    cluster_assignments = kmeans.labels_

    # Calculam scorul silhouette
    silhouette_avg = silhouette_score(X, cluster_assignments)

    # Verificam daca K curent are cel mai bun silhouette score
    if silhouette_avg > best_silhouette_score:
        best_K = K
        best_silhouette_score = silhouette_avg
```

```
# Initializam K-Means cu cea mai buna valoare K
kmeans = KMeans(n_clusters=best_K, init='k-means++', random_state=0)
kmeans.fit(X)


cluster_assignments = kmeans.labels_
for i, cluster in enumerate(cluster_assignments):
    print(f"Fraza {i} face parte din clusterul {cluster}")

print(f"Cel mai bun numar de clustere (K) este {best_K} cu un scor silhouette de {best_s
```

```
Fraza 0 face parte din clusterul 3
Fraza 1 face parte din clusterul 8
Fraza 2 face parte din clusterul 3
Fraza 3 face parte din clusterul 6
Fraza 4 face parte din clusterul 0
Fraza 5 face parte din clusterul 1
Fraza 6 face parte din clusterul 7
Fraza 7 face parte din clusterul 2
Fraza 8 face parte din clusterul 5
Fraza 9 face parte din clusterul 6
Fraza 10 face parte din clusterul 1
Fraza 11 face parte din clusterul 2
Fraza 12 face parte din clusterul 3
Fraza 13 face parte din clusterul 1
Fraza 14 face parte din clusterul 7
Fraza 15 face parte din clusterul 0
Fraza 16 face parte din clusterul 6
Fraza 17 face parte din clusterul 4
Fraza 18 face parte din clusterul 1
Fraza 19 face parte din clusterul 9
Fraza 20 face parte din clusterul 4
Cel mai bun numar de clustere (K) este 10 cu un scor silhouette de 0.03
```

In [71]:

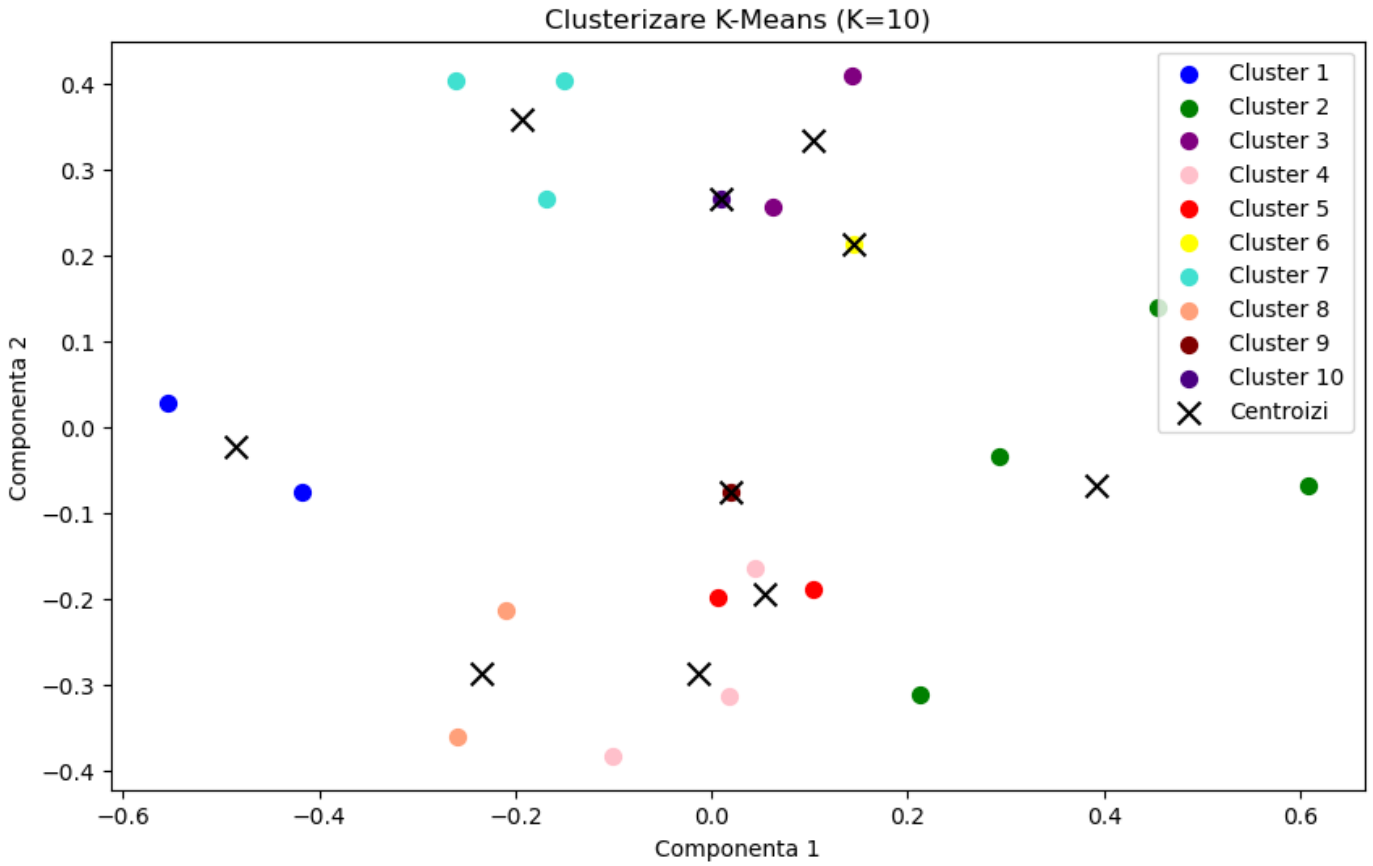
```
# Reducerea dimensionalitatii cu PCA to pentru a vizualiza clusterele
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X.toarray())

# Get the cluster assignments for each data point
cluster_assignments = kmeans.labels_

# Create a scatter plot for each data point and color by cluster assignment
plt.figure(figsize=(10, 6))
colors = ['blue', 'green', 'purple', 'pink', 'red', 'yellow', 'turquoise', 'lightsalmon'
for i in range(K):
    plt.scatter(X_pca[cluster_assignments == i, 0], X_pca[cluster_assignments == i, 1],

# Mark the cluster centers with black 'x' markers
cluster_centers = pca.transform(kmeans.cluster_centers_)
plt.scatter(cluster_centers[:, 0], cluster_centers[:, 1], s=100, c='k', marker='x', labe

plt.title(f'Clusterizare K-Means (K={K})')
plt.xlabel('Componenta 1')
plt.ylabel('Componenta 2')
plt.legend()
plt.show()
```



In [73]:
```
## Predictii
# Adaugarea coloanei 'cluster' care indica din ce cluster face parte fraza
cluster_assignments = kmeans.labels_
a=[]
for i, cluster in enumerate(cluster_assignments):
    a.append(str(cluster))


fraze["cluster"]=pd.DataFrame(a)
fraze
```

Out[73]:

	sentences	sentences_preprocess	cluster
0	Cryptocurrencies have taken the world by storm...	cryptocurrencie storm transform money finance	3
1	These digital assets have become a hot topic, ...	digital asset become hot topic capture attenti...	8
2	In this article, we will delve into the fascin...	article delve fascinating cryptocurrencie expl...	3
3	The cryptocurrency revolution began with the c...	cryptocurrency revolution begin creation Bitco...	6
4	Bitcoin introduced the concept of a decentrali...	Bitcoin introduce concept decentralized digita...	0
5	This technology allows for secure and transpar...	technology allow secure transparent transactio...	1
6	At the heart of cryptocurrencies is blockchain...	heart cryptocurrencie blockchain technology	7
7	A blockchain is a distributed ledger that reco...	blockchain distribute ledger record transactio...	2
8	These transactions are grouped into "blocks" a...	transaction group block link together chronolo...	5
9	This decentralized and tamper-resistant ledger...	decentralized tamper resistant ledger ensure i...	6
10	Cryptocurrencies operate on a peer-to-peer net...	cryptocurrencie operate peer peer network part...	1
11	This process, often referred to as mining, inv...	process often refer mining involve solve compl...	2
12	The future of cryptocurrencies is both promisi...	future cryptocurrencie promising uncertain	3
13	As technology continues to evolve, cryptocurre...	technology continue evolve cryptocurrencie lik...	1
14	Innovations in blockchain technology and the g...	innovation blockchain technology growth crypto...	7

15	Cryptocurrencies are digital or virtual curren...	cryptocurrencie digital virtual currency use c...	0
16	Bitcoin, created in 2009, was the first decent...	Bitcoin create first decentralized cryptocurre...	6
17	The growing popularity of cryptocurrencies has...	grow popularity cryptocurrencie spark debate p...	4
18	Cryptocurrencies enable peer-to-peer transacti...	cryptocurrencie enable peer peer transaction i...	1
19	Ethereum, a prominent cryptocurrency, introduc...	Ethereum prominent cryptocurrency introduce sm...	9
20	Despite their innovative potential, cryptocurr...	despite innovative potential cryptocurrencie a...	4

In [78]:
```
test_data=['GDP growth rate in major economies influences cryptocurrency markets',
           'as traders assess the overall health of the global economy',
           'blockchain technology, the backbone of most cryptocurrencies',
           'economists debate the potential implications for monetary policy financial s
           'the recent market volatility has led investors to diversify their portfolios
          'Cryptocurrency regulations are evolving globally',
          'governments working to strike a balance between fostering innovation and addr
          '"bitcoin the pioneering cryptocurrency experienced a surge in value as invest

# Step 3: Preprocess and vectorize the test data
test_tfidf_matrix = vectorizer.transform(test_data)

# Step 4: Make predictions on the test data
test_predictions = kmeans.predict(test_tfidf_matrix)

# Print the predicted cluster assignments
print("Predicted cluster assignments for test data:")
for i, cluster in enumerate(test_predictions):
    print(f"Text '{test_data[i]}' belongs to Cluster {cluster}")
```

```
Predicted cluster assignments for test data:
Text 'GDP growth rate in major economies influences cryptocurrency markets' belongs to C
luster 6
Text 'as traders assess the overall health of the global economy' belongs to Cluster 4
Text 'blockchain technology, the backbone of most cryptocurrencies' belongs to Cluster 7
Text 'economists debate the potential implications for monetary policy financial stabili
ty' belongs to Cluster 4
Text 'the recent market volatility has led investors to diversify their portfolios' belo
ngs to Cluster 4
Text 'Cryptocurrency regulations are evolving globally' belongs to Cluster 6
Text 'governments working to strike a balance between fostering innovation and addressin
g concerns related to money laundering' belongs to Cluster 3
Text '"bitcoin the pioneering cryptocurrency experienced a surge in value as investors '
belongs to Cluster 6
```

In [80]:
```
# Coeficientul silhouette pentru evaluarea predictiei
from sklearn.metrics import silhouette_score
silhouette_avg = silhouette_score(test_tfidf_matrix, test_predictions)
print(f"Scorul silhouette: {silhouette_avg}") # 11.8%
```

```
Scorul silhouette: 0.1181632096989824
```