I081 Aniruddh Kulkarni NLP Exp4

May 28, 2023

1 Name: Aniruddh Kulkarni

2 Roll no: I081

3 Stream: CS (AI)

4 Division: I

5 Semester: 5th Semester

6 Batch: I-3

7 Subject: NLP

8 Assignment-4

[4]:
```
!pip install nltk
%pip install textblob

import nltk

import pandas as pd
import numpy as np
from collections import Counter
import re,string

from textblob import TextBlob

import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn-white')


from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
from sklearn.decomposition import TruncatedSVD
```

```
%pip install wordcloud
from wordcloud import WordCloud
from textwrap import wrap

%pip install textstat
import textstat
```

```
Requirement already satisfied: nltk in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (3.8.1)
Requirement already satisfied: tqdm in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (from nltk)
(4.65.0)
Requirement already satisfied: click in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (from nltk)
(8.1.3)
Requirement already satisfied: regex>=2021.8.3 in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (from nltk)
(2023.5.5)
Requirement already satisfied: joblib in
/Users/pushpakulkarni/miniconda3/lib/python3.10/site-packages (from nltk)
(1.2.0)
Requirement already satisfied: textblob in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (0.17.1)
Requirement already satisfied: nltk>=3.1 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from textblob)
(3.7)
Requirement already satisfied: regex>=2021.8.3 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
nltk>=3.1->textblob) (2022.7.9)
Requirement already satisfied: joblib in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
nltk>=3.1->textblob) (1.2.0)
Requirement already satisfied: click in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
nltk>=3.1->textblob) (8.0.4)
Requirement already satisfied: tqdm in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
nltk>=3.1->textblob) (4.64.1)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: wordcloud in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (1.9.2)
Requirement already satisfied: pillow in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from wordcloud)
(9.2.0)
Requirement already satisfied: numpy>=1.6.1 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from wordcloud)
```

```
(1.24.3)
Requirement already satisfied: matplotlib in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from wordcloud)
(3.7.1)
Requirement already satisfied: importlib-resources>=3.2.0 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (5.12.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (1.4.2)
Requirement already satisfied: python-dateutil>=2.7 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: packaging>=20.0 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (21.3)
Requirement already satisfied: contourpy>=1.0.1 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (1.0.7)
Requirement already satisfied: cycler>=0.10 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from
matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: zipp>=3.1.0 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from importlib-
resources>=3.2.0->matplotlib->wordcloud) (3.8.0)
Requirement already satisfied: six>=1.5 in
/Users/pushpakulkarni/opt/anaconda3/lib/python3.9/site-packages (from python-
dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Collecting textstat
  Using cached textstat-0.7.3-py3-none-any.whl (105 kB)
Collecting pyphen
  Using cached pyphen-0.14.0-py3-none-any.whl (2.0 MB)
Installing collected packages: pyphen, textstat
Successfully installed pyphen-0.14.0 textstat-0.7.3
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
import spacy
%pip install scattertext
import scattertext as st
```

```
Collecting scattertext
```

```
  Using cached scattertext-0.1.19-py3-none-any.whl (8.2 MB)
Requirement already satisfied: numpy in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (1.23.5)
Requirement already satisfied: scipy in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (1.10.1)
Requirement already satisfied: scikit-learn in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (1.2.2)
Requirement already satisfied: pandas in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (2.0.1)
Collecting statsmodels (from scattertext)
  Using cached statsmodels-0.14.0-cp310-cp310-macosx_11_0_arm64.whl (9.4 MB)
Collecting flashtext (from scattertext)
  Using cached flashtext-2.7-py2.py3-none-any.whl
Collecting gensim>=4.0.0 (from scattertext)
  Using cached gensim-4.3.1-cp310-cp310-macosx_11_0_arm64.whl (24.0 MB)
Requirement already satisfied: spacy>=3.2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (3.5.3)
Requirement already satisfied: tqdm in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scattertext) (4.65.0)
Requirement already satisfied: smart-open>=1.8.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from gensim>=4.0.0->scattertext) (6.3.0)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (1.0.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (1.0.9)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (2.0.7)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (3.0.8)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (8.1.10)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
```

```
(from spacy>=3.2->scattertext) (1.1.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (2.4.6)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (2.0.8)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (0.7.0)
Requirement already satisfied: pathy>=0.10.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (0.10.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (2.29.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (1.10.8)
Requirement already satisfied: jinja2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (3.1.2)
Requirement already satisfied: setuptools in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (67.7.2)
Requirement already satisfied: packaging>=20.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (23.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from spacy>=3.2->scattertext) (3.3.0)
Requirement already satisfied: python-dateutil>=2.8.2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pandas->scattertext) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pandas->scattertext) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pandas->scattertext) (2023.3)
Requirement already satisfied: joblib>=1.1.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scikit-learn->scattertext) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from scikit-learn->scattertext) (3.1.0)
Collecting patsy>=0.5.2 (from statsmodels->scattertext)
  Using cached patsy-0.5.3-py2.py3-none-any.whl (233 kB)
```

```
Requirement already satisfied: six in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from patsy>=0.5.2->statsmodels->scattertext) (1.16.0)
Requirement already satisfied: typing-extensions>=4.2.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->spacy>=3.2->scattertext) (4.5.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from requests<3.0.0,>=2.13.0->spacy>=3.2->scattertext) (3.1.0)
Requirement already satisfied: idna<4,>=2.5 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from requests<3.0.0,>=2.13.0->spacy>=3.2->scattertext) (3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from requests<3.0.0,>=2.13.0->spacy>=3.2->scattertext) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from requests<3.0.0,>=2.13.0->spacy>=3.2->scattertext) (2023.5.7)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from thinc<8.2.0,>=8.1.8->spacy>=3.2->scattertext) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from thinc<8.2.0,>=8.1.8->spacy>=3.2->scattertext) (0.0.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from typer<0.8.0,>=0.3.0->spacy>=3.2->scattertext) (8.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from jinja2->spacy>=3.2->scattertext) (2.1.2)
Installing collected packages: flashtext, patsy, gensim, statsmodels,
scattertext
Successfully installed flashtext-2.7 gensim-4.3.1 patsy-0.5.3 scattertext-0.1.19
statsmodels-0.14.0
Note: you may need to restart the kernel to use updated packages.
```

9 About Dataset:

This contains data of news headlines published over a period of nineteen years.

Sourced from the reputable Australian news source ABC (Australian Broadcasting Corporation)

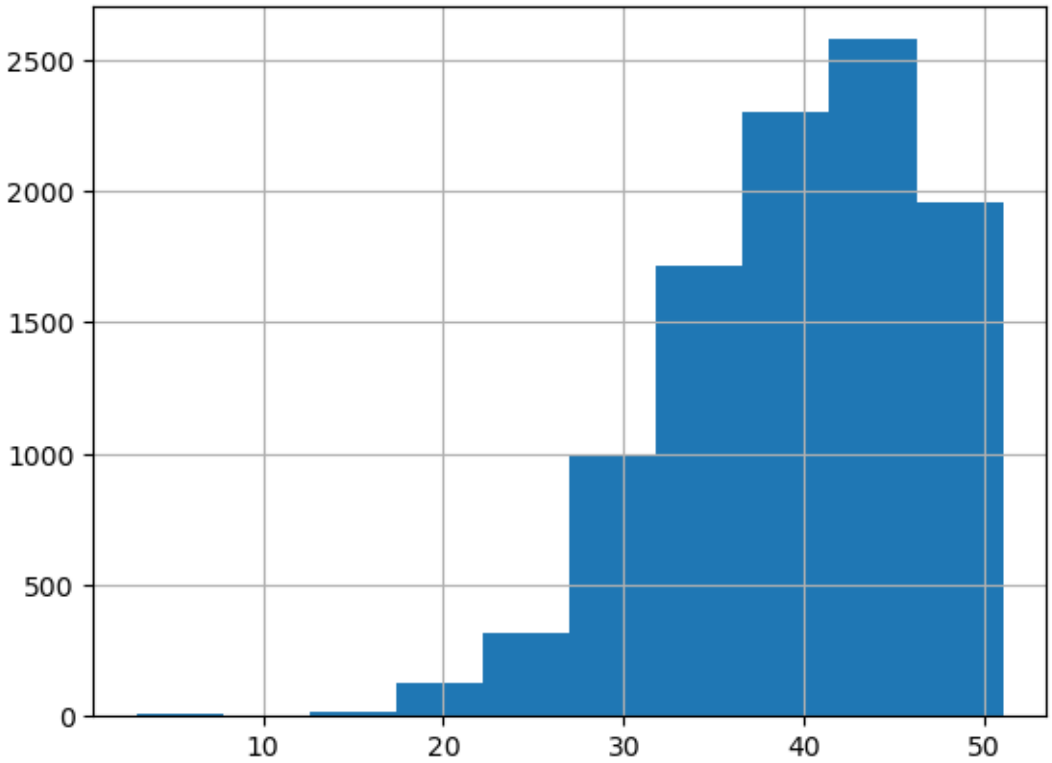
[6]:
```
import pandas as pd
df = pd.read_csv('abcnews-date-text.csv',nrows=10000)
df.head()
```

[6]:

	publish_date	headline_text
0	20030219	aba decides against community broadcasting lic…
1	20030219	act fire witnesses must be aware of defamation
2	20030219	a g calls for infrastructure protection summit
3	20030219	air nz staff in aust strike for pay rise
4	20030219	air nz strike to affect australian travellers

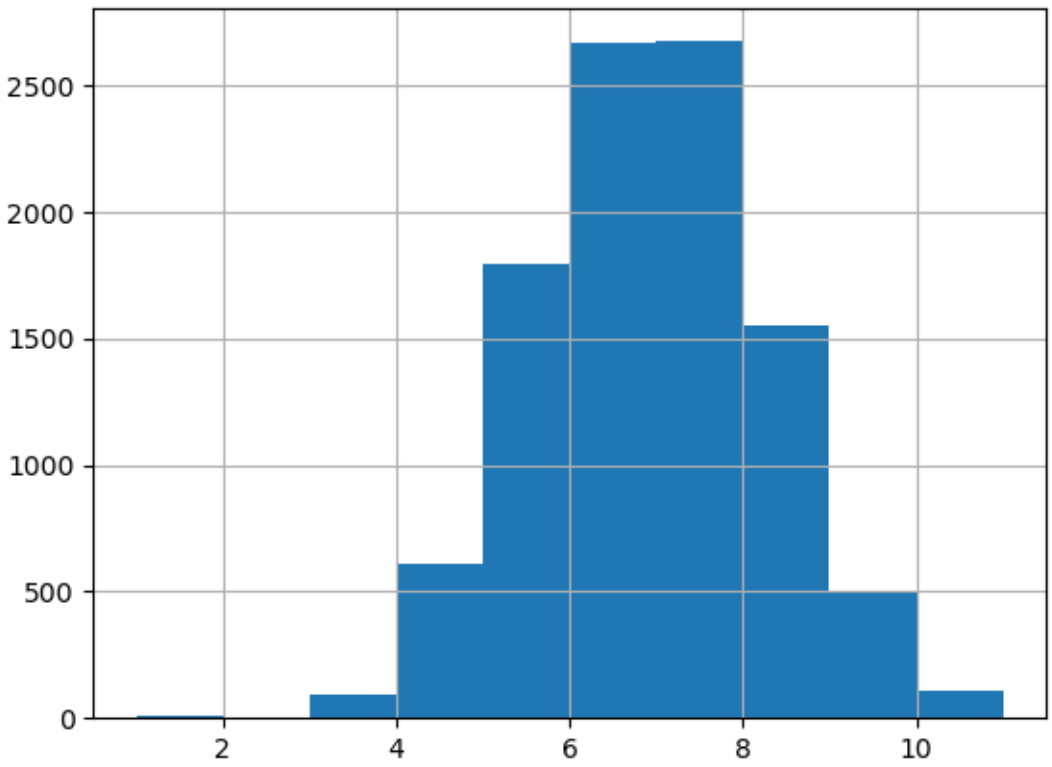
[7]:
```
df['headline_text'].str.len().hist()
```

[7]: <Axes: >



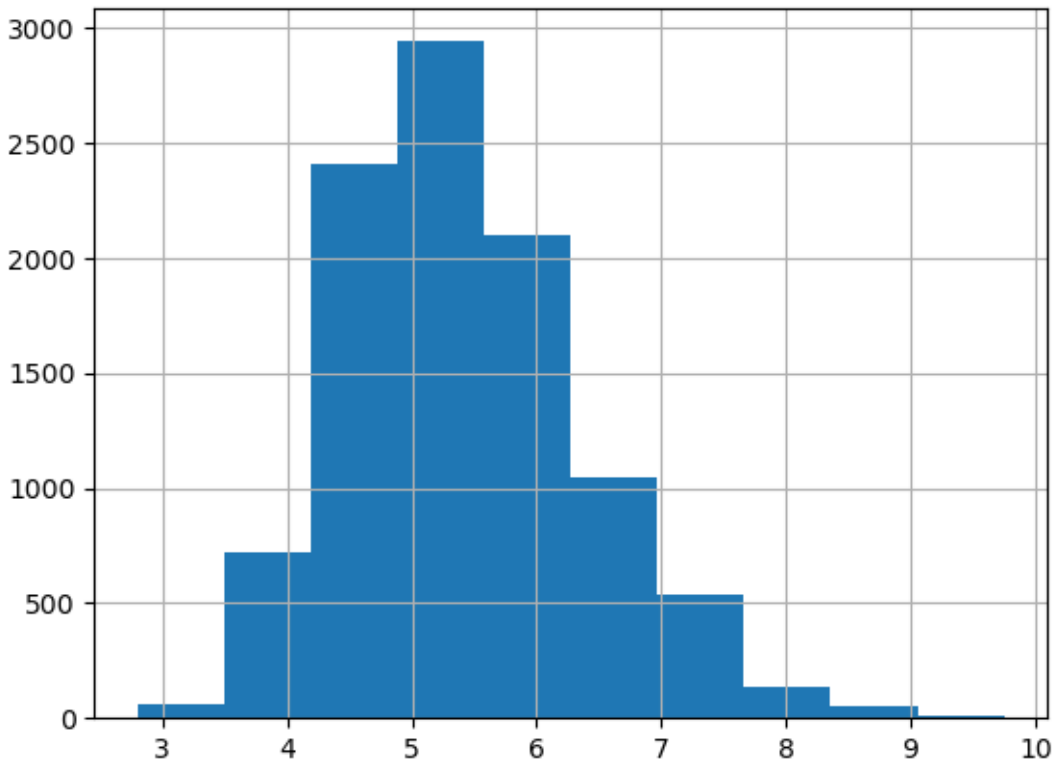
[8]:
```
df['headline_text'].str.split().map(lambda x: len(x)).hist()
```

[8]: <Axes: >



[10]:
```
#average word length in each sentence
import numpy as np
df['headline_text'].str.split().apply(lambda x : [len(i) for i in x]).
  ↪map(lambda x: np.mean(x)).hist()
```

[10]: <Axes: >



[12]:
```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop=set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/pushpakulkarni/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[13]:
```
#Creating Corpus
corpus=[]
new= df['headline_text'].str.split()
new=new.values.tolist()
corpus=[word for i in new for word in i]

from collections import defaultdict
dic=defaultdict(int)
for word in corpus:
    if word in stop:
        dic[word]+=1

corpus
```

[13]:
```
['aba',
 'decides',
 'against',
 'community',
 'broadcasting',
 'licence',
 'act',
 'fire',
 'witnesses',
 'must',
 'be',
 'aware',
 'of',
 'defamation',
 'a',
 'g',
 'calls',
 'for',
 'infrastructure',
 'protection',
 'summit',
 'air',
 'nz',
 'staff',
 'in',
 'aust',
 'strike',
 'for',
 'pay',
 'rise',
 'air',
 'nz',
 'strike',
 'to',
 'affect',
 'australian',
 'travellers',
 'ambitious',
 'olsson',
 'wins',
 'triple',
 'jump',
 'antic',
 'delighted',
 'with',
 'record',
 'breaking',
```

```
'barca',
'aussie',
'qualifier',
'stosur',
'wastes',
'four',
'memphis',
'match',
'aust',
'addresses',
'un',
'security',
'council',
'over',
'iraq',
'australia',
'is',
'locked',
'into',
'war',
'timetable',
'opp',
'australia',
'to',
'contribute',
'10',
'million',
'in',
'aid',
'to',
'iraq',
'barca',
'take',
'record',
'as',
'robson',
'celebrates',
'birthday',
'in',
'bathhouse',
'plans',
'move',
'ahead',
'big',
'hopes',
'for',
'launceston',
```

```
'cycling',
'championship',
'big',
'plan',
'to',
'boost',
'paroo',
'water',
'supplies',
'blizzard',
'buries',
'united',
'states',
'in',
'bills',
'brigadier',
'dismisses',
'reports',
'troops',
'harassed',
'in',
'british',
'combat',
'troops',
'arriving',
'daily',
'in',
'kuwait',
'bryant',
'leads',
'lakers',
'to',
'double',
'overtime',
'win',
'bushfire',
'victims',
'urged',
'to',
'see',
'centrelink',
'businesses',
'should',
'prepare',
'for',
'terrorist',
'attacks',
```

```
'calleri',
'avenges',
'final',
'defeat',
'to',
'eliminate',
'massu',
'call',
'for',
'ethanol',
'blend',
'fuel',
'to',
'go',
'ahead',
'carews',
'freak',
'goal',
'leaves',
'roma',
'in',
'ruins',
'cemeteries',
'miss',
'out',
'on',
'funds',
'code',
'of',
'conduct',
'toughens',
'organ',
'donation',
'regulations',
'commonwealth',
'bank',
'cuts',
'fixed',
'home',
'loan',
'rates',
'community',
'urged',
'to',
'help',
'homeless',
'youth',
```

```
'council',
'chief',
'executive',
'fails',
'to',
'secure',
'position',
'councillor',
'to',
'contest',
'wollongong',
'as',
'independent',
'council',
'moves',
'to',
'protect',
'tas',
'heritage',
'garden',
'council',
'welcomes',
'ambulance',
'levy',
'decision',
'council',
'welcomes',
'insurance',
'breakthrough',
'crean',
'tells',
'alp',
'leadership',
'critics',
'to',
'shut',
'up',
'dargo',
'fire',
'threat',
'expected',
'to',
'rise',
'death',
'toll',
'continues',
'to',
```

```
'climb',
'in',
's',
'korean',
'subway',
'dems',
'hold',
'plebiscite',
'over',
'iraqi',
'conflict',
'dent',
'downs',
'philippoussis',
'in',
'tie',
'break',
'thriller',
'de',
'villiers',
'to',
'learn',
'fate',
'on',
'march',
'5',
'digital',
'tv',
'will',
'become',
'commonplace',
'summit',
'direct',
'anger',
'at',
'govt',
'not',
'soldiers',
'crean',
'urges',
'dispute',
'over',
'at',
'smithton',
'vegetable',
'processing',
'plant',
```

```
'dog',
'mauls',
'18',
'month',
'old',
'toddler',
'in',
'nsw',
'dying',
'korean',
'subway',
'passengers',
'phoned',
'for',
'help',
'england',
'change',
'three',
'for',
'wales',
'match',
'epa',
'still',
'trying',
'to',
'recover',
'chemical',
'clean',
'up',
'costs',
'expressions',
'of',
'interest',
'sought',
'to',
'build',
'livestock',
'fed',
'opp',
'to',
're',
'introduce',
'national',
'insurance',
'firefighters',
'contain',
'acid',
```

```
'spill',
'four',
'injured',
'in',
'head',
'on',
'highway',
'crash',
'freedom',
'records',
'net',
'profit',
'for',
'third',
'successive',
'funds',
'allocated',
'for',
'domestic',
'violence',
'victims',
'funds',
'allocated',
'for',
'youth',
'at',
'risk',
'funds',
'announced',
'for',
'bridge',
'work',
'funds',
'to',
'go',
'to',
'cadell',
'upgrade',
'funds',
'to',
'help',
'restore',
'cossack',
'german',
'court',
'to',
'give',
```

```
'verdict',
'on',
'sept',
'11',
'accused',
'gilchrist',
'backs',
'rest',
'policy',
'girl',
'injured',
'in',
'head',
'on',
'highway',
'crash',
'gold',
'coast',
'to',
'hear',
'about',
'bilby',
'project',
'golf',
'club',
'feeling',
'smoking',
'ban',
'impact',
'govt',
'is',
'to',
'blame',
'for',
'ethanols',
'unpopularity',
'opp',
'greens',
'offer',
'police',
'station',
'alternative',
'griffiths',
'under',
'fire',
'over',
'project',
```

```
'knock',
'back',
'group',
'to',
'meet',
'in',
'north',
'west',
'wa',
'over',
'rock',
'art',
'hacker',
'gains',
'access',
'to',
'eight',
'million',
'credit',
'cards',
'hanson',
'is',
'grossly',
'naive',
'over',
'nsw',
'issues',
'costa',
'hanson',
'should',
'go',
'back',
'where',
'she',
'came',
'from',
'nsw',
'mp',
'harrington',
'raring',
'to',
'go',
'after',
'break',
'health',
'minister',
'backs',
```

```
'organ',
'and',
'tissue',
'storage',
'heavy',
'metal',
'deposits',
'survey',
'nearing',
'end',
'injured',
'rios',
'pulls',
'out',
'of',
'buenos',
'aires',
'open',
'inquest',
'finds',
'mans',
'death',
'accidental',
'investigations',
'underway',
'into',
'death',
'toll',
'of',
'korean',
'investigation',
'underway',
'into',
'elster',
'creek',
'spill',
'iraqs',
'neighbours',
'plead',
'for',
'continued',
'un',
'inspections',
'iraq',
'to',
'pay',
'for',
```

```
'own',
'rebuilding',
'white',
'house',
'irish',
'man',
'arrested',
'over',
'omagh',
'bombing',
'irrigators',
'vote',
'over',
'river',
'management',
'israeli',
'forces',
'push',
'into',
'gaza',
'strip',
'jury',
'to',
'consider',
'verdict',
'in',
'murder',
'case',
'juvenile',
'sex',
'offenders',
'unlikely',
'to',
'reoffend',
'as',
'kelly',
'disgusted',
'at',
'alleged',
'bp',
'ethanol',
'scare',
'kelly',
'not',
'surprised',
'ethanol',
'confidence',
```

```
'low',
'korean',
'subway',
'fire',
'314',
'still',
'missing',
'last',
'minute',
'call',
'hands',
'alinghi',
'big',
'lead',
'low',
'demand',
'forces',
'air',
'service',
'cuts',
'man',
'arrested',
'after',
'central',
'qld',
'hijack',
'attempt',
'man',
'charged',
'over',
'cooma',
'murder',
'man',
'fined',
'after',
'aboriginal',
'tent',
'embassy',
'raid',
'man',
'jailed',
'over',
'keno',
'fraud',
'man',
'with',
'knife',
```

```
'hijacks',
'light',
'plane',
'martin',
'to',
'lobby',
'against',
'losing',
'nt',
'seat',
'in',
'fed',
'massive',
'drug',
'crop',
'discovered',
'in',
'western',
'nsw',
'mayor',
'warns',
'landfill',
'protesters',
'meeting',
'to',
'consider',
'tick',
'clearance',
'costs',
'meeting',
'to',
'focus',
'on',
'broken',
'hill',
'water',
'woes',
'moderate',
'lift',
'in',
'wages',
'growth',
'more',
'than',
'40',
'pc',
'of',
```

```
'young',
'men',
'drink',
'alcohol',
'at',
'more',
'water',
'restrictions',
'predicted',
'for',
'northern',
'tas',
'more',
'women',
'urged',
'to',
'become',
'councillors',
'most',
'highly',
'educated',
'live',
'in',
'nsw',
'wa',
'mp',
'raises',
'hospital',
'concerns',
'in',
'parliament',
'mp',
'rejects',
'ambulance',
'levy',
'claims',
'mugabe',
'to',
'touch',
'down',
'in',
'paris',
'for',
'summit',
'national',
'gallery',
'gets',
```

```
'all',
'clear',
'after',
'nato',
'gives',
'green',
'light',
'to',
'defend',
'turkey',
'nca',
'defends',
'aboriginal',
'tent',
'embassy',
'raid',
'new',
'zealand',
'imposes',
'visa',
'entry',
'for',
'zimbabwe',
'no',
'side',
'effects',
'for',
'new',
'whooping',
'cough',
'vaccine',
'nsw',
'govt',
'under',
'fire',
'for',
'holding',
'back',
'vegetation',
'nsw',
'opp',
'defends',
'claims',
'of',
'running',
'race',
'campaign',
```

```
'nsw',
'opp',
'pledges',
'50m',
'drought',
'relief',
'nt',
'govt',
'boosts',
'nurse',
'number',
'with',
'overseas',
'intake',
'nth',
'koreans',
'seek',
'asylum',
'at',
'japanese',
'embassy',
'nursing',
'student',
'intake',
'down',
'oh',
'brother',
'your',
'times',
'up',
'says',
'ganguly',
'senior',
'omodei',
'to',
'stay',
'in',
'politics',
'onesteel',
'to',
'invest',
'80m',
'in',
'whyalla',
'steelworks',
'opposition',
'urged',
```

```
'to',
'help',
'protect',
'recherche',
'bay',
'orientation',
'begins',
'for',
'uni',
'students',
'osullivan',
'in',
'world',
'cross',
'country',
'doubt',
'pagan',
'says',
'rule',
'changes',
'not',
'necessary',
'pair',
'to',
'face',
'court',
'over',
'ayr',
'murder',
'patterson',
'defends',
'decision',
'not',
'to',
'attend',
'health',
'patterson',
'no',
'show',
'displays',
'govts',
'arrogance',
'crean',
'patterson',
'snubs',
'health',
'meeting',
```

```
'to',
'avoid',
'lions',
'den',
'peace',
'agreement',
'may',
'bring',
'respite',
'for',
'venezuela',
'pienaar',
'shines',
'as',
'ajax',
'frustrate',
'arsenal',
'plan',
'for',
'second',
'skatepark',
'plan',
'to',
'encourage',
'farmers',
'into',
'plantation',
'timber',
'png',
'nurses',
'strike',
'after',
'colleague',
'raped',
'on',
'way',
'to',
'police',
'cracking',
'down',
'on',
'driver',
'safety',
'police',
'defend',
'aboriginal',
'tent',
```

```
'embassy',
'raid',
'policewomen',
'accusations',
'feature',
'at',
'federal',
'crime',
'probe',
'launched',
'into',
'plane',
'crash',
'program',
'to',
'monitor',
'forest',
'harvested',
'areas',
'public',
'urged',
'to',
'check',
'gas',
'cylinders',
'public',
'warned',
'about',
'phone',
'scam',
'qantas',
'international',
'crews',
'to',
'strike',
'over',
'pay',
'qantas',
'war',
'plan',
'to',
'cut',
'2500',
'jobs',
'outrages',
'unions',
'qr',
```

```
'not',
'planning',
'northern',
'route',
'sackings',
'questions',
'public',
'anger',
'grows',
'after',
'korean',
'subway',
'rabbit',
'control',
'program',
'on',
'trial',
'radioactive',
'spill',
'at',
'wmcs',
'olympic',
'dam',
'mine',
'rain',
'eases',
'wheatbelt',
'water',
'woes',
'reading',
'go',
'third',
'in',
'first',
'division',
'record',
'amount',
'for',
'gladstone',
'ventures',
'refshauge',
'wins',
'defamation',
'court',
'case',
'regulator',
'to',
```

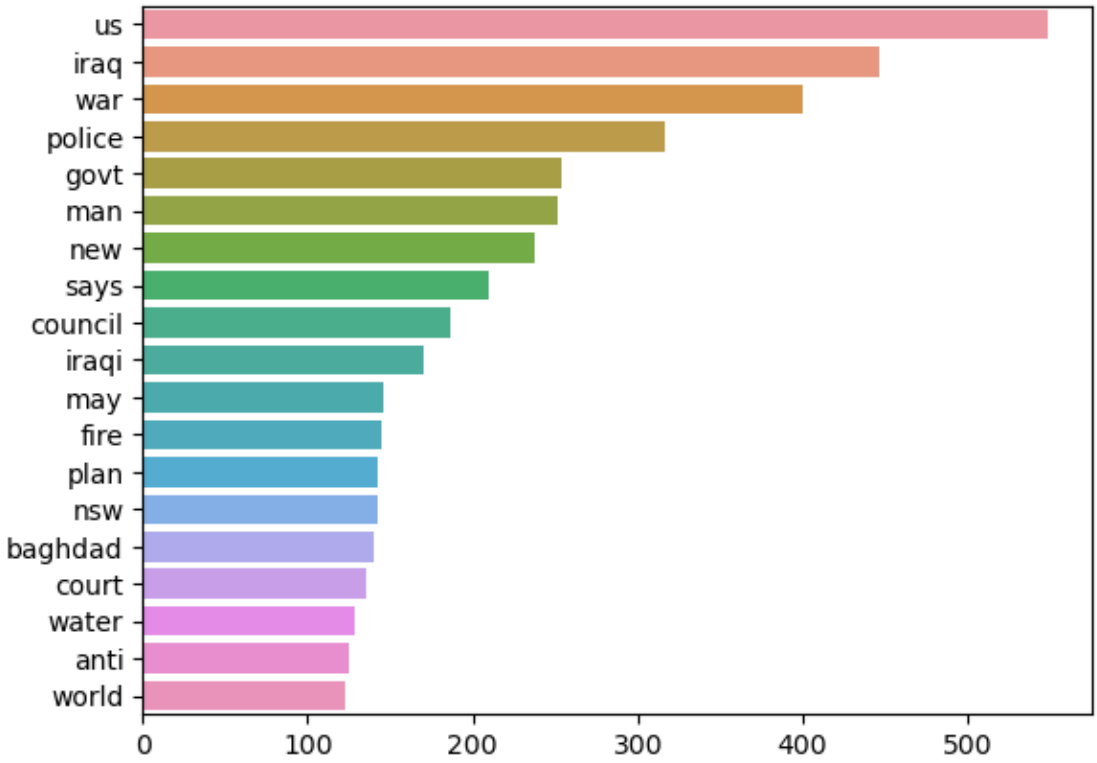
```
 'inspect',
 'gm',
 'canola',
 'trials',
 'report',
 'highlights',
 'container',
 'terminal',
 'potential',
 'resource',
 'stocks',
 'boost',
 'all',
 …]
```

[17]:
```
from collections import Counter
counter=Counter(corpus)
most=counter.most_common()

x, y= [], []
for word,count in most[:40]:
    if (word not in stop):
        x.append(word)
        y.append(count)

%pip install seaborn
import seaborn as sns
sns.barplot(x=y,y=x)
```

```
Collecting seaborn
  Downloading seaborn-0.12.2-py3-none-any.whl (293 kB)
                           293.3/293.3
kB 8.8 MB/s eta 0:00:00
Requirement already satisfied: numpy!=1.24.0,>=1.17 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from seaborn) (1.23.5)
Requirement already satisfied: pandas>=0.25 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from seaborn) (2.0.1)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from seaborn) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (1.0.7)
Requirement already satisfied: cycler>=0.10 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
```

```
(from matplotlib!=3.6.1,>=3.1->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (4.39.4)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (23.1)
Requirement already satisfied: pillow>=6.2.0 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (9.5.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from matplotlib!=3.6.1,>=3.1->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pandas>=0.25->seaborn) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from pandas>=0.25->seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.1->seaborn) (1.16.0)
Installing collected packages: seaborn
Successfully installed seaborn-0.12.2
Note: you may need to restart the kernel to use updated packages.
```

[17]: <Axes: >

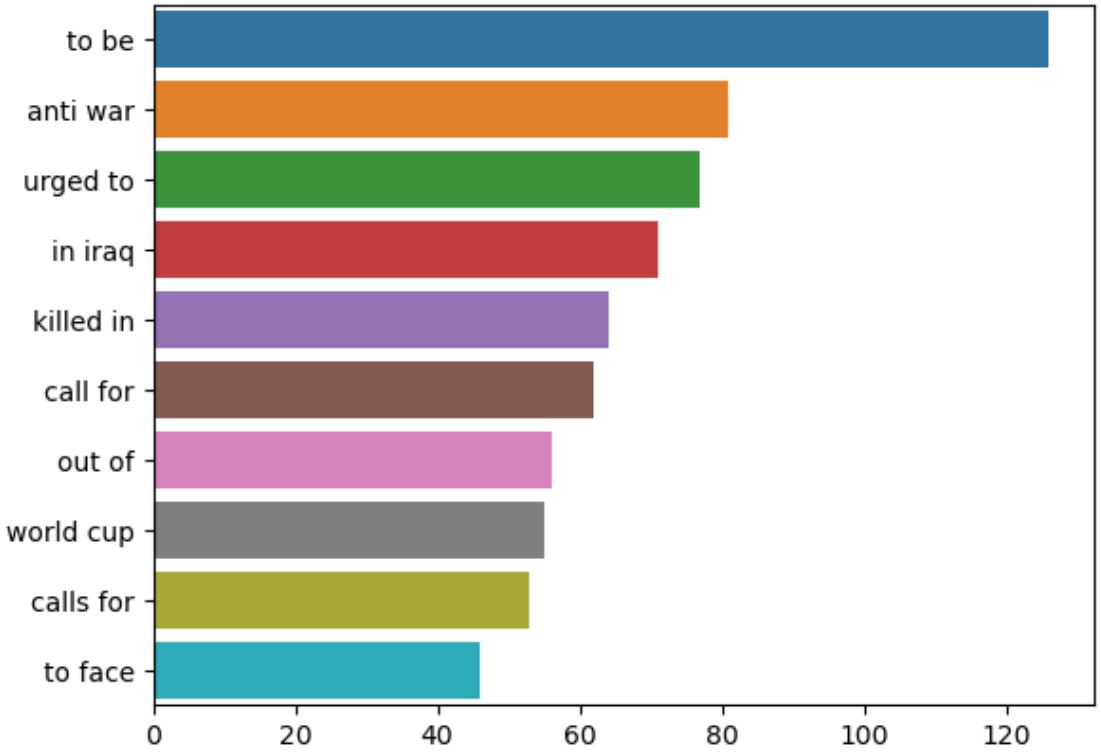


10 Here 'us' could mean either the USA or us (you and me). us is not a stopword, but when we observe other words in the graph they are all related to the US – Iraq war and "us" here probably indicate the USA.

[18]:
```
#NGRAM EXPLORATION
from nltk.util import ngrams
def get_top_ngram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(n, n)).fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx])
                  for word, idx in vec.vocabulary_.items()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:10]
```

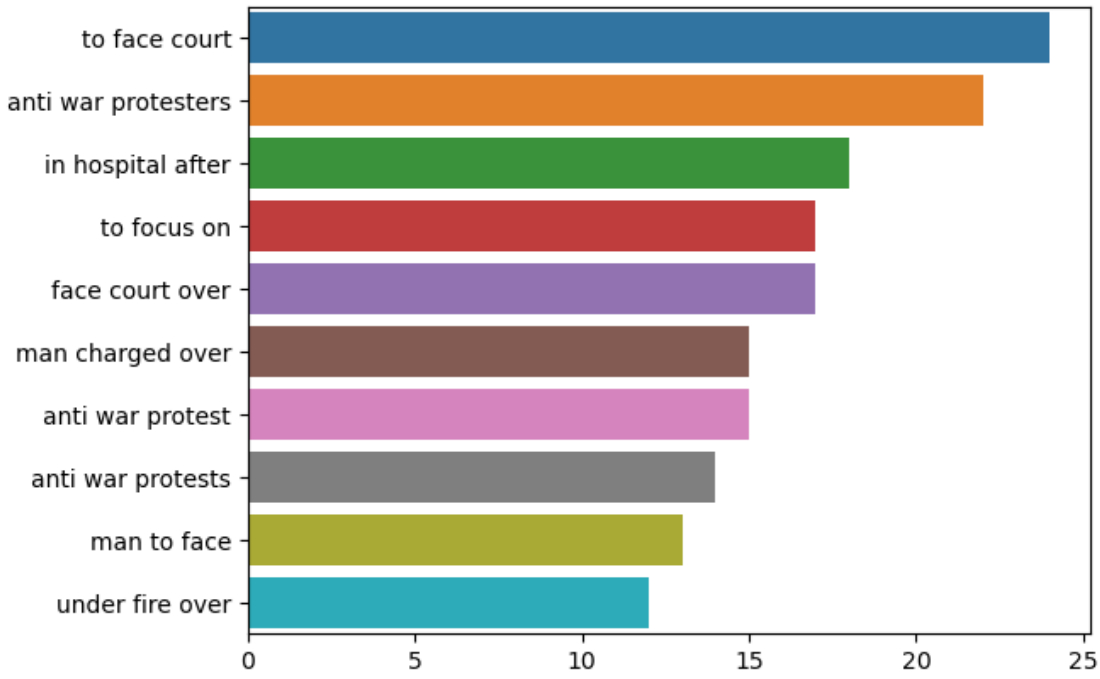
[20]:
```
from sklearn.feature_extraction.text import CountVectorizer
top_bi_grams=get_top_ngram(df['headline_text'],n=2)
x,y=map(list,zip(*top_bi_grams))
sns.barplot(x=y,y=x)
```

[20]: <Axes: >



[21]:
```
top_tri_grams=get_top_ngram(df['headline_text'],n=3)
x,y=map(list,zip(*top_tri_grams))
sns.barplot(x=y,y=x)
```

[21]: <Axes: >



[23]:
```
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
stopwords = set(STOPWORDS)

def show_wordcloud(data):
    wordcloud = WordCloud(
        background_color='white',
        stopwords=stopwords,
        max_words=100,
        max_font_size=30,
        scale=3,
        random_state=1)

    wordcloud=wordcloud.generate(str(data))

    fig = plt.figure(1, figsize=(12, 12))
    plt.axis('off')

    plt.imshow(wordcloud)
    plt.show()

show_wordcloud(corpus)
```

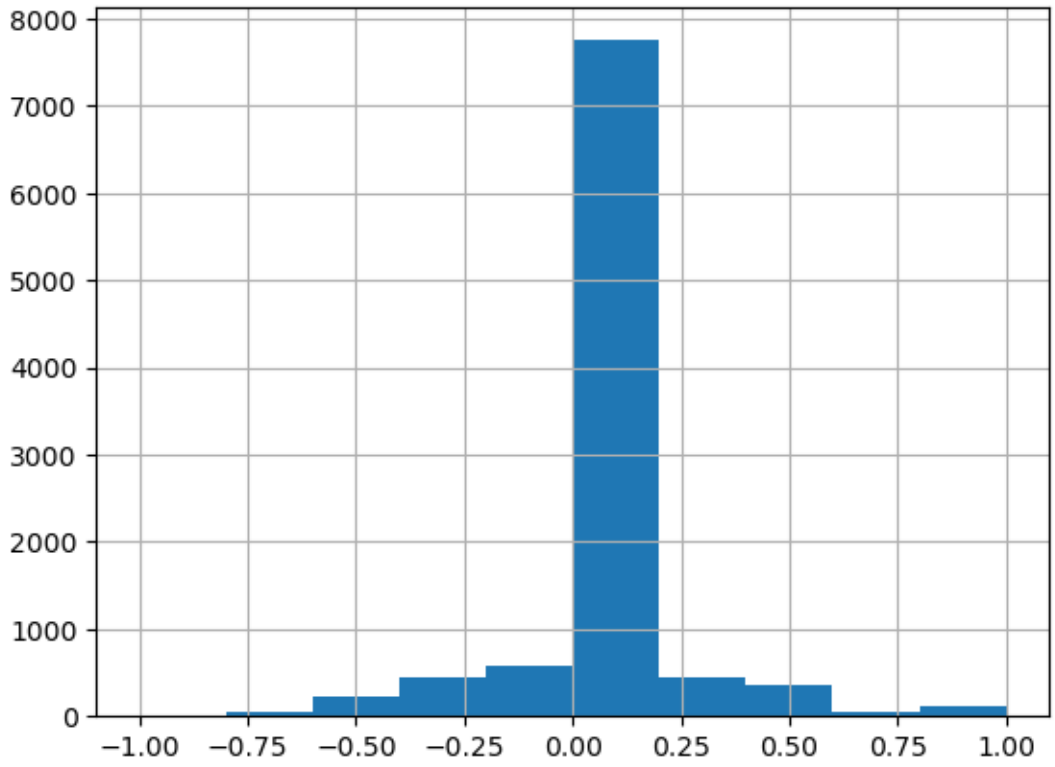
[26]:
```
#Textblob
%pip install textblob
from textblob import TextBlob
TextBlob('100 people killed in Iraq').sentiment #returns polarity[-1,1] &
 ↪subjectivity [0,1]
```

```
Collecting textblob
  Using cached textblob-0.17.1-py2.py3-none-any.whl (636 kB)
Requirement already satisfied: nltk>=3.1 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from textblob) (3.8.1)
Requirement already satisfied: click in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from nltk>=3.1->textblob) (8.1.3)
Requirement already satisfied: joblib in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from nltk>=3.1->textblob) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from nltk>=3.1->textblob) (2023.5.5)
Requirement already satisfied: tqdm in
/Users/pushpakulkarni/miniconda3/envs/tensorflow/lib/python3.10/site-packages
(from nltk>=3.1->textblob) (4.65.0)
Installing collected packages: textblob
Successfully installed textblob-0.17.1
Note: you may need to restart the kernel to use updated packages.
```

[26]:
```
Sentiment(polarity=-0.2, subjectivity=0.0)
```

[27]:
```
def polarity(text):
    return TextBlob(text).sentiment.polarity

df['polarity_score']=df['headline_text'].\
   apply(lambda x : polarity(x))
df['polarity_score'].hist()
```

[27]: <Axes: >

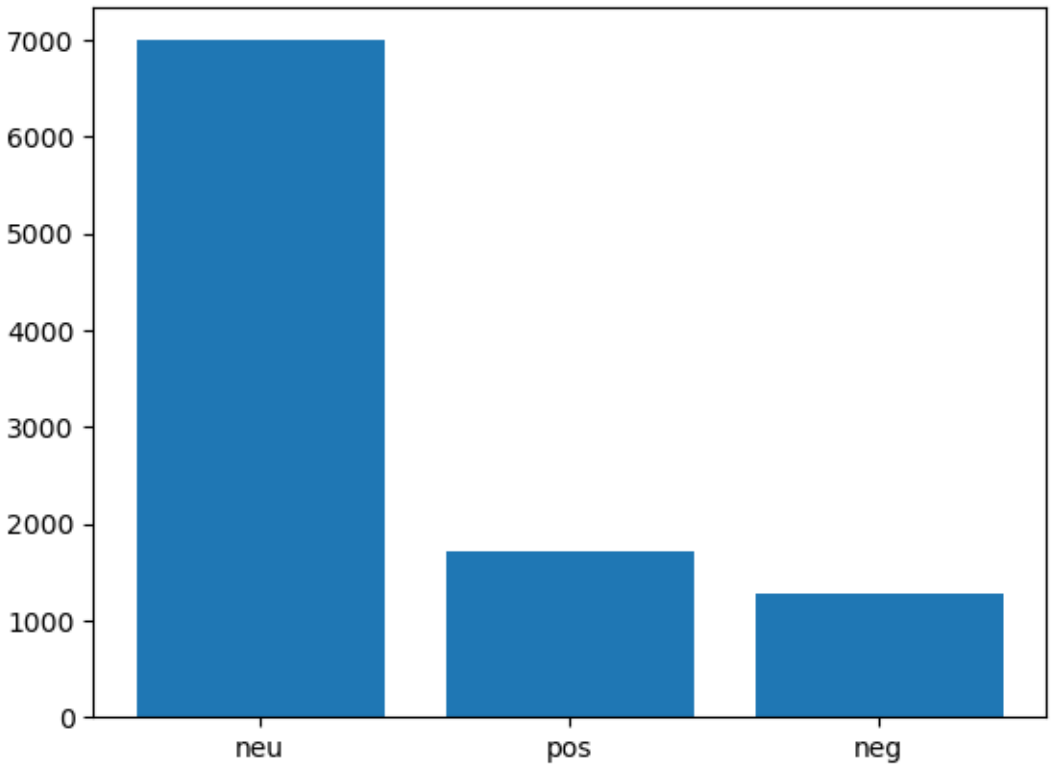


11 We can see that the polarity mainly ranges between 0.00 and 0.20. This indicates that the majority of the news headlines are neutral.

[28]:
```
# classifying the news as negative, positive and neutral based on the scores.
def sentiment(x):
    if x<0:
        return 'neg'
    elif x==0:
        return 'neu'
    else:
        return 'pos'
```

```
df['polarity']=df['polarity_score'].\
   map(lambda x: sentiment(x))

plt.bar(df.polarity.value_counts().index,
        df.polarity.value_counts())
```

[28]: <BarContainer object of 3 artists>



[29]:
```
df[df['polarity']=='pos']['headline_text'].head()
```

[29]:
```
1     act fire witnesses must be aware of defamation
5                  ambitious olsson wins triple jump
6         antic delighted with record breaking barca
18        bryant leads lakers to double overtime win
26      commonwealth bank cuts fixed home loan rates
Name: headline_text, dtype: object
```

[30]:
```
df[df['polarity']=='neg']['headline_text'].head()
```

[30]:
```
7     aussie qualifier stosur wastes four memphis match
23              carews freak goal leaves roma in ruins
```

```
28     council chief executive fails to secure position
34                    dargo fire threat expected to rise
40        direct anger at govt not soldiers crean urges
Name: headline_text, dtype: object
```

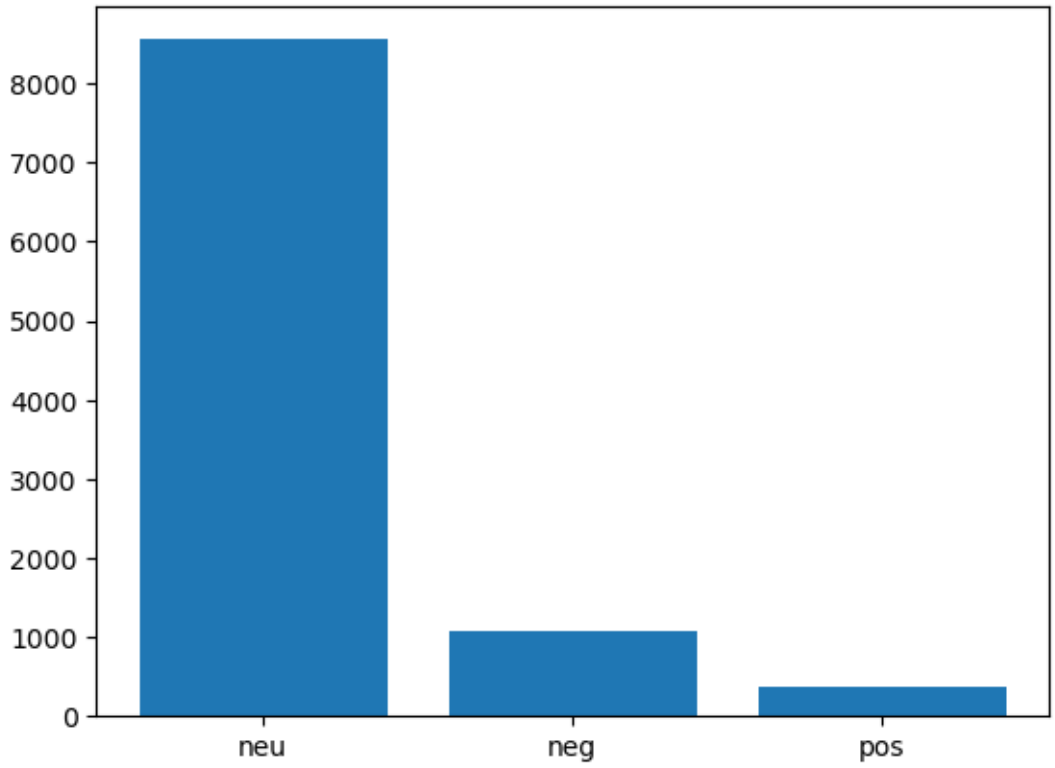
12 Vader sentiment analysis

VADER or Valence Aware Dictionary and Sentiment Reasoner is a rule/lexicon-based, open-source sentiment analyzer pre-built library, VADER sentiment analysis class returns a dictionary that contains the probabilities of the text for being positive, negative and neutral. Then we can filter and choose the sentiment with most probability.

[31]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer

nltk.download('vader_lexicon')
sid = SentimentIntensityAnalyzer()

def get_vader_score(sent):
    # Polarity score returns dictionary
    ss = sid.polarity_scores(sent)
    #return ss
    return np.argmax(list(ss.values())[:-1])

df['polarity']=df['headline_text'].\
    map(lambda x: get_vader_score(x))
polarity=df['polarity'].replace({0:'neg',1:'neu',2:'pos'})

plt.bar(polarity.value_counts().index,
        polarity.value_counts())
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/pushpakulkarni/nltk_data…
[nltk_data]   Package vader_lexicon is already up-to-date!
```

[31]:
```
<BarContainer object of 3 artists>
```



13 Named Entity Recognition

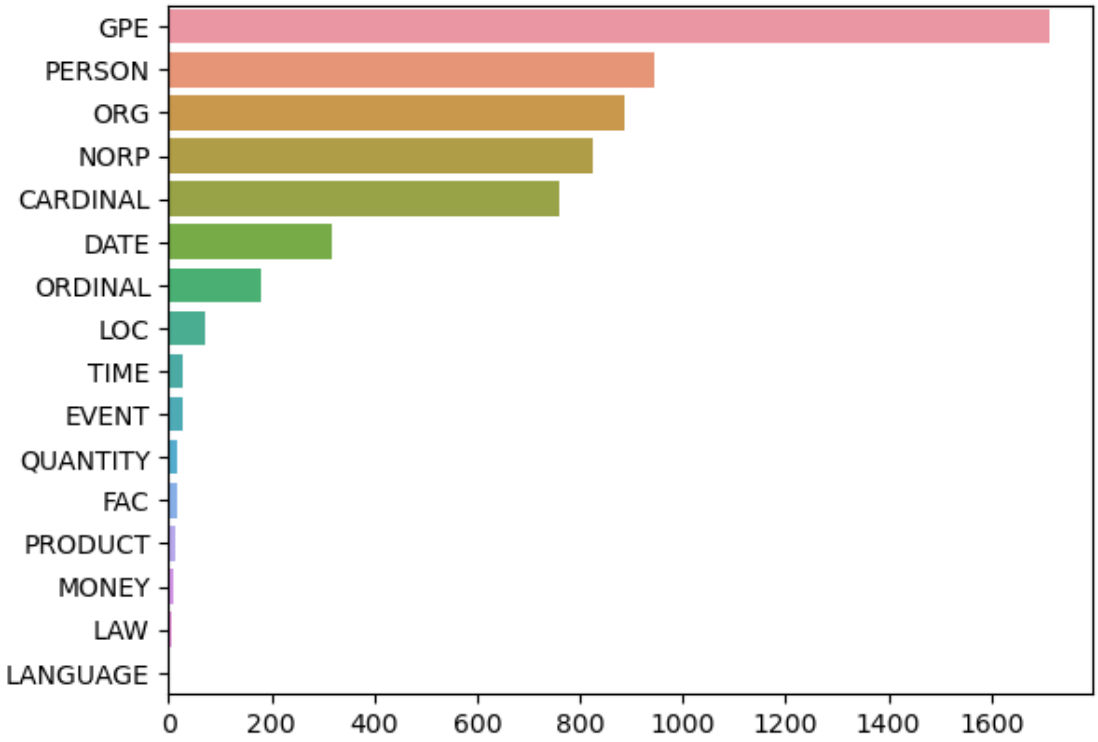
Named entity recognition is an information extraction method in which entities that are present in the text are classified into predefined entity types like “Person”,” Place”,” Organization”, etc. By using NER we can get great insights about the types of entities present in the given text dataset.

[32]:
```
import spacy

nlp = spacy.load("en_core_web_sm")
```

[33]:
```
def ner(text):
    doc=nlp(text)
    return [X.label_ for X in doc.ents]

ent=df['headline_text'].\
    apply(lambda x : ner(x))
ent=[x for sub in ent for x in sub]

counter=Counter(ent)
count=counter.most_common()
```

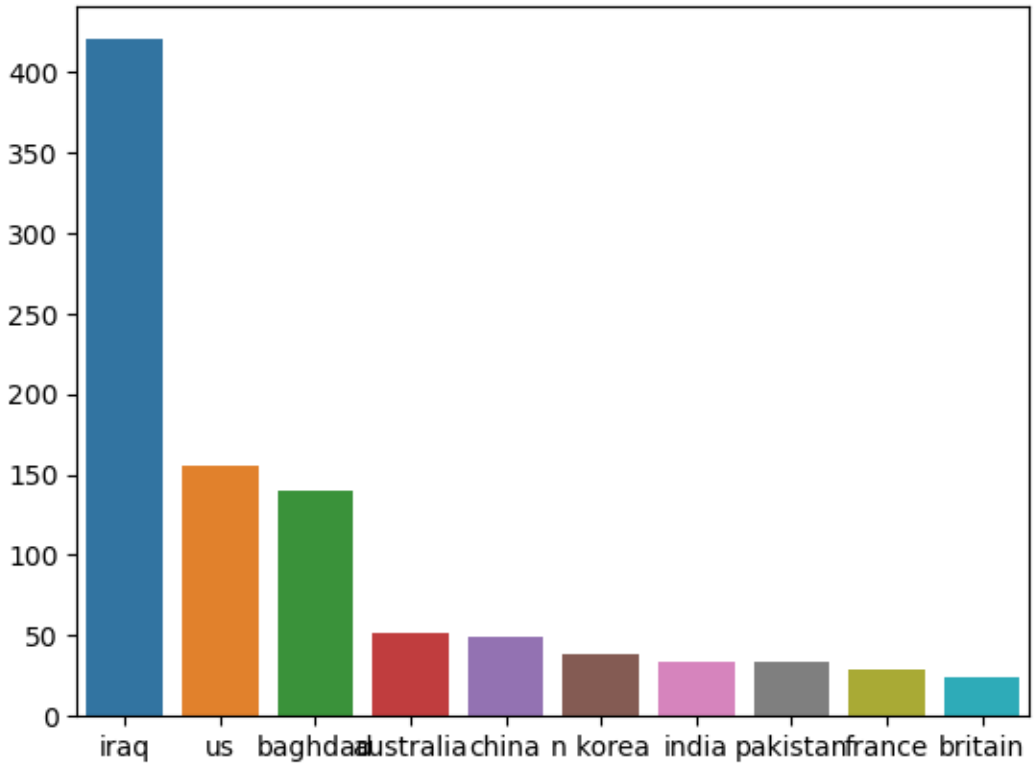
[34]:
```
x,y=map(list,zip(*count))
sns.barplot(x=y,y=x)
```

[34]: <Axes: >



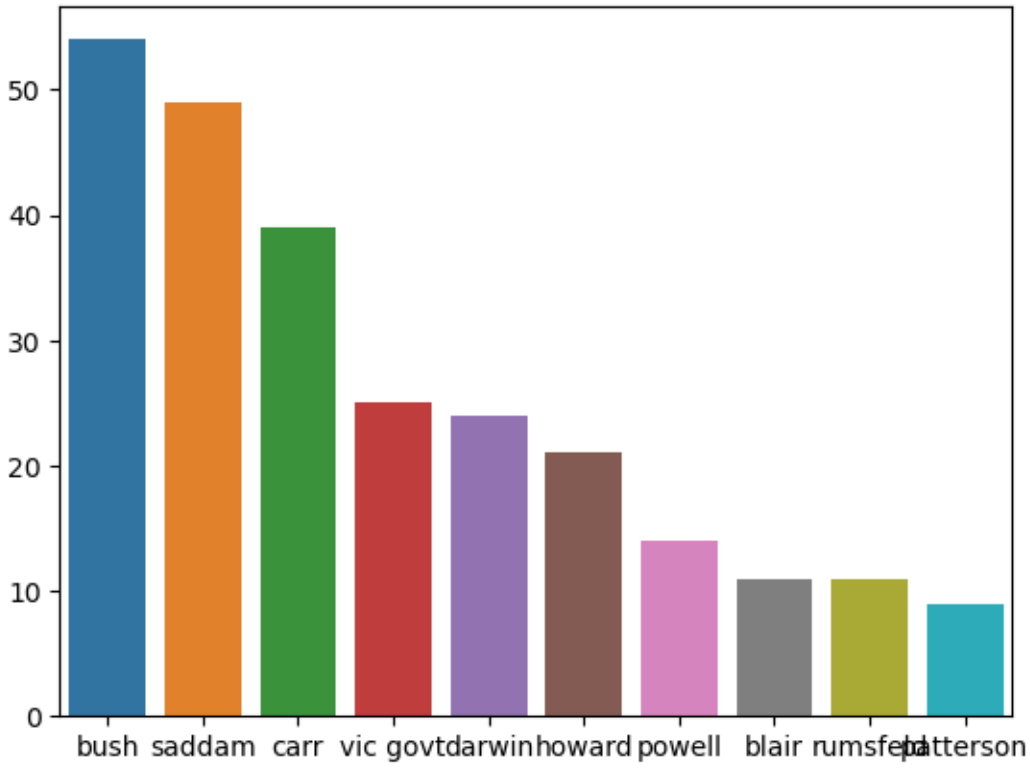
[36]:
```
#visualizing the most common tokens per entity.
def ner(text,ent="GPE"):
    doc=nlp(text)
    return [X.text for X in doc.ents if X.label_ == ent]

gpe=df['headline_text'].apply(lambda x: ner(x))
gpe=[i for x in gpe for i in x]
counter=Counter(gpe)

x,y=map(list,zip(*counter.most_common(10)))
sns.barplot(y = y,x = x)
```

[36]: <Axes: >



[38]:
```
per=df['headline_text'].apply(lambda x: ner(x,"PERSON"))
per=[i for x in per for i in x]
counter=Counter(per)

x,y=map(list,zip(*counter.most_common(10)))
sns.barplot(y = y,x = x)
```

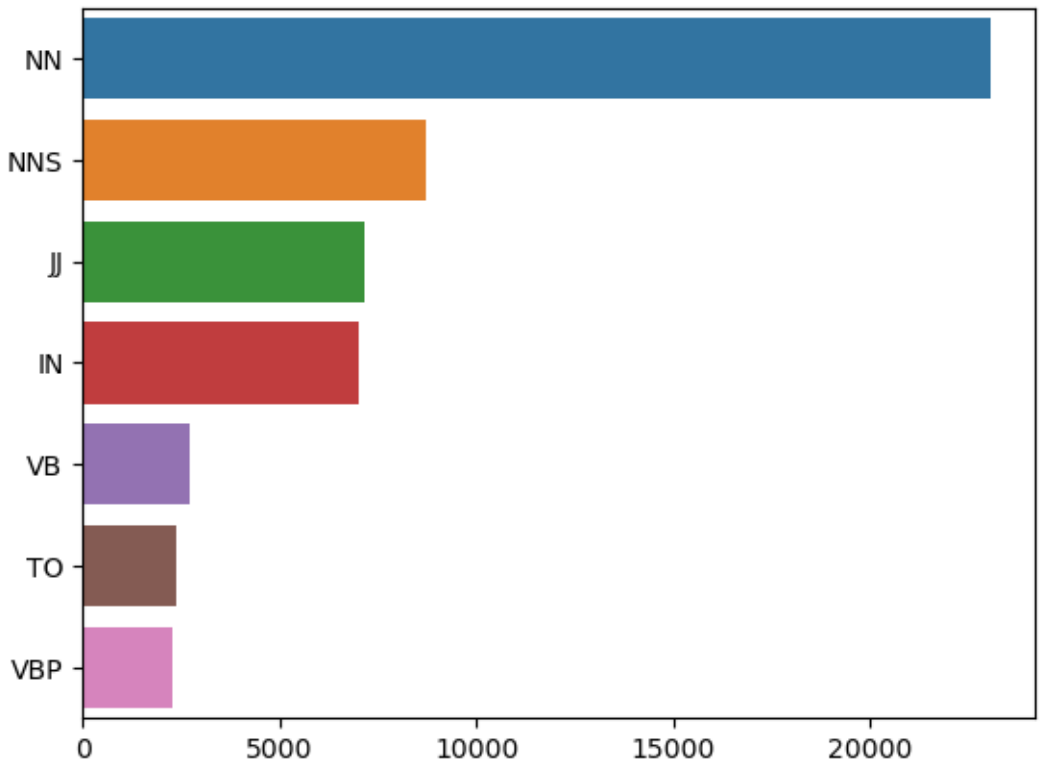
[38]: <Axes: >



[39]:
```
#POS TAGGING
nltk.download('averaged_perceptron_tagger')
from nltk import pos_tag
from nltk.tokenize import word_tokenize
def pos(text):
    pos=nltk.pos_tag(word_tokenize(text))
    pos=list(map(list,zip(*pos)))[1]
    return pos

tags=df['headline_text'].apply(lambda x : pos(x))
tags=[x for l in tags for x in l]
counter=Counter(tags)

x,y=list(map(list,zip(*counter.most_common(7))))
sns.barplot(x=y,y=x)
```

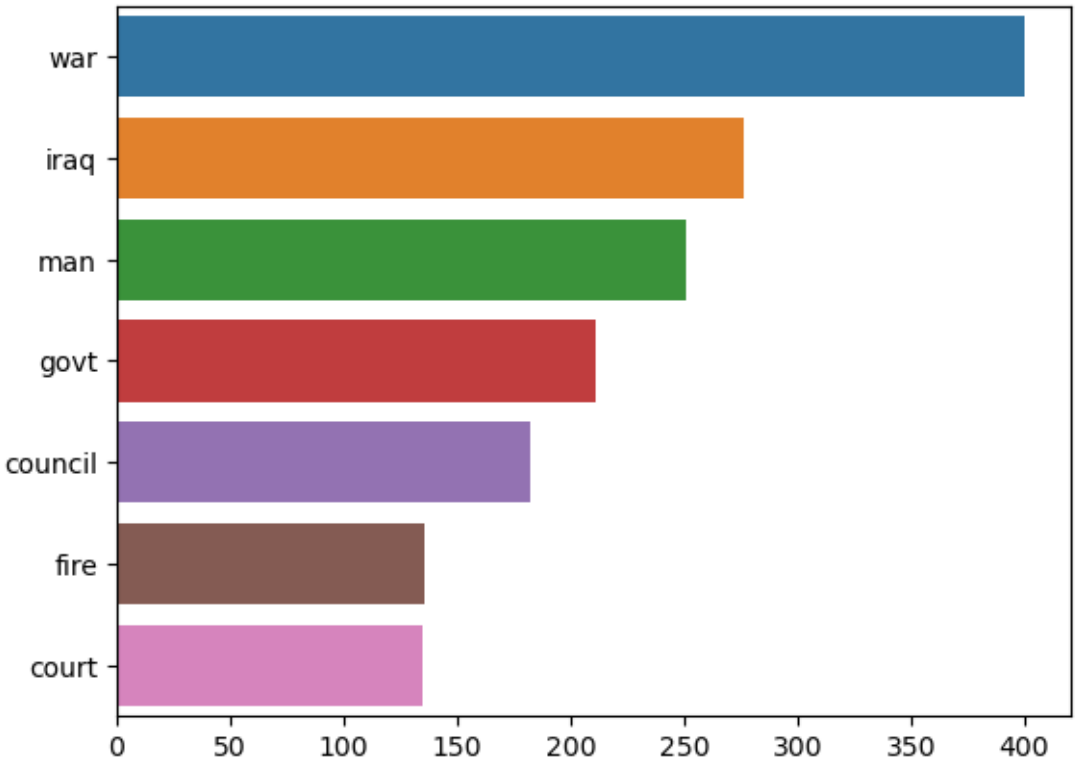
```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/pushpakulkarni/nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

[39]: <Axes: >



[40]:
```
#which singular noun occur most commonly in news headlines.
def get_adjs(text):
    adj=[]
    pos=nltk.pos_tag(word_tokenize(text))
    for word,tag in pos:
        if tag=='NN':
            adj.append(word)
    return adj


words=df['headline_text'].apply(lambda x : get_adjs(x))
words=[x for l in words for x in l]
counter=Counter(words)

x,y=list(map(list,zip(*counter.most_common(7))))
sns.barplot(x=y,y=x)
```

[40]: <Axes: >



14 Exploring through text complexity

It can be very informative to know how readable (difficult to read) the text is and what type of reader can fully understand it. Do we need a college degree to understand the message or a first-grader can clearly see what the point is?

There are many readability score formulas available for the English language like:

ARI = 4.71 * (characters/words) + 0.5 * (words/sentence) -21.43

FRE = 206.835 − 1.015 * (total words/total sentences) − 84.6 * (total syllables/ total words)

FKGL = 0.39 * (total words/ totalsentences) + 11.8 (total syllables/total words) -15.59

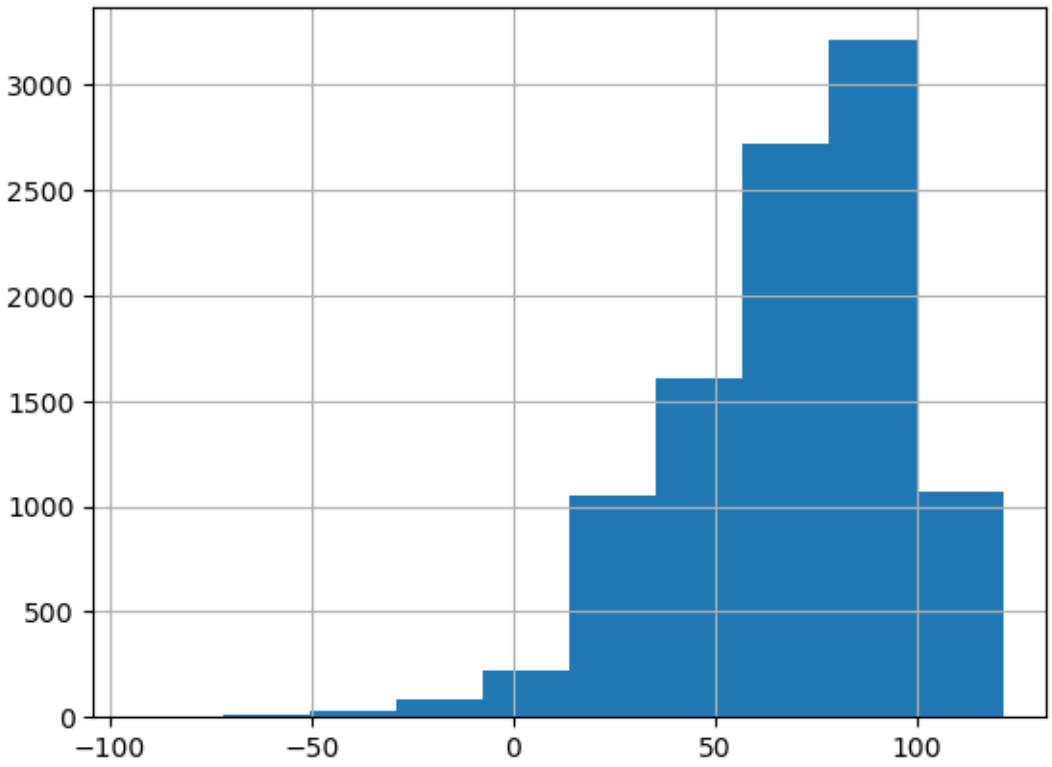
GFI = 0.4 * ((words/ sentence) + 100 * (complex words/ words))

[42]:
```
%pip install textstat
from textstat import flesch_reading_ease

df['headline_text'].\
   apply(lambda xi : flesch_reading_ease(xi)).hist()
```

```
Collecting textstat
  Using cached textstat-0.7.3-py3-none-any.whl (105 kB)
```

```
Collecting pyphen (from textstat)
  Using cached pyphen-0.14.0-py3-none-any.whl (2.0 MB)
Installing collected packages: pyphen, textstat
Successfully installed pyphen-0.14.0 textstat-0.7.3
Note: you may need to restart the kernel to use updated packages.
```

[42]: <Axes: >



[43]:
```
li=[]
li=(df['headline_text'].apply(lambda xi : flesch_reading_ease(xi)))
```

15 Almost all of the readability scores fall above 60. This means that an average 11-year-old student can read and understand the news headlines. Let's check all news headlines that have a readability score below 5.

[44]:
```
x=[i for i in range(len(li)) if li[i]<5]
df.iloc[x]['headline_text'].head()
```

[44]:
```
134    policewomen accusations feature at federal crime
150      report highlights container terminal potential
285       groups praise outgoing opposition agriculture
298            investigations underway into qantas skid
308      landholder contribution still under discussion
Name: headline_text, dtype: object
```

16 Conclusion:

1) EDA (Exploratory Data Analysis) is an important step to understand the data and its characteristics so that proper pre-processing can be done before making the model.

2) We can see that even though negative news is less, but its affect is more than the neutral and positive ones.