```
Start coding or generate with AI.
```

## Part 1) Aim:

For user to pick any one person from list as input and output the 10 other people who's overview are "closest" to the person in a Natural Language Processing sense. Also output the sentiment of the overview of the person

- Load the CSV file: df named "step1"

1. Data Cleaning
2. Bag of Words Using CountVectorizer: to compute similarity between overviews ("text")
3. TD-IDF: to find closest overviews based on similarity
4. K Nearest Neighbors: to analyze sentiment of specified overview
5. Output sentiment

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords, wordnet
nltk.download('stopwords')

# Set up stop words
stop_words = set(stopwords.words('english'))

from nltk.sentiment.vader import SentimentIntensityAnalyzer
# nltk.download('omw-1.4')
nltk.download('punkt_tab')
nltk.download('averaged_perceptron_tagger_eng')

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# nlkt stop words and wordnet (lemmatizer)

pd.options.display.max_columns = 100




```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
```

```
%%capture
# Install textblob
!pip install -U textblob
from textblob import TextBlob
```

```
%%capture
# Download corpora
!python -m textblob.download_corpora
```

```
url= "https://ddc-datascience.s3.amazonaws.com/Projects/Project.5-NLP/Data/NLP.csv"
nlp = pd.read_csv(url)
step1= nlp.copy()
```

```
nlp.info()
step1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42786 entries, 0 to 42785
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
```

```
---  ------  --------------  -----
 0   URI     42786 non-null  object
 1   name    42786 non-null  object
 2   text    42786 non-null  object
dtypes: object(3)
memory usage: 1002.9+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42786 entries, 0 to 42785
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  -----
 0   URI     42786 non-null  object
 1   name    42786 non-null  object
 2   text    42786 non-null  object
dtypes: object(3)
memory usage: 1002.9+ KB
```

Cell content is object/string data type.

"URI" column appears to be an ID and will not be used for vectorization, use for Part 2

"text" column to be used for vectorization

```
step1.head()
```

| | URI | name | text |
|---|---|---|---|
| 0 | <http://dbpedia.org/resource/Digby_Morrell> | Digby Morrell | digby morrell born 10 october 1979 is a former... |
| 1 | <http://dbpedia.org/resource/Alfred_J._Lewy> | Alfred J. Lewy | alfred j lewy aka sandy lewy graduated from un... |
| 2 | <http://dbpedia.org/resource/Harpdog_Brown> | Harpdog Brown | harpdog brown is a singer and harmonica player... |
| 3 | <http://dbpedia.org/resource/Franz_Rottensteiner> | Franz Rottensteiner | franz rottensteiner born in waidmannsfeld lowe... |
| 4 | <http://dbpedia.org/resource/G-Enka> | G-Enka | henry krvits born 30 december 1974 in tallinn ... |

Next steps: Generate code with step1 View recommended plots New interactive sheet

Shape: 3 columns, 42k+rows

Zero nulls, Zero duplicates

Object key. However, cell content is string

```
print(step1.shape)
print(step1.columns)

print(step1.nunique())
print(step1.duplicated().value_counts())

print(step1.isnull().sum())
```

```
(42786, 3)
Index(['URI', 'name', 'text'], dtype='object')
URI     42786
name    42785
text    42786
dtype: int64
False    42786
Name: count, dtype: int64
URI     0
name    0
text    0
dtype: int64
```

I decided to follow one file to run various data cleaning methods and attributes. I used iloc[0] to select the 1st person on the list

```
#RETURN all contents for row 1, column 2 ('text')
pd.set_option("display.max_colwidth", None)  # Set max column width to unlimited
print(step1.iloc[0]['text'])
```

```
digby morrell born 10 october 1979 is a former australian rules footballer who played with the kangaroos and carlton in the australian f
```

I wanted to see the contents in their entirety to see changes after each method.

I saw lots of filler words like "is, for, and" etc

```
#RETURN all contents for row 1, column 2 ('text') but wrapped
import textwrap

pd.set_option("display.max_colwidth", None)  # Set maximum column width to unlimited

text_0 = step1.loc[step1.index[0], 'text']
wrapped_text0 = textwrap.fill(text, width=80)  # Wrap text to 80 characters wide
print(wrapped_text0)
```

```
harpdog brown is a singer and harmonica player who has been active in canadas
blues scene since 1982 hailing from vancouver he crossed tens of thousands of
miles playing club dates and festivals in canada the northwestern united states
and germanyover the years he has issued seven cds in 1995 his home is where the
harp is won the muddy award for the best nw blues release from the cascade blues
association in portland oregon as well that year it was nominated for a canadian
juno for the best bluesgospel recording teamed up with graham guest on piano his
cd naturally was voted 1 canadian blues album of 2010 by the blind lemon
surveybrown tours extensively with his guitarist j arthur edmonds performing
their electric mid1950s chicago blues either as a duo or with the full band
while he is home he juggles a few combos working many venues big and small he
also leads the harpdog brown band which is a gutsy traditional chicago blues
band in 2014 they released what it is comprising mainly original songs and a few
classic covers influential blues promoter and broadcaster holger petersen called
what it is browns best albumhe was just awarded the maple blues award in toronto
for best harmonica player in canada 2014 and was honored with a life time
membership to the hamilton blues society
```

## Data Cleaning

I ran each method individually to see the changes.

All can be compiled into one code cell.

```
#Remove special characters, make lower case, remove stop words
```

After applying lower case method multiple \n appeared in output, indicating there was a break in line and these would have to be removed.

```
#make lower case
wrapped_text0.lower()
```

```
'harpdog brown is a singer and harmonica player who has been active in canadas\nblues scene since 1982 hailing from vancouver he crosse
d tens of thousands of\nmiles playing club dates and festivals in canada the northwestern united states\nand germanyover the years he h
as issued seven cds in 1995 his home is where the\nharp is won the muddy award for the best nw blues release from the cascade blues\nas
sociation in portland oregon as well that year it was nominated for a canadian\njuno for the best bluesgospel recording teamed up with
graham guest on piano his\ncd naturally was voted 1 canadian blues album of 2010 by the blind lemon\nsurveybrown tours extensively with
his guitarist j arthur edmonds performing\ntheir electric mid1950s chicago blues either as a duo or with the full band\nwhile he is hom
e he juggles a few combos working many venues big and small he\nalso leads the harpdog brown band which is a gutsy traditional chicago
```

Stop words were removed, including the \n.

This variable works backward-- it's telling the words to join back together after they have been split AND only if they are not stop words.

It also includes turning the string to lower case.

```
# Remove stop words from wrapped_text
wrapped_text_stop_lower = ' '.join([word for word in wrapped_text0.split() if word.lower() not in stop_words])
wrapped_text_stop_lower
```

```
'harpdog brown singer harmonica player active canadas blues scene since 1982 hailing vancouver crossed tens thousands miles playing clu
b dates festivals canada northwestern united states germanyover years issued seven cds 1995 home harp muddy award best nw blues release
cascade blues association portland oregon well year nominated canadian juno best bluesgospel recording teamed graham guest piano cd nat
urally voted 1 canadian blues album 2010 blind lemon surveybrown tours extensively guitarist j arthur edmonds performing electric mid19
50s chicago blues either duo full band home juggles combos working many venues big small also leads harpdog brown band gutsy traditiona
l chicago blues band 2014 released comprising mainly original songs classic covers influential blues promoter broadcaster holger peters
en called browns best albumhe awarded maple blues award toronto best harmonica player canada 2014 honored life time membership hamilton
```

```
#Turn string into text blob for further cleaning
wrapped_textblob = TextBlob(wrapped_text_stop_lower)
```

```
#Output sentiment
wrapped_sentiment = wrapped_textblob.sentiment
wrapped_sentiment
```

```
Sentiment(polarity=0.30722222222222223, subjectivity=0.4277777777777778)
```

```
#RETURN string as list of words
words = wrapped_textblob.words
words
```

```
WordList(['harpdog', 'brown', 'singer', 'harmonica', 'player', 'active', 'canadas', 'blues', 'scene', 'since', '1982', 'hailing',
'vancouver', 'crossed', 'tens', 'thousands', 'miles', 'playing', 'club', 'dates', 'festivals', 'canada', 'northwestern', 'united',
'states', 'germanyover', 'years', 'issued', 'seven', 'cds', '1995', 'home', 'harp', 'muddy', 'award', 'best', 'nw', 'blues', 'release',
'cascade', 'blues', 'association', 'portland', 'oregon', 'well', 'year', 'nominated', 'canadian', 'juno', 'best', 'bluesgospel',
'recording', 'teamed', 'graham', 'guest', 'piano', 'cd', 'naturally', 'voted', '1', 'canadian', 'blues', 'album', '2010', 'blind',
'lemon', 'surveybrown', 'tours', 'extensively', 'guitarist', 'j', 'arthur', 'edmonds', 'performing', 'electric', 'mid1950s', 'chicago',
'blues', 'either', 'duo', 'full', 'band', 'home', 'juggles', 'combos', 'working', 'many', 'venues', 'big', 'small', 'also', 'leads',
'harpdog', 'brown', 'band', 'gutsy', 'traditional', 'chicago', 'blues', 'band', '2014', 'released', 'comprising', 'mainly', 'original',
'songs', 'classic', 'covers', 'influential', 'blues', 'promoter', 'broadcaster', 'holger', 'petersen', 'called', 'browns', 'best',
'albumhe', 'awarded', 'maple', 'blues', 'award', 'toronto', 'best', 'harmonica', 'player', 'canada', '2014', 'honored', 'life', 'time',
'membership', 'hamilton', 'blues', 'society'])
```

Applying noun phrases method returned the string in neat noun phrases.

Like looking for themes.

Most years were not returned.

```
noun_phrases = wrapped_textblob.noun_phrases
noun_phrases
```

```
WordList(['brown singer harmonica player', 'active canadas blues scene', 'tens thousands miles', 'club dates festivals canada
northwestern', 'states germanyover years', 'home harp muddy award', 'nw blues release cascade blues association portland oregon',
'canadian juno', 'graham guest piano cd', 'canadian blues album', 'blind lemon surveybrown tours', 'guitarist j arthur edmonds',
'electric mid1950s chicago blues', 'full band home juggles combos', 'brown band gutsy', 'traditional chicago blues band', 'original
songs', 'influential blues promoter broadcaster holger petersen', 'maple blues', 'harmonica player canada', 'life time membership
hamilton blues society'])
```

re.sub is a function from the re module (regular expressions) in Python which finds and replaces patterns within strings

I asked it to find and remove [^A-Za-z0-9\w\s] vs [^\w\s] because the first sweeps for and removes EVERYTHING that is not a string in upper/lower case or a number, a general word, and any whitespace. The second option does the same but only general word search and can leave things embedded in text such as _underscores.

```
For example - wild_horses would remain the same using [^\w\s]
```

```
# Remove special characters which are not alphanumeric and whitespace

special = re.sub(r'[^A-Za-z0-9\w\s]', '', wrapped_text_stop_lower)
special
```

```
'harpdog brown singer harmonica player active canadas blues scene since 1982 hailing vancouver crossed tens thousands miles playing clu
b dates festivals canada northwestern united states germanyover years issued seven cds 1995 home harp muddy award best nw blues release
cascade blues association portland oregon well year nominated canadian juno best bluesgospel recording teamed graham guest piano cd nat
urally voted 1 canadian blues album 2010 blind lemon surveybrown tours extensively guitarist j arthur edmonds performing electric mid19
50s chicago blues either duo full band home juggles combos working many venues big small also leads harpdog brown band gutsy traditiona
l chicago blues band 2014 released comprising mainly original songs classic covers influential blues promoter broadcaster holger peters
en called browns best albumhe awarded maple blues award toronto best harmonica player canada 2014 honored life time membership hamilton
```

Will did it all in one code cell

```
# def process_data (text):
#   '''
#   This function will take a value and apply the lower case method, remove white spaces and characters and convert the text to TextBlob
#   '''
#   text = text.lower()
#   text = re.sub(r"[^\w\s]", "", text)
#   text = TextBlob(text)
#   return text
```

```
# df['text'] = df['text'].apply(process_data)
```