NLP Assignment #3 - Luke Schwenke

In [ ]:

```
import pandas as pd
import nltk
from nltk.util import ngrams
from nltk.tokenize import word_tokenize
from scipy.spatial.distance import jaccard
from nltk import TweetTokenizer
from nltk.stem import WordNetLemmatizer

pd.set_option('display.max_rows', 100)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', 500)
```

Determine which news articles (news_df) are similar to each other and which tweets (tweets_df) are more similar to each other. In order to accomplish this you need to create n-grams and compare the similarity of the text using Jaccard distance.

Read news data

In [ ]:

```
news_path = 'https://storage.googleapis.com/msca-bdp-data-open/news/nlp_a_3_
news_df = pd.read_json(news_path, orient='records', lines=True, encoding='ut

print(f'Sample contains {news_df.shape[0]:,.0f} news articles')
news_df.head(2)
```

```
Sample contains 1,018 news articles
```

Out[ ]:

	url	date	language	title	t
0	https://auto.hindustantimes.com/lml-bikes/dealers/bodh-gaya	2022-01-21	en	Lml Bikes Car Dealers - Lml Bikes Showrooms in India	Lml Bikes Car Dealers - Bikes Showrooms in Ir Explore Friday, 21 Janu 2022 Log in/Sign SearchNotifications SectionsAuto News NewsBike NewsLatestA NewsPhotosVideosElec VehiclesTrending ReadsOffersnewF carsFind bikesComp carsCompare bikesI calculatorDealersExpl AutoAbout UsCont UsSITEMAPRSSTerms UsePrivacy PolicyCopyri © HT Media Limited All rig reserved.HomeOffersnewF carsFind bikesComp carsCompare bikesI calculatorDealers NewsB
1	https://auto.hindustantimes.com/pure-ev-bikes/dealers/avadi	2022-01-21	en	Pure Ev Bikes Car Dealers - Pure Ev Bikes Showrooms in India	Pure Ev Bikes Car Deale Pure Ev Bikes Showroom India Explore Friday January 2022 Log in/Sign SearchNotifications SectionsAuto News NewsBike NewsLatestA NewsPhotosVideosElec VehiclesTrending ReadsOffersnewF carsFind bikesComp carsCompare bikesI calculatorDealersExpl AutoAbout UsCont UsSITEMAPRSSTerms UsePrivacy PolicyCopyri © HT Media Limited All rig reserved.HomeOffersnewF carsFind bikesComp carsCompare bikesI calculatorDealersCa

Read Tweets data

In [ ]:

```
tweets_path = 'https://storage.googleapis.com/msca-bdp-data-open/tweets/nlp_
tweets_df = pd.read_json(tweets_path, orient='records', lines=True)
print(f'Sample contains {tweets_df.shape[0]:,.0f} tweets')
tweets_df.head(2)
```

```
Sample contains 1,020 tweets
```

Out[ ]:

	id	lang	date	name	retweeted	text
0	1491880241782005777	en	2022-02-10	Sir Botalot dropping the Mike!		@singervehicles Will you do a straight swap for my Renault Twingo 1.2 Extreme? \n\nThat is awesome! 🔥
1	1501237946590142469	en	2022-03-08	Sabrina Ghio	RT	Australian GP Qualifying, Melbourne, 8th March 1997. #F1 The Grid ⬇️\n\nRow 3:\n5. Eddie Irvine (Ferrari) +2.512s\n6. Mika Hakkinen (McLaren-Mercedes) +2.602s\n\nRow 4:\n7. Johnny Herbert (Sauber-Petronas) +2.918s\n8. Jean Alesi (Benetton-Renault) +3.224s https://t.co/i4IC4LzVVN

Cleaning / Preparation

In [ ]:

```
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

import re
def clean_text(text):
    # Remove mentions
    text = re.sub(r'@[A-Za-z0-9_]+', '', text)
    # Remove hashtags (but keep the text after #)
    text = re.sub(r'#', '', text)
    # Remove RT (retweet symbol)
    text = re.sub(r'RT[\s]+', '', text)
    # Remove hyperlinks
    text = re.sub(r'https?:\/\/\S+', '', text)
    # Remove newline characters
    text = re.sub(r'\n', ' ', text)
    # Remove carriage return characters
    text = re.sub(r'\r', '', text)
    # Remove "&amp;"
    text = re.sub(r'&amp;', '', text)
    # Remove other special characters and numbers
    text = re.sub(r'[^A-Za-z\s]', '', text)
    # Optionally, remove single characters (mostly left from removing mentic
    text = re.sub(r'\s+[a-zA-Z]\s+', ' ', text)
    # Convert multiple spaces to a single space
    text = re.sub(r'\s+', ' ', text)
    # Optionally, convert to lowercase
    text = text.lower()
    # Remove stopwords
    text = ' '.join([word for word in text.split() if word not in stop_words

    return text.strip()
```

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

In [ ]:

```
tweets_df['text_clean'] = tweets_df['text'].apply(clean_text)
news_df['text_clean'] = news_df['text'].apply(clean_text)
news_df['title_clean'] = news_df['title'].apply(clean_text)
```

In [ ]:

```
# Examine a cleaned sample tweet
tweets_df['text_clean'][0]
```

Out[ ]:

```
'straight swap renault twingo extreme awesome'
```

In [ ]:

```
# Examine a cleaned sample news article
news_df['text_clean'][0][0:100]
```

Out[ ]:

```
'lml bikes car dealers lml bikes showrooms india explore friday january log
insign searchnotification'
```

In [ ]:

```
news_df['title_clean'][0]
```

Out[ ]:

```
'lml bikes car dealers lml bikes showrooms india'
```

In [ ]:

```
# Create clean dataset copies
news_df_c = news_df
tweets_df_c = tweets_df
```

Define a function that will create n number of ngram columns

In [ ]:

```
def create_ngrams(text, n=10):
    n_grams = list(nltk.ngrams(text, n))
    return n_grams
```

Section #1: Tweets

Start by tokenizing the tweets

In [ ]:

```
def tweet_tokenize(text):
    tokenizer = TweetTokenizer()
    tokens = tokenizer.tokenize(text)
    return tokens
```

Lemmatize the tokens to the root word

In [ ]:

```
tweets_df_c['tokens'] = tweets_df_c['text_clean'].apply(tweet_tokenize)

# Lemmatize tokens
lemmatizer = WordNetLemmatizer()
tweets_df_c['tokens'] = tweets_df_c['tokens'].apply(lambda tokens: [lemmatiz

tweets_df_c.head(5)
```

Out[ ]:

	id	lang	date	name	retweeted	text	
0	1491880241782005777	en	2022-02-10	Sir Botalot dropping the Mike!		@singervehicles Will you do a straight swap for my Renault Twingo 1.2 Extreme? \n\nThat is awesome! 🔥	straig
1	1501237946590142469	en	2022-03-08	Sabrina Ghio	RT	Australian GP Qualifying, Melbourne, 8th March 1997. #F1 The Grid ⬇️\n\nRow 3:\n5. Eddie Irvine (Ferrari) +2.512s\n6. Mika Hakkinen (McLaren-Mercedes) +2.602s\n\nRow 4:\n7. Johnny Herbert (Sauber-Petronas) +2.918s\n8. Jean Alesi (Benetton-Renault) +3.224s https://t.co/i4IC4LzVVN	qualif th ed m row saub
2	1505982695129718784	en	2022-03-21	Colin N. Walker 🏴󠁧󠁢󠁳󠁣󠁴󠁿 🇪🇺 💙😷 #FBPE	RT	#BoycottRenault\n\nThink of the blood of thousands of Ukrainian women and children pouring from every Renault car. \n\nhttps://t.co/rbU01Sy9DU	boyc b u c e
3	1516744110463463426	en	2022-04-20	Yvette Lissman	RT	Almost 200,000 workers in Russia still on western payrolls\n\nMcDonald’s, IKEA, Renault, Levi Strauss, &amp;others pay salaries to thousands of their employees while their operations in🇷🇺are suspended\nCoca-Cola, Yum Brands,KFC didn't confirm if they still pay🇷🇺s https://t.co/KWDhCo1dM0	rus pay str sal c susp yum c
4	1493777143347630086	en	2022-02-16	Andile Xaba 🇿🇦	RT	Take a selfie with the New #Renault #ClioV and WIN R1000 fuel voucher 💥 tag @tableviewrenault &amp; #renaulttableviewcliov random winner announced 28.02.2022 @BradAtRenault 0825662336 to book a test drive 🇿🇦 https://t.co/D7V2GnJa8B	renau renau a

Create n number of ngrams columns based on the lemmatized token tweets

In [ ]:

```
for i in range(1, 11):
    tweets_df_c['ngrams_'+str(i)] = tweets_df_c['tokens'].apply(create_ngram
```

In [ ]:

```
tweets_df_c.head(1)
```

Out[ ]:

	id	lang	date	name	retweeted	text	text_clean	to
0	1491880241782005777	en	2022-02-10	Sir Botalot dropping the Mike!		@singervehicles Will you do a straight swap for my Renault Twingo 1.2 Extreme? \n\nThat is awesome! 🔥	straight swap renault twingo extreme awesome	[str re tv ext awes

In [ ]:

```
from nltk.metrics import jaccard_distance

def calculate_jaccard_distance(ngrams1, ngrams2):
    # Convert n-gram lists to sets
    set_ngrams1 = set(ngrams1)
    set_ngrams2 = set(ngrams2)

    distance = 1 - jaccard_distance(set_ngrams1, set_ngrams2)

    return distance
```

In [ ]:

```
def compute_max_jaccard(df, num):
    # Create new columns for storing the highest Jaccard similarity and corr
    df['highest_sim_'+str(num)] = 0.0  # Initialize with 0.0 (no similarity)
    df['highest_sim_ind_'+str(num)] = -1  # Initialize with -1 (no index)

    from itertools import combinations

    # Iterate over all unique pairs of rows in the dataframe
    for i, j in combinations(df.index, 2):
        # Calculate Jaccard similarity, handle cases with empty n-grams
        if len(df['ngrams_'+str(num)][i]) == 0 or len(df['ngrams_'+str(num)]
            similarity = 0  # If either n-gram set is empty, similarity is 0
        else:
            # Calculate Jaccard similarity for non-empty n-gram sets
            similarity = calculate_jaccard_distance(df.at[i, 'ngrams_'+str(n

        # Ignore perfect matches (similarity = 1), considering them as zero
        if similarity == 1:
            similarity = 0

        # Update the record for the highest similarity for each row
        # If current similarity is higher than the stored value, update it a
        if similarity > df.at[i, 'highest_sim_'+str(num)]:
            df.at[i, 'highest_sim_ind_'+str(num)] = j  # Update index
            df.at[i, 'highest_sim_'+str(num)] = similarity  # Update similar

        if similarity > df.at[j, 'highest_sim_'+str(num)]:
            df.at[j, 'highest_sim_ind_'+str(num)] = i  # Update index for th
            df.at[j, 'highest_sim_'+str(num)] = similarity  # Update similar

    # At the end of this function, each row in df will have the highest Jacc
    # and the index of the row with which this highest similarity is achieve
```

In [ ]:

```
for i in range(1,11):
    compute_max_jaccard(tweets_df_c, i)
```

In [ ]:

```
sim_columns
```

Out[ ]:

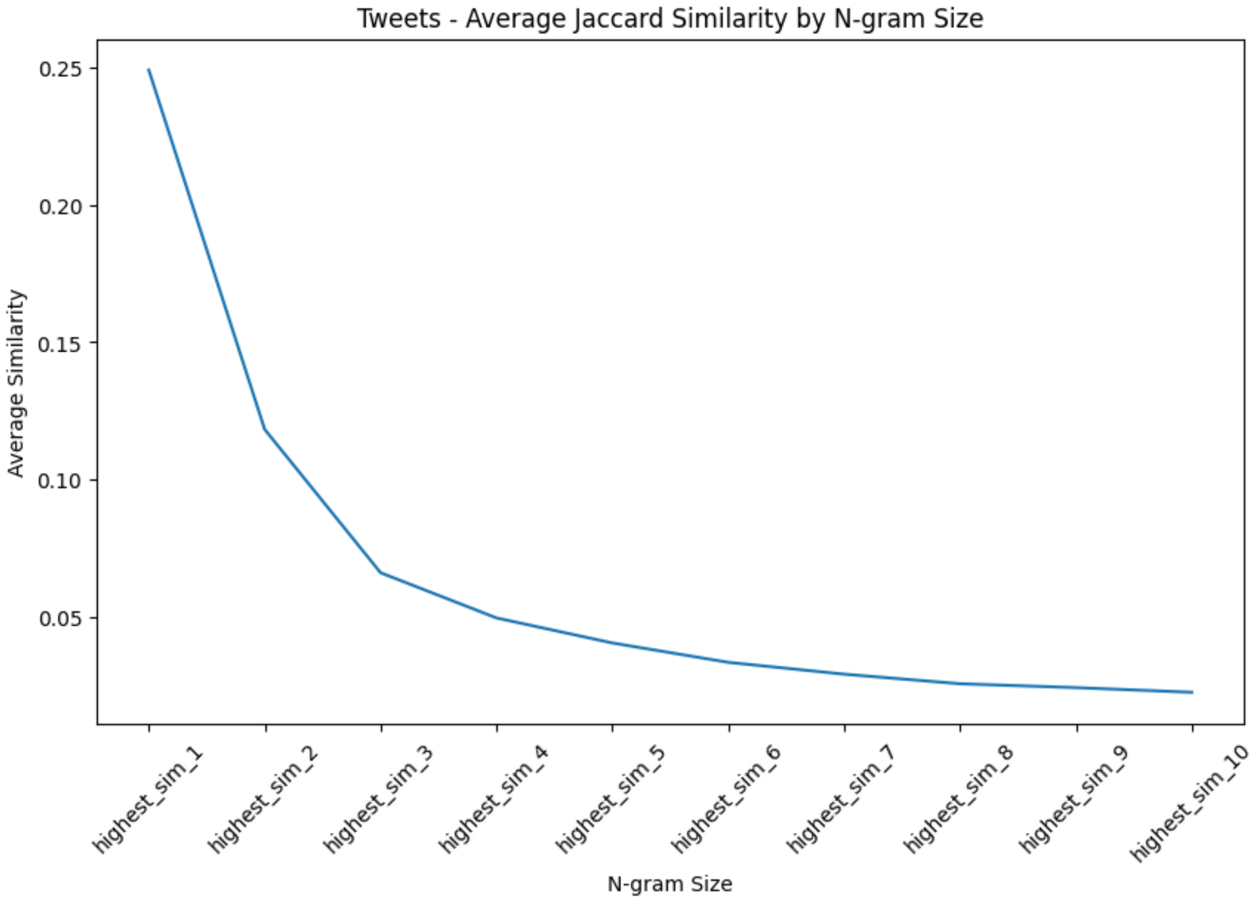
```
Index(['highest_sim_1', 'highest_sim_2', 'highest_sim_3', 'highest_sim_4',
       'highest_sim_5', 'highest_sim_6', 'highest_sim_7', 'highest_sim_8',
       'highest_sim_9', 'highest_sim_10'],
      dtype='object')
```

In [ ]:

```
import matplotlib.pyplot as plt

sim_columns = tweets_df_c.filter(regex='^highest_sim_\d+').columns
avg_sim = tweets_df_c[sim_columns].mean()

# Plot using seaborn's lineplot
plt.figure(figsize=(10, 6))  # Adjust the figure size as necessary
sns.lineplot(data=avg_sim)
plt.xticks(rotation=45)  # Rotate x-axis labels for better visibility
plt.xlabel('N-gram Size')  # Set title for the x-axis
plt.ylabel('Average Similarity')  # Set title for the y-axis
plt.title('Tweets - Average Jaccard Similarity by N-gram Size')  # Set the t
plt.show()
```



Based on the above plot, I would conclude an ngram value of 3 is best for tweets. This makes sense because ~5 is usually ideal for shorter texts whereas a higher number like 10 is better for long articles / books. So tweets aligning with n=3 seems correct.

In [ ]:

```
tweets_df[['highest_sim_3','highest_sim_ind_3','text_clean','text']].nlarges
```

Out[ ]:

	highest_sim_3	highest_sim_ind_3	text_clean	text
15	0.947368	378	renault kiger stunning yet muscular suv stance crafted complement free spirit renault cars nagercoil sportysmart renaultcars renaultindia bestcars buycarsnagercoil morespacing renaultkiger	Renault Kiger is stunning yet muscular SUV stance is crafted to complement your free spirit. \n#renault #cars #nagercoil #Sportysmart #renaultcars #renaultindia #bestcars #buycarsnagercoil #morespacing #renaultkiger https://t.co/0gaRyxca7h
378	0.947368	15	kiger stunning yet muscular suv stance crafted complement free spirit renault cars nagercoil sportysmart renaultcars renaultindia bestcars buycarsnagercoil morespacing renaultkiger	https://t.co/DFfqSVOypZ \nKiger is stunning yet muscular SUV stance is crafted to complement your free spirit.\n#renault #cars #nagercoil #Sportysmart #renaultcars #renaultindia #bestcars #buycarsnagercoil #morespacing #renaultkiger https://t.co/LzLJ0IjpyT
181	0.909091	877	cnn renault announced departure week russian auto market amid countrys war ukraine moscows mayor announced factory used restart defunct sovietera moskvich car brand via	CNN: After Renault announced its departure this week from the Russian auto market amid the country's war with Ukraine, Moscow's mayor announced its factory will be used to restart the defunct Soviet-era Moskvich car brand. https://t.co/QtIpxQ9d6I (via https://t.co/IgfjwGPwdG)
877	0.909091	181	renault announced departure week russian auto market amid countrys war ukraine moscows mayor announced factory used restart defunct sovietera moskvich car brand	After Renault announced its departure this week from the Russian auto market amid the country's war with Ukraine, Moscow's mayor announced its factory will be used to restart the defunct Soviet-era Moskvich car brand. https://t.co/F3pRGrtYQX
140	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the

			majority owner avtovaz russias largest car manufacturer famous lada brand	majority owner of AvtoVaz, Russia’s largest car manufacturer famous for the Lada brand.
241	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault majority owner avtovaz russias largest car manufacturer famous lada brand	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the majority owner of AvtoVaz, Russia’s largest car manufacturer famous for the Lada brand.
242	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault majority owner avtovaz russias largest car manufacturer famous lada brand	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the majority owner of AvtoVaz, Russia’s largest car manufacturer famous for the Lada brand.
406	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault majority owner avtovaz russias largest car manufacturer famous lada brand	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the majority owner of AvtoVaz, Russia’s largest car manufacturer famous for the Lada brand.
430	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault majority owner avtovaz russias largest car manufacturer famous lada brand	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the majority owner of AvtoVaz, Russia’s largest car manufacturer famous for the Lada brand.

447	0.846154	659	reuters french automaker renault resumes production moscow march company decided resume operations country renault majority owner avtovaz russias largest car manufacturer famous lada brand	⚡ Reuters: French automaker Renault resumes production in Moscow.\n\nOn March 22, the company decided to resume operations in the country. Renault is the majority owner of AvtoVaz, Russia's largest car manufacturer famous for the Lada brand.

The highest similarity tweets are related to the french automatker, Renault cars, and mentions of Moscow.

In [ ]:

```
non_zero_tweets = tweets_df[tweets_df['highest_sim_3'] > 0]
non_zero_tweets[['highest_sim_3','highest_sim_ind_3', 'text_clean', 'text']]
```

Out[ ]:

	highest_sim_3	highest_sim_ind_3	text_clean	
667	0.019231	455	jean alesi finished nd first podium points since moving benettonrenault michael schumacher also took first podium finish points ferrari driver rd lap behind race winner damon hill brazilian gp interlagos st march	Jean Alesi finished 2nd, his first &amp; points with since m Benetton-Renault. Michael Schumac took his first podium finish &amp; p a Ferrari driver with 3rd in his F3 behind race winner Damon Hill. Braz Interlagos, 31st March 19 https://t.co/leJw
664	0.022222	898	rothmans williamsrenault launch estoril chassis fw engine renault rs v tyres goodyear eagle damon hill jacques villeneuve test driver jeanchristophe boullion	Rothmans Williams-Renault Launch, 1996. #F1 🇬🇧 🇫🇷 \nChassis: FW18\n Renault RS8, 3.0L V10\nTyres: G Eagle F1 \n5. Damon Hill 🇬🇧 \n6. Villeneuve 🇨🇦 \nTest Driver: Jean-Ch Boullion 🇫🇷 https://t.co/giO
			mild seven renault car launch event monaco chassis engine mecachromebuilt	Mild Seven Renault F1 2005 car launc at Monaco\n\nChassis R25\ Mecachrome-built Renault RS25 3.

898	0.022222	664	renault rs v naturally aspirated power hp rpm weight kg tyres michelin points race wins podiums world titles elplan	72° naturally aspirated\nPower 800- @ 19,000 RPM\nWeight 605 kg Michelin\n\n191 points, 8 race podiums and 2 World Titles\n\n https://t.co/L77O
641	0.023810	877	car news today mr va auto legendary sovietera moskvich car could revived renault exit mr va auto car buying hero autoblog news car check autoblogs news	Car News of Today 🚗 Auto\n.\n'Legendary' Soviet-era M car could be revived after Renault exi VA Auto - Your Car Hero\nhttps://t.co/MWNhcYg5LC\n#A #Car\nhttps://t.co/tNfYThIk8q\nchec AutoBlogs news here\nhttps://t.co/fpt https://t.co/unQ3
771	0.023810	1012	michael schumacher mild seven benettonrenault lap qualifying couldnt beat session best german claimed pole position spanish gp qualifying barcelona th may	Michael Schumacher (Mild Seven Be Renault B195) on a lap in qualifyi couldn't beat his session best German claimed Pole Position.\nSpa Qualifying, Barcelona, 13th Ma #F1\nhttps://t.co/DMcxB
849	0.023810	61	tailored meet demands uk professionals requiring highspec vehicle reflects business values lifestyle new renault trucks trafic red edition tonnes prioritises productivity safety driver comfort renaulttrucks	Tailored to meet the demand professionals requiring a high-spec that reflects their own business val lifestyle, the new Renault Trucks Tra Edition from 2.8 - 3.1 tonnes pr productivity, safety and driver com \n\n#RenaultTrucks https://t.co/jM69
21	0.026316	718	behind powerfully elegant suv silhouette hides technological marvel allnew renault austral almost ready closeup	Behind this powerfully elega silhouette hides a technological mar new Renault #Austral is almost read close-up! #comingsoon\nLear https://t.co/eMQE https://t.co/ETN

			comingsoon learn	
60	0.026316	200	riccardo patrese brought williamsrenaultfwc home rd position italians fourth consecutive podium finish french grand prix paul ricard july motorsport images	Riccardo Patrese brought his W Renault-FW12C home in 3rd positior was the Italian's fourth consecutive finish. \n\nFrench Grand Prix, Paul R July 1989.\n\n© Motorsport Images https://t.co/w00
61	0.026316	822	today cest participate tunemyt challenge share beautiful renault trucks high pimped truck within new renault trucks dlc win incredible suprises merch eshop	🚨 You have until today 12am (participate in the #T challenge.\nShare you most b Renault Trucks T &amp; T High pimpe within the new Renault Trucks @SCSsoftware\n to win some inc suprises from our merch https://t.co/I03MpE
718	0.026316	21	allnew renault austral sound quality renault group brand news renault follow mycarnewsonline see latest news reviews tech french manufacturer right car news watch full video youtube fol	The all-new Renault Austral: The s quality | Renault Group\n\nFor mor news from Renault follow MyCarNews See the latest news, reviews and te the French manufacturer right her Car News.\n\nWatch Full V YouTube\n\nFol...\nhttps://t.co/vqXN

The lowest similarity rows, as expected do not appear related. For example, the first few tweets talk about Renault and Ricciardo and do not share as much simmilar text/language

Section #2: Articles

In [ ]:

```
def news_tokenize(text):
    # Directly tokenize the text using the word_tokenize function
    tokens = word_tokenize(text)
    return tokens
```

In [ ]:

```
# Apply the cleaning functio nto the column that we defined earlier
news_df_c['text_clean'] = news_df_c['text'].apply(clean_text)
news_df_c['title_clean'] = news_df_c['title'].apply(clean_text)
```

In [ ]:

```
# Tokenize the news articles text and titles
news_df_c['text_tokens'] = news_df_c['text_clean'].apply(news_tokenize)
news_df_c['title_tokens'] = news_df_c['title_clean'].apply(news_tokenize)
```

In [ ]:

```
# Lemmatize the tokens to the root
lemmatizer = WordNetLemmatizer()
news_df_c['text_tokens'] = news_df_c['text_tokens'].apply(lambda tokens: [le
news_df_c['title_tokens'] = news_df_c['title_tokens'].apply(lambda tokens: [

news_df_c[['text_tokens','title_tokens']].head(5)
```

Out[ ]:

	text_tokens	title_tokens
0	[lml, bike, car, dealer, lml, bike, showroom, india, explore, friday, january, log, insign, searchnotificationstop, sectionsauto, newscar, newsbike, newslatestauto, newsphotosvideoselectric, vehiclestrendingmy, readsoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculatordealersexplore, autoabout, uscontact, ussitemaprssterms, useprivacy, policycopyright, ht, medium, limited, right, reservedhomeoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculatordealerscar...	[lml, bike, car, dealer, lml, bike, showroom, india]
1	[pure, ev, bike, car, dealer, pure, ev, bike, showroom, india, explore, friday, january, log, insign, searchnotificationstop, sectionsauto, newscar, newsbike, newslatestauto, newsphotosvideoselectric, vehiclestrendingmy, readsoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculatordealersexplore, autoabout, uscontact, ussitemaprssterms, useprivacy, policycopyright, ht, medium, limited, right, reservedhomeoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculator...	[pure, ev, bike, car, dealer, pure, ev, bike, showroom, india]
2	[syncron, price, selected, mitsubishi, motor, corporation, boost, enhanced, service, part, pricing, strategyskip, contentcircle, country, music, lifestyleadvertise, usteacher, tributeask, expertthank, nursebe, excellentwatch, livenewselectionsvaccine, trackervideoweathersportscommunitycontestsabout, uscovidsearchhomesee, snap, send, itnewsstorm, centurynationalstateeditorialinvestigateeast, texas, ag, newscrimeeast, texas, nowthe, next, normalsept, thweathersign, thundercalllake, levelsproje...	[syncron, price, selected, mitsubishi, motor, corporation, boost, enhanced, service, part, pricing, strategy]
3	[mahindra, tease, future, electric, lineup, motoroids, motoroids, blogmotoroids, forum, authorscontact, ussubmit, storyadvertise, usprivacy, policy, search, homeauto, newsfeatureslaunchesupcoming, carsupcoming, suvsupcoming, bikesrecent, launchesreviewsmodsmodified, bikesmodified, carsinteresting, offbeatlistscc, bikescc, bikescc, bikescc, bikescc, bikescc, bikescc, bikescc, cc, bikesbikes, indiacars, indiaforums, trending, triumph, trident, get, pricier, homenewsmahindra, tease, future, ele...	[mahindra, tease, future, electric, lineup, motoroids]
4	[jawa, bike, car, dealer, jawa, bike, showroom, india, explore, saturday, january, log, insign, searchnotificationstop, sectionsauto, newscar, newsbike, newslatestauto, newsphotosvideoselectric, vehiclestrendingmy, readsoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculatordealersexplore, autoabout, uscontact, ussitemaprssterms, useprivacy, policycopyright, ht, medium, limited, right, reservedhomeoffersnewfind, carsfind, bikescompare, carscompare, bikesemi, calculatordealer...	[jawa, bike, car, dealer, jawa, bike, showroom, india]

Examine the most common tokens

In [ ]:

```
from collections import Counter

Counter(news_df_c['text_tokens'].explode()).most_common(10)
```

```
Out[ ]: [('petrol', 15661),
         ('lakh', 13718),
         ('mt', 9260),
         ('offer', 8434),
         ('turbo', 8338),
         ('car', 6568),
         ('complete', 5837),
         ('cc', 5216),
         ('xv', 5113),
         ('aprview', 5053)]
```

```
In [ ]: Counter(news_df_c['title_tokens'].explode()).most_common(10)
```

```
Out[ ]: [('car', 555),
         ('bike', 336),
         ('offer', 255),
         ('discount', 254),
         ('march', 254),
         ('dealer', 251),
         ('showroom', 247),
         ('india', 223),
         ('news', 95),
         ('tata', 78)]
```

The title tokens appear more valuable with less noise -- we will use these title tokens for the remaining analysis

```
In [ ]: for i in range(1, 11):
            news_df_c['ngrams_'+str(i)] = news_df_c['title_tokens'].apply(create_ngr
```

```
In [ ]: news_df_c.head(1)
```

Out[ ]:

	url	date	language	title	te
0	https://auto.hindustantimes.com/lml-bikes/dealers/bodh-gaya	2022-01-21	en	Lml Bikes Car Dealers - Lml Bikes Showrooms in India	Lml Bikes Car Dealers - Lr Bikes Showrooms in Ind Explore Friday, 21 Janua 2022 Log in/Sign u SearchNotificationsTc SectionsAuto NewsC NewsBike NewsLatestAu NewsPhotosVideosElectr VehiclesTrendingM ReadsOffersnewFir carsFind bikesCompa carsCompare bikesEM calculatorDealersExplo AutoAbout UsConta UsSITEMAPRSSTerms UsePrivacy PolicyCopyrig © HT Media Limited All righ reserved.HomeOffersnewFir carsFind bikesCompa carsCompare bikesEM calculatorDealersC NewsBik

In [ ]:

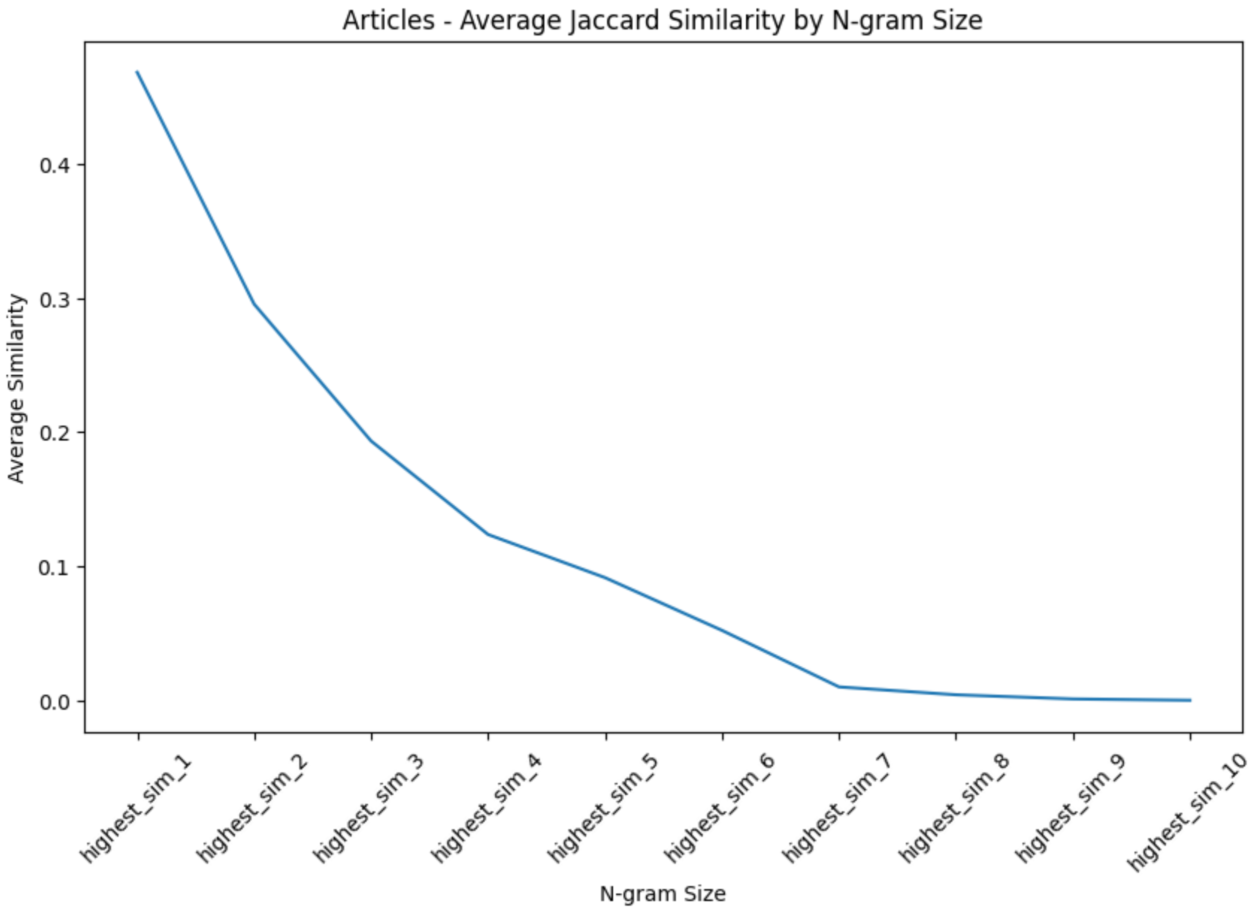
```python
for i in range(1,11):
    compute_max_jaccard(news_df_c, i)
```

In [ ]:

```python
import matplotlib.pyplot as plt

sim_columns = news_df_c.filter(regex='^highest_sim_\d+').columns
avg_sim = news_df_c[sim_columns].mean()

# Plot using seaborn's lineplot
plt.figure(figsize=(10, 6))  # Adjust the figure size as necessary
sns.lineplot(data=avg_sim)
plt.xticks(rotation=45)  # Rotate x-axis labels for better visibility
plt.xlabel('N-gram Size')  # Set title for the x-axis
plt.ylabel('Average Similarity')  # Set title for the y-axis
plt.title('Articles - Average Jaccard Similarity by N-gram Size')  # Set the
plt.show()
```



Based on the above plot, I would conclude an ngram value of 4 is best for the article titles. I will continue with this n value going forward.

```
In [ ]: news_df_c[['highest_sim_4','highest_sim_ind_4','title_clean', 'text_clean']]
```

Out[ ]:

	highest_sim_4	highest_sim_ind_4	title_clean	
6	0.857143	349	exnissan us exec kelly gets suspended sentence go home	exnissan us exec kelly gets suspende news us us politics politics world world cc
349	0.857143	6	exnissan us exec kelly gets suspended sentence go home ourquadcities	exnissan us exec kelly gets suspende termprimary menunewslocal newslocal
			france issues arrest	

35	0.833333	760	warrant disgraced auto tycoon ghosn	newsbusinesseducationweathertra
534	0.833333	760	france issues arrest warrant disgraced auto tycoon ghosn	france issues arrest warrant di uscirculationnewslett
664	0.833333	790	russias war spurs corporate exodus exposes business risks	russias war spurs corporate exodu usadvertisesubscribeprivacy noticeterms
760	0.833333	35	france issues arrest warrant disgraced auto tycoon ghosn kesq	buttonchevronrightchevronleftchevronups valley questions answerededucationiteam
790	0.833333	664	russias war spurs corporate exodus exposes business risks khon	russias war spurs corporate exodus expos menunewshawaii new
800	0.800000	912	japan prosecutors appeal exnissan executive kelly trial	japan prosecutors appeal ex regionweatherschool closingsobitua
912	0.800000	800	japan prosecutors appeal exnissan executive kelly trial wtmj	japan pr homenewslocalnationalcoronavirus
294	0.714286	35	france issues arrest warrant disgraced auto tycoon ghosn knwa fox	france issues arrest warrant di menunewsknwafoxaround arkansasruss

The highest similarity news articles by their title primarily are centered around france and nissans. A lof of the language between them is expectedly similar.

In [ ]:
```
non_zero_articles = news_df_c[news_df_c['highest_sim_4'] > 0]
non_zero_articles[['highest_sim_4','highest_sim_ind_4', 'title_clean', 'text
```

Out[ ]:

	highest_sim_4	highest_sim_ind_4	title_clean	text_clean
186	0.041667	396	pm gramin digital saksharta abhiyaan renault india partners csc egovernance services support pm gramin digital saksharta abhiyaan auto news et auto	pm gramin digital saksharta abhiyaan renault india partners csc egovernance services support pm gramin digital saksharta abhiyaan auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose igno...
237	0.047619	396	hero motocorp price hike hero motocorp hike motorcycle scooter prices inr july auto news et auto	hero motocorp price hike hero motocorp hike motorcycle scooter prices inr july auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies et aut...
185	0.050000	396	renault exports renaults made india products cross one lakh export milestone	renault exports renaults made india products cross one lakh export milestone auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see

			auto news et auto	privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies et autoa...
386	0.050000	396	biofuel us biofuel industry defends record biden administration mulls policy reform auto news et auto	biofuel us biofuel industry defends record biden administration mulls policy reform auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies e...
492	0.050000	396	global energy transition global energy transition cause shortterm economic pain report auto news et auto	global energy transition global energy transition cause shortterm economic pain report auto news et auto etautonewsnews passenger vehicle commercial vehicle two wheelers automotive components industry tyres aftermarket policy auto technology people movement oil lubes new launches raw material financial results auto finance featuresfeatures trends autopreneur etauto tv industryspeakindustryspeak interviews autologue etauto insights dealersdata analyticsdata analytics etautolytics reports etau...
660	0.050000	396	india fuel demand india expects fuel demand grow next fiscal year auto news et auto	india fuel demand india expects fuel demand grow next fiscal year auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies et autoanalyticsnec...

961	0.052632	396	lg energy solution gm sets bln michigan electric vehicle plants auto news et auto	lg energy solution gm sets bln michigan electric vehicle plants auto news et auto etautonewsnews passenger vehicle commercial vehicle two wheelers automotive components industry tyres aftermarket policy auto technology people movement oil lubes new launches raw material financial results auto finance featuresfeatures trends autopreneur etauto tv industryspeakindustryspeak interviews autologue etauto insights dealersdata analyticsdata analytics etautolytics reports etautotvbrand solutionsbran...
266	0.055556	396	msme promoters aima introduces month management course msme promoters auto news et auto	msme promoters aima introduces month management course msme promoters auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies et autoanalytic...
396	0.058824	452	ev policy state governments push ahead ev road auto news et auto	ev policy state governments push ahead ev road auto news et auto updated terms conditions privacy policy click continue accept continue et autoaccept updated privacy cookie policydear user et auto privacy cookie policy updated align new data regulations european union please review accept changes continue using websiteyou see privacy policy cookie policy use cookies ensure best experience websiteif choose ignore message well assume happy receive cookies et autoanalyticsnecessarynewsletter na...
			rupee value	rupee value rupee rises paise close us dollar auto news et auto etautonewsnews passenger vehicle commercial vehicle two wheelers automotive components industry tyres aftermarket policy auto

452	0.058824	396	rupee rises paise close us dollar auto news et auto	technology people movement oil lubes new launches raw material financial results auto finance featuresfeatures trends autopreneur etauto tv industryspeakindustryspeak interviews autologue etauto insights dealersdata analyticsdata analytics etautolytics reports etautotvbrand solutionsbrand solutions etauto...

The most dissimilar news articles/titles do not show much relation and span various topics like india, scooter prices, and fuel

Summary: The n for tweets and articles were different. n for tweets was 3 and n for article titles was 4 but it could have been 3 as well. Even though I set them to different n values, they both were close and were smaller than 5 ngrams. This makes sense because tweets and titles in general do not have long text, so setting a 10+ ngram value could encompass the whole text itself in some instances.