**Exercise 08: Web page mining**

**Ashutosh Laxminarayan Gor**

**EM 624: Informatics for Engineering management**

**Interpretation:**

* I selected CNN news website for my web page mining particularly the sports section.
* Initially I imported all the required libraries for the assignment.
* Then I targeted the website link using the urllib library.
* Using the url library the website was targeted and http unverified url certificate error was clarified using ssl library.
* Using the beautiful soup library, I captured the headlines specifically from the websites sports section.
* After capturing the headlines, I imported the stopword file provided by professor to remove unwanted words from the captured data.
* Then I created an empty list to populate the headlines data captured.
* Post which I converted all the list into lower alphabets and removed all the words which matched the data from stopword file.
* After which I removed all the non-alphabetical words from the captured data.
* Then I removed all the common frequently words being used in the captured data.
* After cleaning the captured data, I used the sentiment analyzer to analyse the sentiment behind the words being used in the captured data.
* I converted the data into an string and processed the data into analyzer then printed the output accordingly.
* Post that I converted the data again into a list and then again into a string and captured the most positive and most negative line in the captured data from the sentiment analyzer.
* Post that I used the cleaned data to create the bigram using the tokenization method and nltk library.
* After that I converted the bigram extracted into a string and replaced the commas with an underscore between 2 bigram words.
* After printing the extracted bigram data, I used the wordcloud library and created a wordcloud using the cleaned data with the words limitation of 100 words.

**Expected Output:**

These are the headlines captured from the news website:

Cristiano Ronaldo to leave Manchester United with immediate effect

Saudi Arabia stuns Lionel Messi's Argentina in one of the biggest upsets in World Cup history

US journalist says he was initially denied entry to a stadium because of rainbow t-shirt

Bale saves Wales to frustrate USMNT at Qatar 2022

Iran players remain silent during national anthem at World Cup in apparent protest at Iranian regime

England's Harry Kane and several other European captains told not to wear 'OneLove' armband at World Cup

England vs. Iran: For Iranians, this World Cup is about more than football

Cristiano Ronaldo defends timing of controversial interview, says it 'won't shake' Portugal team

A brief history of World Cup shocks

Their countries' teams aren't even in the World Cup. So why are these fans traveling to Qatar?

Worker deaths, exploitation, and anger casts shadow over Qatar's historic World Cup

What fans can't do in Qatar

World Cup guide: Teams and players to watch

Qatar says all are welcome but some LGBTQ fans are staying away

For Lionel Messi and Cristiano Ronaldo, a final shot at World Cup glory

Ronaldo risks ruining his legacy after TV interview slamming Man Utd

How a NASA scientist and a record-breaking mountaineer conquered an unknown mountain

'I run to find peace,' says Marvel actress Lauren Ridloff after finishing her first marathon

'I'm essentially starting a new life,' says Isaac Humphries as he becomes the only openly gay man playing top-flight basketball

Wayne Rooney as you've never seen him before -- rolling sushi

Scotland recognized as world's best golf destination

Golfer gives fans $100 to buy beers for moving out of shadows

Golfer Becky Brewerton's mental health journey back to the top

Ecuador dampens Qatar's party as controversial World Cup gets underway

British comedian Joe Lycett did not shred $11,000 over David Beckham's role as Qatar World Cup ambassador

FIFA president criticized for speech on Qatar's human rights

Wimbledon relaxes all-white clothing rules for women players

Carlos Alcaraz becomes youngest year-end No. 1 in men's tennis

Australia to overturn Djokovic visa ban, source says

Bills and Browns escape blizzards in Buffalo

'It was a learning journey': Brooklyn Nets star Kyrie Irving issues another apology as possible return from suspension nears

NBA superstar Luka Doncic says he prefers chess to Twitter

Max Verstappen says abuse of family is 'unacceptable' following São Paulo Grand Prix

George Russell wins maiden Grand Prix after collision-ridden race in Brazil

This Michael Schumacher Ferrari just smashed auction records

Mikaela Shiffrin clinches her 75th World Cup win, breaking a Lindsey Vonn record

World Anti-Doping Agency refers Russian figure skater Kamila Valieva's case to Court of Arbitration for Sport

International Skating Union proposes raising minimum age to 17 for competitions

Boxing world champion Oleksandr Usyk shares images from family home in Ukrainian area previously held by Russians

'Please don't forget about us,' says Ukrainian football legend Andriy Shevchenko

Ukrainian tennis star on a 'mission' to help war-torn country

How a quadriplegic former rugby player conquered a Himalayan mountain

How feeling 'vulnerable' helped climber Lor Sabourin explore their identity

Omani climber wants to 'die empty' in her quest to share knowledge

What it's like to circumnavigate Lesotho on foot -- in 16 days

Briton breaks world record for fastest female solo row across Atlantic

Ultramarathoner with autism inspiring others after crushing 100 mile goal

Bodybuilder wants to change the view of what women should look like

Olympian pressured to 'perform better' and lose a few pounds

'Oh you must be heavy boned.' Athlete's fight for body acceptance

Emmy nominated: Running as equals

Meet the Black women pushing for equality in swimming

'I had PTSD': Culture of abuse in British gymnastics leaves lasting legacy on its victims

She apologized for verbally abusing an Olympic judge, but questions over the conduct of China's 'Iron Lady' of diving remain

The Winter Olympics don't really represent the world: Costs, climate and quotas keep the majority off the podium

More than 1,000 students were sexually abused at this university. An ex-NFL player wants their stories to be heard

Marathon man's remarkable journey to Tokyo 2020

Navid Afkari: Executed Iranian wrestling star's voice is 'everywhere now'

The blind skateboarder challenging misconceptions about sight and sport

Nedd Brockmann 'went through hell and back 10 times' as he ran nearly 2,500 miles across Australia in 47 days

10 years on from Felix Baumgartner's near-cosmic achievement

Freestyle football great makes history with ninth world title

The day Roger Federer couldn't stop laughing at CNN correspondent's Spanish phrases

Atlético Madrid midfielder's comeback from brain cancer and mother's paralysis

World's only openly gay active pro footballer is concerned for LGBTQ community ahead of Qatar 2022

The 'raw' photo of Federer and Nadal which captures their enduring friendship

The former footballer who picked up running and ended up crossing America

Record-breaking strongwoman says powerlifting 'saved me from myself'﻿

How tech allows this paralyzed former racing driver to hit the track again﻿

Meet the Black women pushing for equality in swimming

Everything you need to know (and more!) about pickleball

In pictures: The Olympics opening ceremony in Beijing

Iconic moments from Winter Olympics history

25 athletes to watch in the Beijing Winter Olympics

In pictures: Soccer legend Diego Maradona

Dodgers win their first World Series since 1988

In photos: The sports world has been taking a stand

When 'Air Jordan' took flight

Cristiano Ronaldo to leave Manchester United with immediate effect

Saudi Arabia stuns Lionel Messi's Argentina in one of the biggest upsets in World Cup history

US journalist says he was initially denied entry to a stadium because of rainbow t-shirt

Gareth Bale saves Wales to frustrate USMNT at Qatar 2022

Iran players remain silent during national anthem at World Cup in apparent protest at Iranian regime

Cristiano Ronaldo defends timing of controversial interview, says it 'won't shake' Portugal team

World Cup fans acclimatise to desert accommodation -- in tents and portacabins

England's Harry Kane and several other European captains told not to wear 'OneLove' armband at World Cup

Can the USMNT's golden generation shine on the world stage?

Ecuador dampens Qatar's party as controversial World Cup gets underway

These are the headlines captured from the news website:

Ecuador dampens Qatar's party as controversial World Cup gets underway

The headlines are rated accordingly:

12.6% headlines have positive sentiment

9.6% headlines have negative sentiment &

77.9% headlines have neutral sentiment.

Most positive & most negative headlines in the data:

Australia overturn Djokovic visa source says

was learning Brooklyn Nets star Kyrie Irving issues apology possible return suspension nears

These are the bigrams extracted from the llist of headlines:

['Cristiano\_Ronaldo', 'Ronaldo\_leave', 'leave\_Manchester', 'Manchester\_United', 'United\_immediate', 'immediate\_effect', 'effect\_Saudi', 'Saudi\_Arabia', 'Arabia\_stuns', 'stuns\_Lionel', 'Lionel\_Argentina', 'Argentina\_one', 'one\_the', 'the\_biggest', 'biggest\_upsets', 'upsets\_World', 'World\_Cup', 'Cup\_history', 'history\_journalist', 'journalist\_says', 'says\_was', 'was\_initially', 'initially\_denied', 'denied\_entry', 'entry\_a', 'a\_stadium', 'stadium\_of', 'of\_rainbow', 'rainbow\_Bale', 'Bale\_saves', 'saves\_Wales', 'Wales\_frustrate', 'frustrate\_USMNT', 'USMNT\_Qatar', 'Qatar\_Iran', 'Iran\_players', 'players\_remain', 'remain\_silent', 'silent\_national', 'national\_anthem', 'anthem\_World', 'World\_Cup', 'Cup\_apparent', 'apparent\_protest', 'protest\_Iranian', 'Iranian\_regime', 'regime\_Harry', 'Harry\_Kane', 'Kane\_several', 'several\_European', 'European\_captains', 'captains\_told', 'told\_to', 'to\_wear', 'wear\_armband', 'armband\_World', 'World\_Cup', 'Cup\_brief', 'brief\_history', 'history\_World', 'World\_Cup', 'Cup\_shocks', 'shocks\_fans', 'fans\_do', 'do\_Qatar', 'Qatar\_World', 'World\_Cup', 'Cup\_Teams', 'Teams\_players', 'players\_watch', 'watch\_Qatar', 'Qatar\_says', 'says\_welcome', 'welcome\_some', 'some\_LGBTQ', 'LGBTQ\_fans', 'fans\_are', 'are\_staying', 'staying\_away', 'away\_Lionel', 'Lionel\_Messi', 'Messi\_Cristiano', 'Cristiano\_a', 'a\_final', 'final\_shot', 'shot\_World', 'World\_Cup', 'Cup\_glory', 'glory\_Ronaldo', 'Ronaldo\_risks', 'risks\_ruining', 'ruining\_legacy', 'legacy\_TV', 'TV\_interview', 'interview\_slamming', 'slamming\_Man', 'Man\_Utd', 'Utd\_a', 'a\_NASA', 'NASA\_scientist', 'scientist\_a', 'a\_mountaineer', 'mountaineer\_conquered', 'conquered\_unknown', 'unknown\_mountain', 'mountain\_run', 'run\_find', 'find\_says', 'says\_Marvel', 'Marvel\_actress', 'actress\_Lauren', 'Lauren\_Ridloff', 'Ridloff\_finishing', 'finishing\_first', 'first\_marathon', 'marathon\_essentially', 'essentially\_starting', 'starting\_new', 'new\_says', 'says\_Isaac', 'Isaac\_Humphries', 'Humphries\_he', 'he\_the', 'the\_openly', 'openly\_gay', 'gay\_man', 'man\_playing', 'playing\_basketball', 'basketball\_Golfer', 'Golfer\_gives', 'gives\_fans', 'fans\_to', 'to\_buy', 'buy\_beers', 'beers\_moving', 'moving\_of', 'of\_shadows', 'shadows\_Golfer', 'Golfer\_Becky', 'Becky\_mental', 'mental\_health', 'health\_journey', 'journey\_to', 'to\_top', 'top\_Ecuador', 'Ecuador\_dampens', 'dampens\_party', 'party\_controversial', 'controversial\_World', 'World\_Cup', 'Cup\_gets', 'gets\_underway', 'underway\_British', 'British\_comedian', 'comedian\_Joe', 'Joe\_Lycett', 'Lycett\_did', 'did\_shred', 'shred\_over', 'over\_David', 'David\_role', 'role\_Qatar', 'Qatar\_World', 'World\_Cup', 'Cup\_ambassador', 'ambassador\_Wimbledon', 'Wimbledon\_relaxes', 'relaxes\_clothing', 'clothing\_rules', 'rules\_women', 'women\_players', 'players\_Australia', 'Australia\_overturn', 'overturn\_Djokovic', 'Djokovic\_visa', 'visa\_source', 'source\_says', 'says\_Bills', 'Bills\_Browns', 'Browns\_escape', 'escape\_blizzards', 'blizzards\_Buffalo', 'Buffalo\_was', 'was\_learning', 'learning\_Brooklyn', 'Brooklyn\_Nets', 'Nets\_star', 'star\_Kyrie', 'Kyrie\_Irving', 'Irving\_issues', 'issues\_apology', 'apology\_possible', 'possible\_return', 'return\_suspension', 'suspension\_nears', 'nears\_NBA', 'NBA\_superstar', 'superstar\_Luka', 'Luka\_Doncic', 'Doncic\_says', 'says\_prefers', 'prefers\_chess', 'chess\_Twitter', 'Twitter\_Michael', 'Michael\_Schumacher', 'Schumacher\_Ferrari', 'Ferrari\_just', 'just\_smashed', 'smashed\_auction', 'auction\_records', 'records\_World', 'World\_Agency', 'Agency\_refers', 'refers\_Russian', 'Russian\_figure', 'figure\_skater', 'skater\_Kamila', 'Kamila\_case', 'case\_Court', 'Court\_Arbitration', 'Arbitration\_Sport', 'Sport\_Boxing', 'Boxing\_world', 'world\_champion', 'champion\_Oleksandr', 'Oleksandr\_Usyk', 'Usyk\_shares', 'shares\_images', 'images\_family', 'family\_home', 'home\_Ukrainian', 'Ukrainian\_area', 'area\_previously', 'previously\_held', 'held\_Russians', 'Russians\_Ukrainian', 'Ukrainian\_tennis', 'tennis\_star', 'star\_a', 'a\_to', 'to\_help', 'help\_country', 'country\_quadriplegic', 'quadriplegic\_rugby', 'rugby\_player', 'player\_conquered', 'conquered\_a', 'a\_Himalayan', 'Himalayan\_mountain', 'mountain\_feeling', 'feeling\_helped', 'helped\_climber', 'climber\_Lor', 'Lor\_Sabourin', 'Sabourin\_explore', 'explore\_identity', 'identity\_Briton', 'Briton\_breaks', 'breaks\_world', 'world\_record', 'record\_fastest', 'fastest\_female', 'female\_solo', 'solo\_row', 'row\_Atlantic', 'Atlantic\_Ultramarathoner', 'Ultramarathoner\_autism', 'autism\_inspiring', 'inspiring\_after', 'after\_crushing', 'crushing\_mile', 'mile\_goal', 'goal\_Bodybuilder', 'Bodybuilder\_wants', 'wants\_change', 'change\_view', 'view\_what', 'what\_women', 'women\_look', 'look\_like', 'like\_Emmy', 'Emmy\_Running', 'Running\_equals', 'equals\_Meet', 'Meet\_Black', 'Black\_women', 'women\_pushing', 'pushing\_equality', 'equality\_swimming', 'swimming\_had', 'had\_Culture', 'Culture\_abuse', 'abuse\_British', 'British\_gymnastics', 'gymnastics\_leaves', 'leaves\_lasting', 'lasting\_legacy', 'legacy\_its', 'its\_victims', 'victims\_than', 'than\_students', 'students\_sexually', 'sexually\_abused', 'abused\_this', 'this\_An', 'An\_player', 'player\_wants', 'wants\_stories', 'stories\_be', 'be\_heard', 'heard\_Marathon', 'Marathon\_remarkable', 'remarkable\_journey', 'journey\_Tokyo', 'Tokyo\_blind', 'blind\_skateboarder', 'skateboarder\_challenging', 'challenging\_misconceptions', 'misconceptions\_sight', 'sight\_sport', 'sport\_Freestyle', 'Freestyle\_football', 'football\_great', 'great\_makes', 'makes\_history', 'history\_ninth', 'ninth\_world', 'world\_title', 'title\_day', 'day\_Roger', 'Roger\_Federer', 'Federer\_stop', 'stop\_laughing', 'laughing\_CNN', 'CNN\_Spanish', 'Spanish\_phrases', 'phrases\_only', 'only\_openly', 'openly\_gay', 'gay\_active', 'active\_pro', 'pro\_footballer', 'footballer\_concerned', 'concerned\_LGBTQ', 'LGBTQ\_community', 'community\_ahead', 'ahead\_Qatar', 'Qatar\_former', 'former\_footballer', 'footballer\_picked', 'picked\_running', 'running\_ended', 'ended\_crossing', 'crossing\_America', 'America\_tech', 'tech\_allows', 'allows\_paralyzed', 'paralyzed\_racing', 'racing\_driver', 'driver\_hit', 'hit\_track', 'track\_Meet', 'Meet\_Black', 'Black\_women', 'women\_pushing', 'pushing\_equality', 'equality\_swimming', 'swimming\_Iconic', 'Iconic\_moments', 'moments\_Winter', 'Winter\_Olympics', 'Olympics\_history', 'history\_athletes', 'athletes\_watch', 'watch\_the', 'the\_Beijing', 'Beijing\_Winter', 'Winter\_Olympics', 'Olympics\_Cristiano', 'Cristiano\_Ronaldo', 'Ronaldo\_leave', 'leave\_Manchester', 'Manchester\_United', 'United\_immediate', 'immediate\_effect', 'effect\_Saudi', 'Saudi\_Arabia', 'Arabia\_stuns', 'stuns\_Lionel', 'Lionel\_Argentina', 'Argentina\_one', 'one\_the', 'the\_biggest', 'biggest\_upsets', 'upsets\_World', 'World\_Cup', 'Cup\_history', 'history\_journalist', 'journalist\_says', 'says\_was', 'was\_initially', 'initially\_denied', 'denied\_entry', 'entry\_a', 'a\_stadium', 'stadium\_of', 'of\_rainbow', 'rainbow\_Gareth', 'Gareth\_Bale', 'Bale\_saves', 'saves\_Wales', 'Wales\_frustrate', 'frustrate\_USMNT', 'USMNT\_Qatar', 'Qatar\_Iran', 'Iran\_players', 'players\_remain', 'remain\_silent', 'silent\_national', 'national\_anthem', 'anthem\_World', 'World\_Cup', 'Cup\_apparent', 'apparent\_protest', 'protest\_Iranian', 'Iranian\_regime', 'regime\_World', 'World\_Cup', 'Cup\_fans', 'fans\_acclimatise', 'acclimatise\_desert', 'desert\_accommodation', 'accommodation\_in', 'in\_tents', 'tents\_portacabins', 'portacabins\_Harry', 'Harry\_Kane', 'Kane\_several', 'several\_European', 'European\_captains', 'captains\_told', 'told\_to', 'to\_wear', 'wear\_armband', 'armband\_World', 'World\_Cup', 'Cup\_the', 'the\_golden', 'golden\_generation', 'generation\_shine', 'shine\_the', 'the\_world', 'world\_Ecuador', 'Ecuador\_dampens', 'dampens\_party', 'party\_controversial', 'controversial\_World', 'World\_Cup', 'Cup\_gets', 'gets\_underway']

Wordcloud for the list of headlines:

Text

Description automatically generated

Process finished with exit code 0