BCSC 4227 DATA SCIENCE

BITC01/0054/2019

KIMUTAI GRIFFINS

Project Scope: Analyzing tweets collected using the query term "World Cup Finals"

Project Report

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

Data Collection

The first step of text mining is gathering the data. With this, the python tweepy library came in handy to facilitate the pulling of data from Twitter. The data pulled from the API is unstructured data- (This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.)

The data collected has to be cleaned and transformed into usable format.It involves the use of techniques such as language identification, tokenization and syntax parsing.

```
In [2]: # Authentication
        API_KEY = "tPQsvRTgOXj6LdwWEE6jU5Z7A"
        API_SECRET_KEY = "ebO0BgMIrZwJ4BxTW6bTWbAhm7bTzaRdhZNxED1UdNYcUtofS4"
        ACCESS_TOKEN = "770261890257686528-yV49uHpcD3HN3FZBbiijFqw8IkjxVOb"
        ACCESS_SECRET_TOKEN = "YNItLRZclGYcidN5pGDYBtprFQnKnuyjD0oMM963JZzQM"

        auth = tweepy.OAuthHandler(API_KEY, API_SECRET_KEY)
        auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET_TOKEN)
        api = tweepy.API(auth)
        print(api)

        <tweepy.api.API object at 0x7f0df1866470>
```

Information retrieval

Returning relevant information based on a pre-defined set of queries.

Tasks performed include :

Tokenization - breaking out long-form text into sentences

```
In [17]: # Tokenization of words
         tweetWords = word_tokenize(cleanTweets)
         print(tweetWords)

['[', '[', '```', ':', 'FIFAWorldCup', 'quarter-finals', 'start', 'today', '!', 'It', "'s", 'Brazil', 'vs', 'Croatia
', 'starting', 'at', '8:30', 'PM', 'IST', 'which', 'will', 'be', 'followed', 'by', 'Argentina', 'vs', 'N…', "'",
',', '0.0', ',', '0.1', ']', ',', '```', ':', 'FIFAWorldCup', 'quarter-finals', 'start', 'today', '!', 'It', "'s", 'B
razil', 'vs', 'Croatia', 'starting', 'at', '8:30', 'PM', 'IST', 'which', 'will', 'be', 'followed', 'by', 'Argentina
', 'vs', 'N…', "'", ',', '[', "'500", ':', 'Today', ',', 'the', 'FIFA', 'World', 'Cup', 'quarter-finals', 'begin',
'!', '⚽', '\\n\\nWho', '’', 's', 'going', 'to', 'the', 'semi-finals', '?', '🏆', '\\n\\nCroatia', '🇭🇷', 'vs', '🇧🇷',
'Brazil\\nNetherlands…', "'", ',', '0.0', ',', '0.0', ']', ',', "'500", ':', 'Today', ',', 'the', 'FIFA', 'World', '
Cup', 'quarter-finals', 'begin', '!', '⚽', '\\n\\nWho', '’', 's', 'going', 'to', 'the', 'semi-finals', '?', '🏆',
'\\n\\nCroatia', '🇭🇷', 'vs', '🇧🇷', 'Brazil\\nNetherlands…', "'", ',', '[', "'", ':', 'The', 'last', 'time', 'these',
'two', 'met', 'at', 'the', 'quarter-finals', 'of', 'the', 'World', 'Cup', 'was', 'in', '1998', '.', 'Dennis', 'Bergk
amp', 'scored', 'a', 'clutch', 'goal', 'on', 'th…', "'", ',', '0.0', ',', '0.06666666666666667', ']', ',', "'", ':',
'The', 'last', 'time', 'these', 'two', 'met', 'at', 'the', 'quarter-finals', 'of', 'the', 'World', 'Cup', 'was', 'in
', '1998', '.', 'Dennis', 'Bergkamp', 'scored', 'a', 'clutch', 'goal', 'on', 'th…', "'", ',', '[', "'", ':', '🚨|',
'2022', 'World', 'Cup', 'quarter-finals', 'games', 'schedule', 'for', 'today', ':', '\\n\\n•', 'Croatia', '🇭🇷', 'vs
', '🇧🇷', 'Brazil', '(', '4pm', 'CET', ')', '.\\n\\n•', 'Netherlands', '🇳🇱', 'vs', '🇦🇷', 'Ar…', "'", ',', '0.0', ',',
'0.0', ']', ',', "'", ':', '🚨|', '2022', 'World', 'Cup', 'quarter-finals', 'games', 'schedule', 'for', 'today',
':', '\\n\\n•', 'Croatia', '🇭🇷', 'vs', '🇧🇷', 'Brazil', '(', '4pm', 'CET', ')', '.\\n\\n•', 'Netherlands', '🇳🇱', 'vs',
'🇦🇷', 'Ar…', "'", ',', '[', '```', ':', 'On', 'the', 'eve', 'of', 'the', 'WorldCup2022', 'Quarter', 'Finals', ',', 'w
e', "'re", 'giving', 'away', 'a', '2022', 'Football', 'World', 'Cup', 'Shirt', 'of', 'your', 'choice', '!', '⚽\\n\\
```

Tokenization of sentences

```
In [16]: # Tokenization of sentences
         tweetSent = sent_tokenize(cleanTweets)
         print(tweetSent)
```

```
ember I watched the semi finals with my family at home.", "I very much enjoyed…\", 0.13, 0.63], \": Alexis Mac Allis
ter on the 2014 semi final: "I remember I watched the semi finals with my family at home.', "I very much enjoyed…',
['23: Messi could be out of the World Cup tomorrow night... Or in the Semi Finals probably against Brazil (or Croati
a)...", 'Sigh.', "', 0.0, 0.0], '23: Messi could be out of the World Cup tomorrow night... Or in the Semi Finals pro
bably against Brazil (or Croatia)...", 'Sigh.', "', ['The last time these two met at the quarter-finals of the World
Cup was in 1998.", "Dennis Bergkamp scored a clutch goa…  0.0, 0.06666666666666667], 'The last time these two met at
the quarter-finals of the World Cup was in 1998.", "Dennis Bergkamp scored a clutch goa…  ['3: The Netherlands and A
rgentina will meet in the quarter-finals of the World Cup today.", "Live on GOtv at 19pm ch 31 \\nWhat…', 0.13636363
636363635, 0.5], '3: The Netherlands and Argentina will meet in the quarter-finals of the World Cup today.", 'Live o
n GOtv at 19pm ch 31 \\nWhat…\', [\'FIFA World Cup Quarter-Finals 2022: Live streaming, Teams, Venues, Dates - Prati
din Time  -…  0.13636363636363635, 0.5], \'FIFA World Cup Quarter-Finals 2022: Live streaming, Teams, Venues, Dates
- Pratidin Time  -…  [": Luis Enrique\'s biggest mistake was not to call up these two in Spain\'s World Cup Squad.',
'**\\n\\nIf these two couldn\'t even m…", 0.0, 0.0], ": Luis Enrique\'s biggest mistake was not to call up these two
in Spain\'s World Cup Squad.', '**\\n\\nIf these two couldn\'t even m…", [\': Holo's World Cup - Quarter-finals\\u2
00d\\n\\nWin prizes by predicting the outcome of World Cup matches!', "\\n\\nTo Enter:\\n\\n- Comple…', 1.0,
0.4], ': Holo's World Cup - Quarter-finals\\u200d\\n\\nWin prizes by predicting the outcome of World Cup matche
s!", "\\n\\nTo Enter:\\n\\n- Comple…', ['500: Today, the FIFA World Cup quarter-finals begin!", '\\n\\nWho's g
oing to the semi-finals?', "\\n\\nCroatia  vs  Brazil\\nNetherlands…', 0.0, 0.0], '500: Today, the FIFA World
Cup quarter-finals begin!", '\\n\\nWho's going to the semi-finals?', "\\n\\nCroatia  vs  Brazil\\nNetherlan
ds…', ['1: World Cup 2022: Watch: Messi, Ronaldo, Mbappe, Neymar and other stars set for quarter-finals\\nThreeLion
```

Stemming: separating the prefixes and suffixes from words to derive the root word form and meaning.

Natural language processing (NLP)

Ability to understand text and spoken words in much the same way human beings can.

Tasks involved include :

Stops words: These are actually the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and does not add much information to the

text

```
In [18]: # removing  stopwords
         from nltk.corpus import stopwords
         # print(stopwords.words('english')) All the stopwords in english
         text_with_no_stopwords = [word for word in tweetWords if word not in stopwords.words('english')]
         text_with_no_stopwords
```

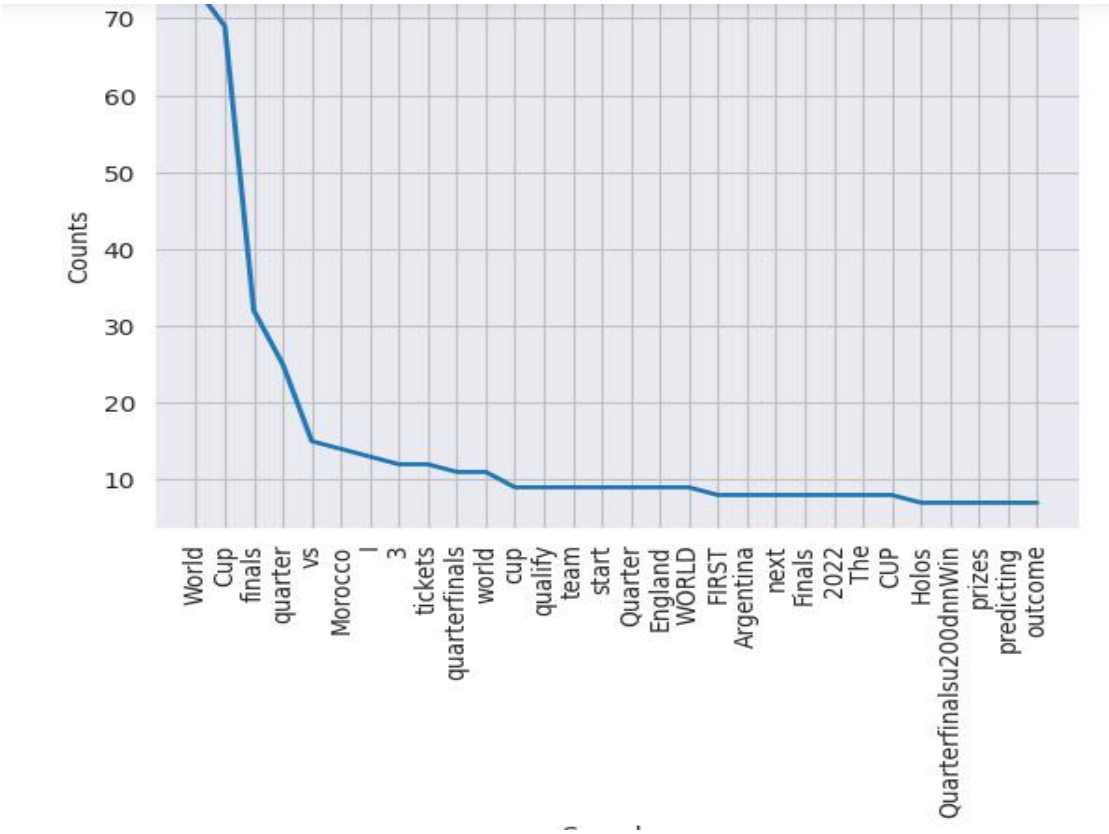
```
Out[18]: ['[',
          '[',
          '``',
          ':',
          'FIFAWorldCup',
          'quarter-finals',
          'start',
          'today',
          '!',
          'It',
          "'s",
          'Brazil',
          'vs',
          'Croatia',
          'starting',
          '8:30',
          'PM',
          'IST',
          'followed',
          'Argentina',
```

Sentiment analysis: This task detects positive or negative sentiment from data sources, allowing you to track changes in customer attitudes over time.

Summarization : provides a synopsis of long pieces of text to create a concise, coherent summary of a document's main points.

Information extraction

Word frequency -counts the *frequency* of each and every *word* in a text, helping you understand the keyword density for SEO or the rate of *word* repetition



Extracting structured information from free text and storing these entities, attributes, and relationship information in a database.

Feature selection, or attribute selection, is the process of selecting the important features (dimensions) to contribute the most to output of a predictive analytics model.

Feature extraction is the process of selecting a subset of features to improve the accuracy of a classification task. This is particularly important for dimensionality reduction.

Named-entity recognition (NER) also known as entity identification or entity extraction, aims to find and categorize specific entities in text, such as names

or locations. For example, NER identifies “California” as a location and “Mary” as a woman’s name.

Importance of the performed analysis

Sentiment analysis-

provides information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

Providing a mechanism for companies to prioritize key pain points for their customers, allowing businesses to respond to urgent issues in real-time and increase customer satisfaction.

Risk management

Give insights into industry trends and financial markets by tracking shifts in sentiment and pulling information from analyst reports and whitepapers This is especially beneficial to financial organizations since it gives them more confidence when contemplating company investments in diverse areas.

Text mining tools analyze documents to identify entities and extract relationships between them, unlocking hidden information to help researchers:

Challanges encountered during the proccess

Short Informal texts - one of the difficulties in sentiment analysis is short informal text.They are restricted in length, usually spanning one or less sentences.They frequently use slang words, misspellings, and truncated word forms.

Many of the collected tweets are stripped of context or too short to serve as real carriers of meaning. Most of the tweets have the meaning hidden in images or links to websites.

Twitter API has a request limit that limits the amount of data to collect and analyse.(TooManyRequests: 429 Too Many Requests 88 - Rate limit exceeded)