Data Analysis of Textual Data

CIS 3330

A module in review

In the last module, we learn about descriptive and predictive analytics

We learned some supervised and unsupervised algorithms

In terms of reporting analytics, we learn how to create visualizations, obtain summaries from models, and evaluate algorithms.

Exam 2 is due 04/23.

Exam 3 survey is open and due Today at 11:00 PM.

Text Analysis

Text Analysis – Introduction

- Daily information systems collect data of many kinds (e.g., numerical, categorical).
- Computers can interpretate some data better than others.
- You want to be careful not to confuse numerical data with categorical data expressed in numerical values (e.g., student number, primary key).
- In general terms, computers are limited to solely store and retrieve textual information.
- Therefore, auxiliary methods have been developed to obtain numerical data from textual data.

Textual data – What is it?

Data created by humans or machines that is expressed in a natural language

Customer reviews

Social media posts and comments

Conference calls

Text Analysis – How is it accomplished?

- Natural language is difficult to interpret.
- Machines rely on natural language processing to examine textual data.
- In the last decade, great advances in the world of natural language processing have been developed (e.g., ChatGPT, Bard, IBM Watson).
- Lexicons are generally used to identify elements (e.g., part of speech) and emotions (e.g., sentiment analysis)
- There are many API and Python Packages for querying and processing data.

  We will learn how to use an API to query data and a Python Package to process it.

Important terminology in text analysis

- Text can be divided into letters, words, and paragraphs.
- In NLP, we text data is also divided in:

Corpus

Collection of text documents

Lexicon

List of words with meanings

Token

A word or element of a word

Natural Language Toolkit (NLTK) – Python Package

- NLTK is a powerful Python package that has over 50 corpora and lexical resources.
- NLTK is the Pandas of text analysis.
- To install NLTK, you need to execute the following:

  pip install -U nltk

- Then, depending on what you will utilize, you will be required to download lexicons and corpora.

Natural Language Toolkit (NLTK) – Python Package

- If you have space in your computer, you could download all by executing the following commands:

  import nltk

  nltk.download('all') #Alternatively nltk.download('package-name')

- Or you can install them on demand by paying attention to what errors you get from executing some python statement.

```
  Resource punkt not found.
  Please use the NLTK Downloader to obtain
ce:

  >>> import nltk
  >>> nltk.download('punkt')
```

NLTK - Tokenization

- The following code shows a Python script to tokenize textual data.

```
import nltk

reviews = open('ice_cream_reviews.txt')

for review in reviews:
    # print(review)
    tokens = nltk.word_tokenize(review)
    print(tokens)
```

NLTK – Part of speech (POS)

- First you need to have decomposed your textual data in tokens.
- Then, you can execute the following code:

```
tokens = nltk.word_tokenize(review)

pos_tags = nltk.pos_tag(tokens)
print (pos_tags)
```

NLTK – Part of speech (POS) dictionary

CC	coordinating conjunction
CD	cardinal number
DT	determiner
FW	foreign word
IN	preposition or subordinating conjunction
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
MD	modal verb
NN	noun, singular or mass
NNP	proper noun, singular
NNS	noun, plural
PRP	personal pronoun
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
VB	verb, base form
VBD	verb, past tense
VBG	verb, gerund or present participle
VBN	verb, past participle
VBP	verb, non-3rd person singular present
VBZ	verb, 3rd person singular present
WDT	wh-determiner
WRB	wh-adverb
WP	wh-pronoun

NLTK – Searching for all adjectives in a review

```
pos_tags = nltk.pos_tag(tokens)
for tag in pos_tags:
    if tag[1] == 'JJ' or tag[1] == 'JJS':
        print(tag[0])
```

NLTK – Stop words

- Natural languages content can be simplified easily if we remove words that appear often but carry little value to a conversation.
- This words are called stop words and when removed can make the analysis of text more efficient.
- To get a list of the stop words you can execute the following code after importin NLTK:

  from nltk.corpus import stopwords

  print(stopwords.words('english'))

- See the examples below of text processed by removing stop words:

```
Que Rico! The server was the sweetest lady!! Makes the best mangonadas!! Absolutely amazing flavor selection. Open space, bright and lovely ice cream theme...
```

```
Que Rico ! The server sweetest lady ! ! Makes best mangonadas ! ! Absolutely amazing flavor selection . Open space , bright lovely ice cream theme ...
```

NLTK - Stop Words

```
import nltk
from nltk.corpus import stopwords

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

```
new_text = []
for tag in pos_tags:
    if tag[0] not in stop_words:
        print(tag[0])
        new_text.append(tag[0])

print("\nOriginal")
print(review)
print("\nNew")
print(" ".join(new_text))
```

VADER
(Valence Aware Dictionary and sEntiment Reasoner)

To install execute the following command:

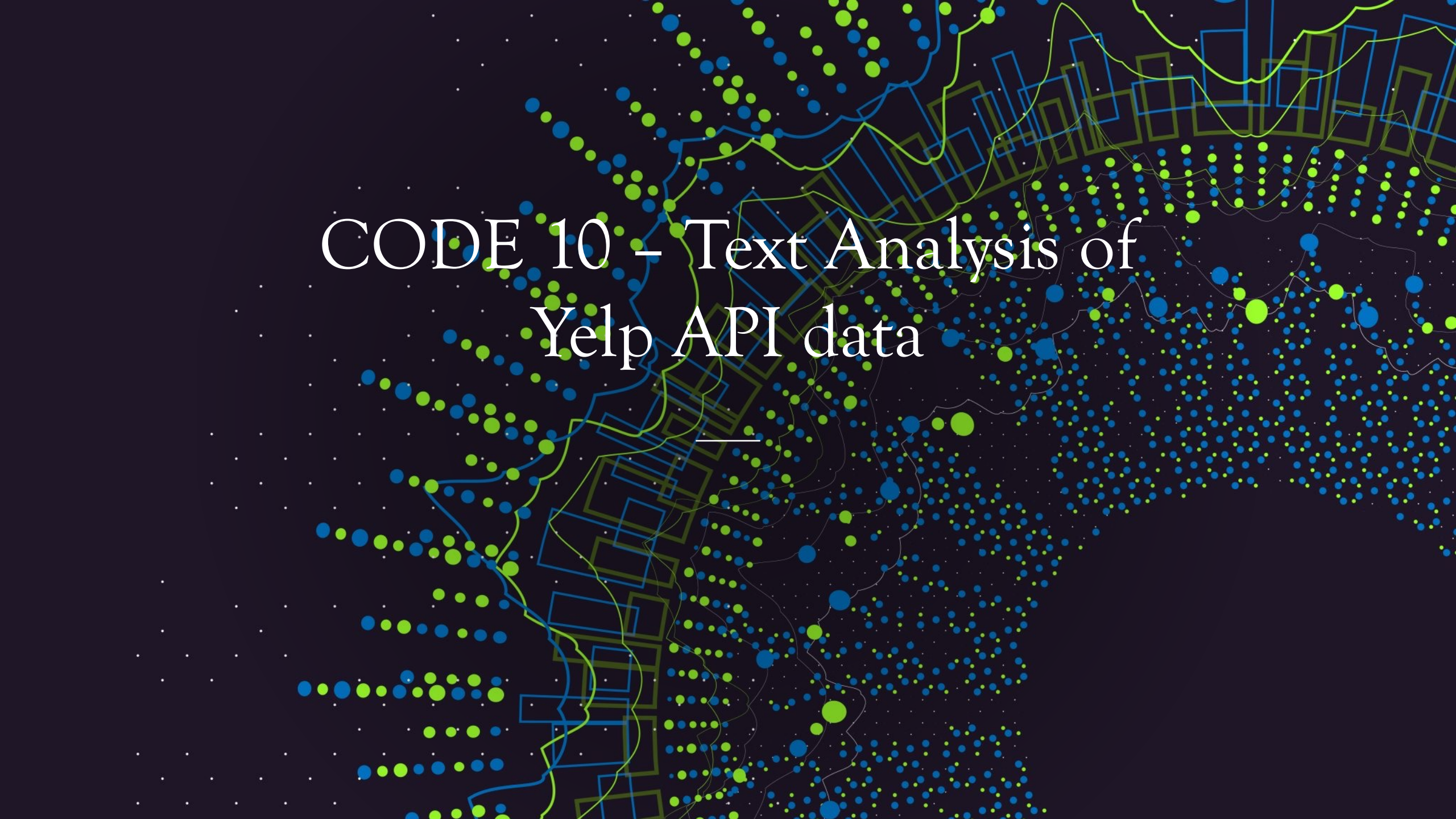
pip install vaderSentiment

Source Code

https://github.com/cjhutto/vaderSentiment

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
reviews = open('ice_cream_reviews.txt')

for review in reviews:
    sentiment_score = analyzer.polarity_scores(review)
    print(review)
    print(sentiment_score)
    print('\n')
```

```
The man behind the counter had the patience of a monk when dealing with my daughter
 who couldn't make up her mind what she wanted.  The ice cream and...

{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Que Rico! The server was the sweetest lady!! Makes the best mangonadas!! Absolutely
 amazing flavor selection. Open space, bright and lovely ice cream theme...

{'neg': 0.0, 'neu': 0.553, 'pos': 0.447, 'compound': 0.9528}
Do you want a different ice cream place with original flavors. ????    this is it I
 tried the pine nut , coffee and velvet ....the three ice cream's scoops...

{'neg': 0.0, 'neu': 0.856, 'pos': 0.144, 'compound': 0.5514}
```



CODE 10 - Text Analysis of Yelp API data

Deliverables

Deliverables

In a Word or PDF, deliver a report of your analysis. Your report must include the following sections.

1. Plan for data retrieval and analysis

- Offer a quick summary of the insights you want to discover using the Yelp Fusion API
- Explain the search parameters (e.g., term, location, sort criteria) that you plan to use
- Explain how you plan to use that information from the Yelp Fusion API to provide the business insights you want to discover
- Explain what text analysis (e.g., sentiment, most used words) you plan to use for obtaining the desired business insights

2. Text insights report

- Report the results of your analysis
- Offer a conclusion with the information you obtained from the analysis

You need to submit the report in Blackboard and in your code repository. Finally, **do not forget** to submit all the code you use for your analysis, and that will be needed to replicate your work in the code repository.

Plan for data retrieval

- You must use the Yelp API to retrieve data about food businesses.
- The Yelp API will allow you to search businesses by “keywords” and location “El Paso, TX”
- Additionally, it allows you to retrieve the most recent three reviews from a business.
- Your job is to plan what data you can obtain for starting a business in the food industry.
- For example, I want to open a new ice cream shop. Then I can learn from knowing my competition and what customers are talking about them.