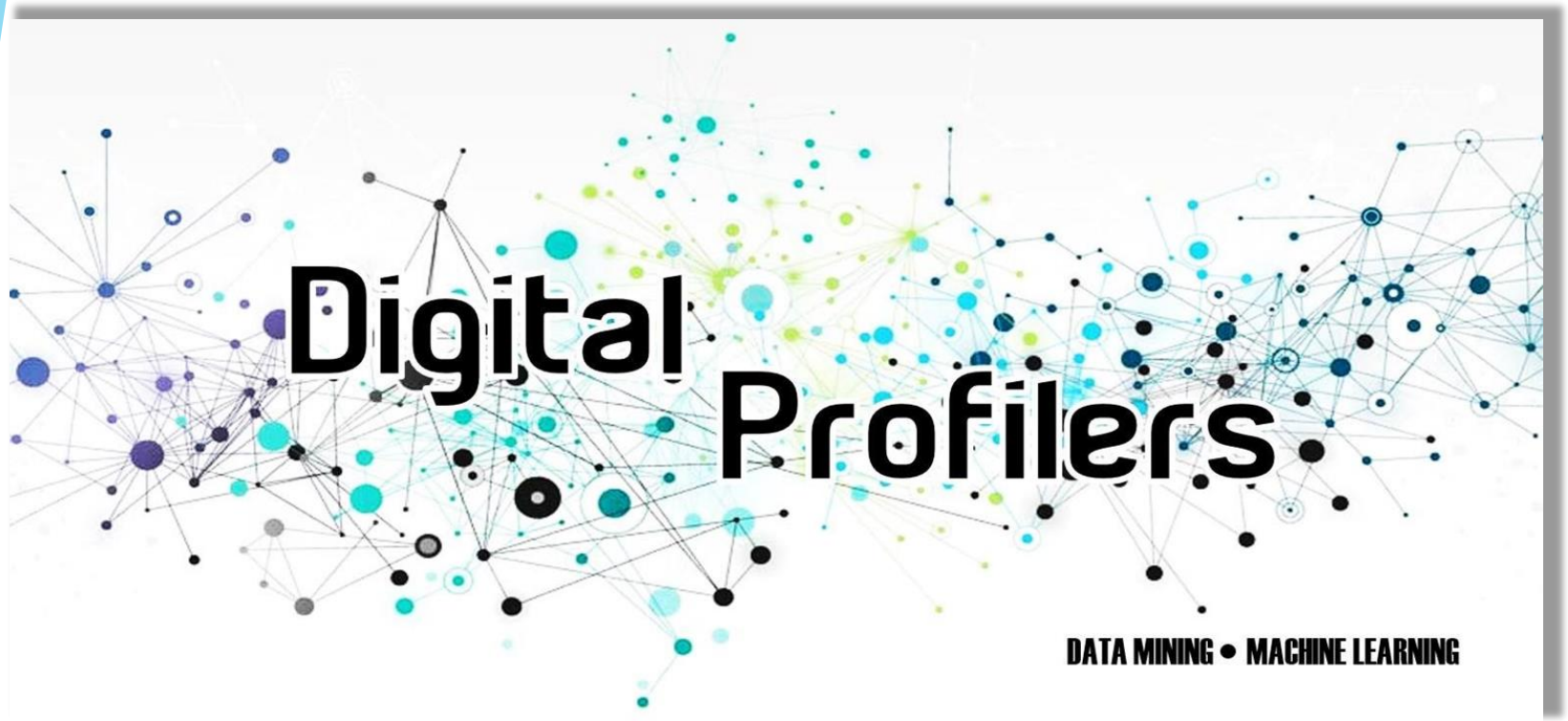


Team: Austin Sass, Lise Welch, Paden Rumsey, Seth Forrest

Clients: Dr. Saied Hemati, Dr. James Frenzel

Lead Instructor: Bruce Bolden

Purpose

- Design a process that will data mine social media streams and index that data in a database. Input that data into IBM Watson's Personality Insights and use the results to train a machine learning algorithm to do personality analysis based on the Five Factor Model: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN).

Requirements

Table 1. Data/Web Mining Requirements

No.	Need	Metric
1	Text Sample	1200-3000 words
2	Quality of Text	Raw/Unprocessed
3	Author	Singular/Separable
4	Retain Information	Copyable to Repository
5	Semi-Automated	10000's tweets/crawl
6	Location Retention (ie the URL)	Scraper Pulls URL

Table 2. Database Requirements

No.	Need	Metric
1	Text Storage	1,000,000's of Entries
2	Links to local copy	Cataloged
3	Author of Text	Cataloged
4	Watson's Output	Numerics/Classifications indexed

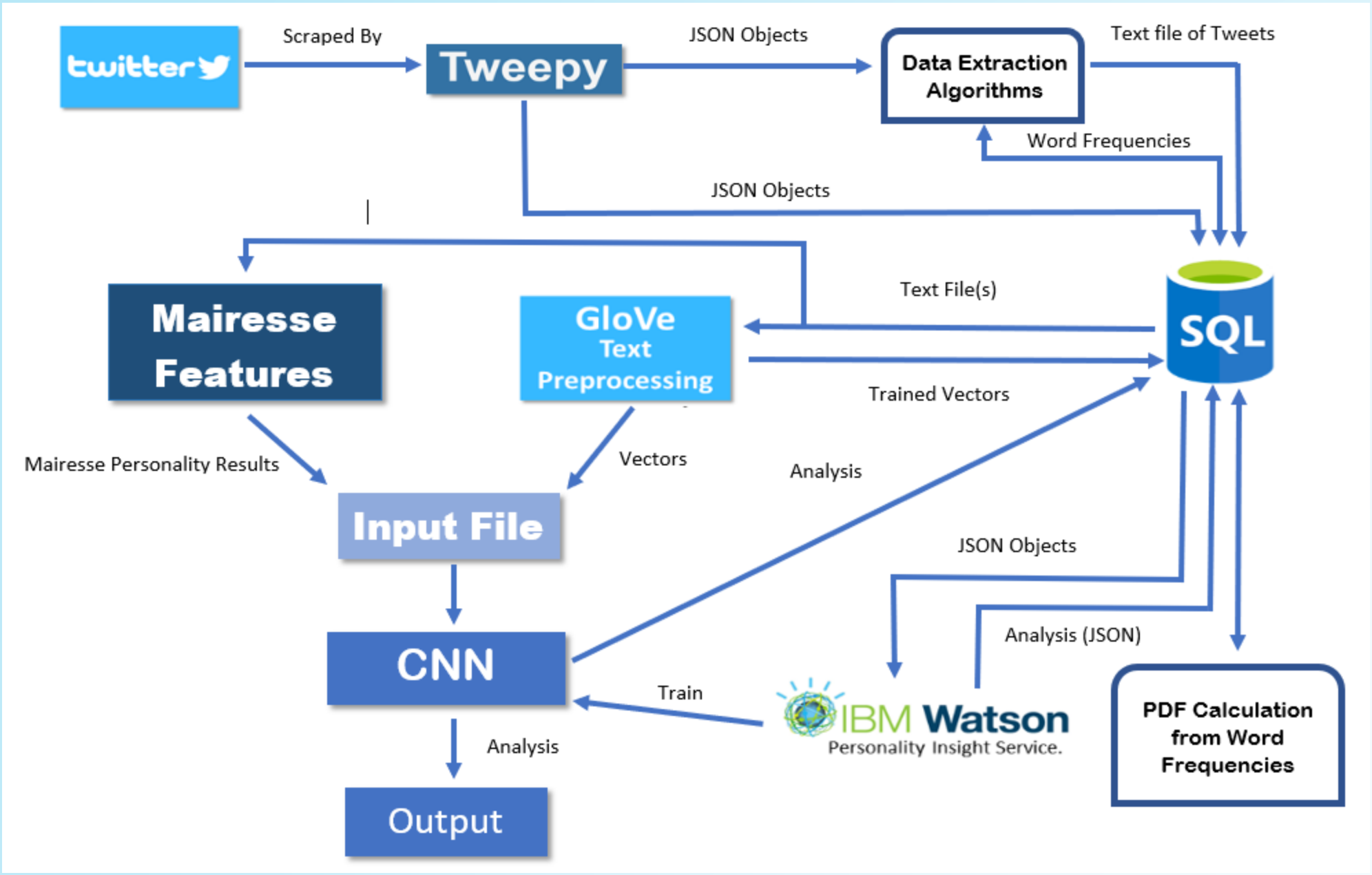
Requirements

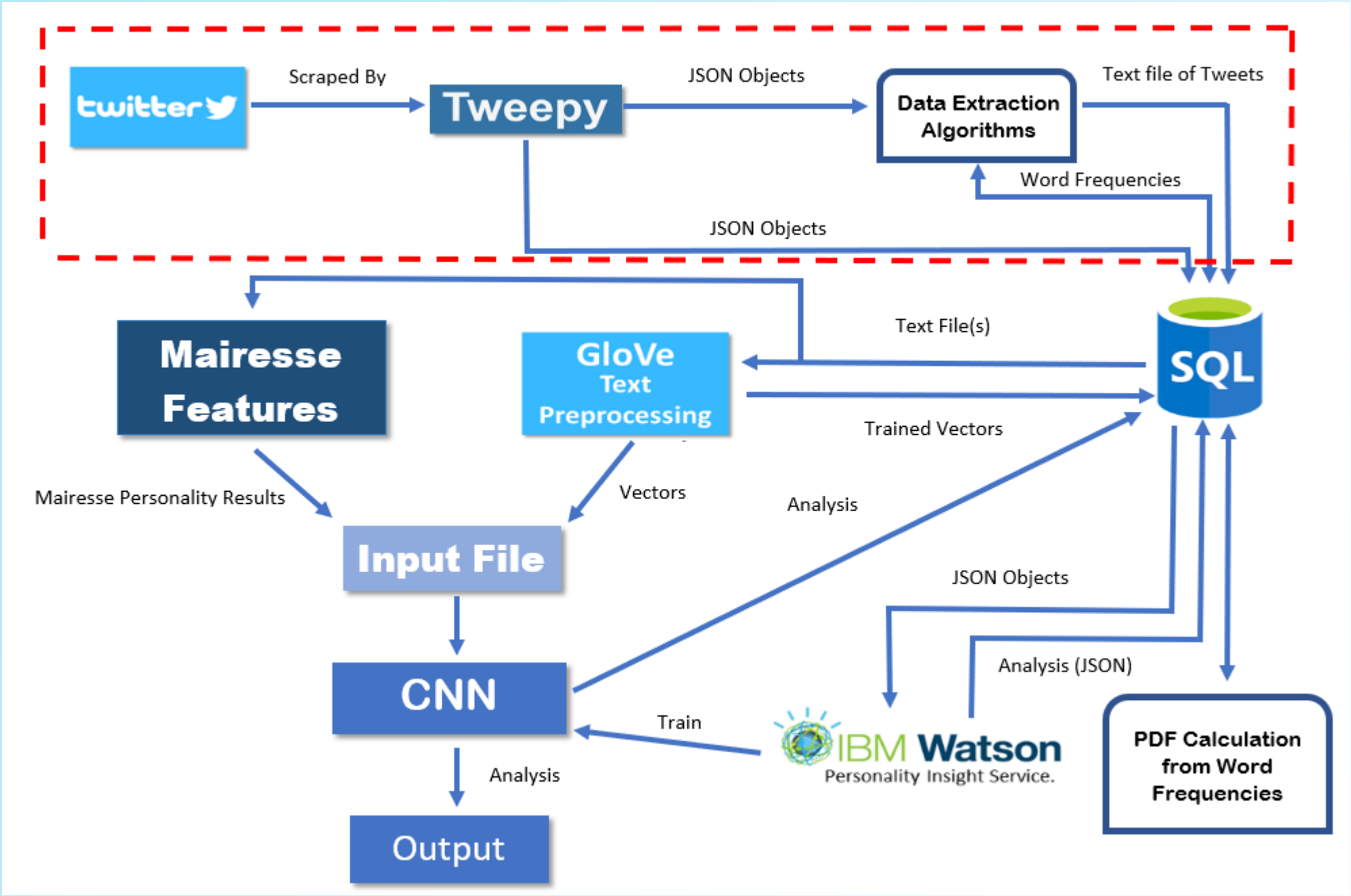
Table 3. Trained Machine Requirements

No.	Need	Metric
1	Process Entries	1,000,000's to Train
2	Approximate Watson's Output	12.3 Mean Average Error

Table 4. Miscellaneous

No.	Need	Metric
1	Probability Density Function for Dictionary Adjectives	# Adjectives in Dict.

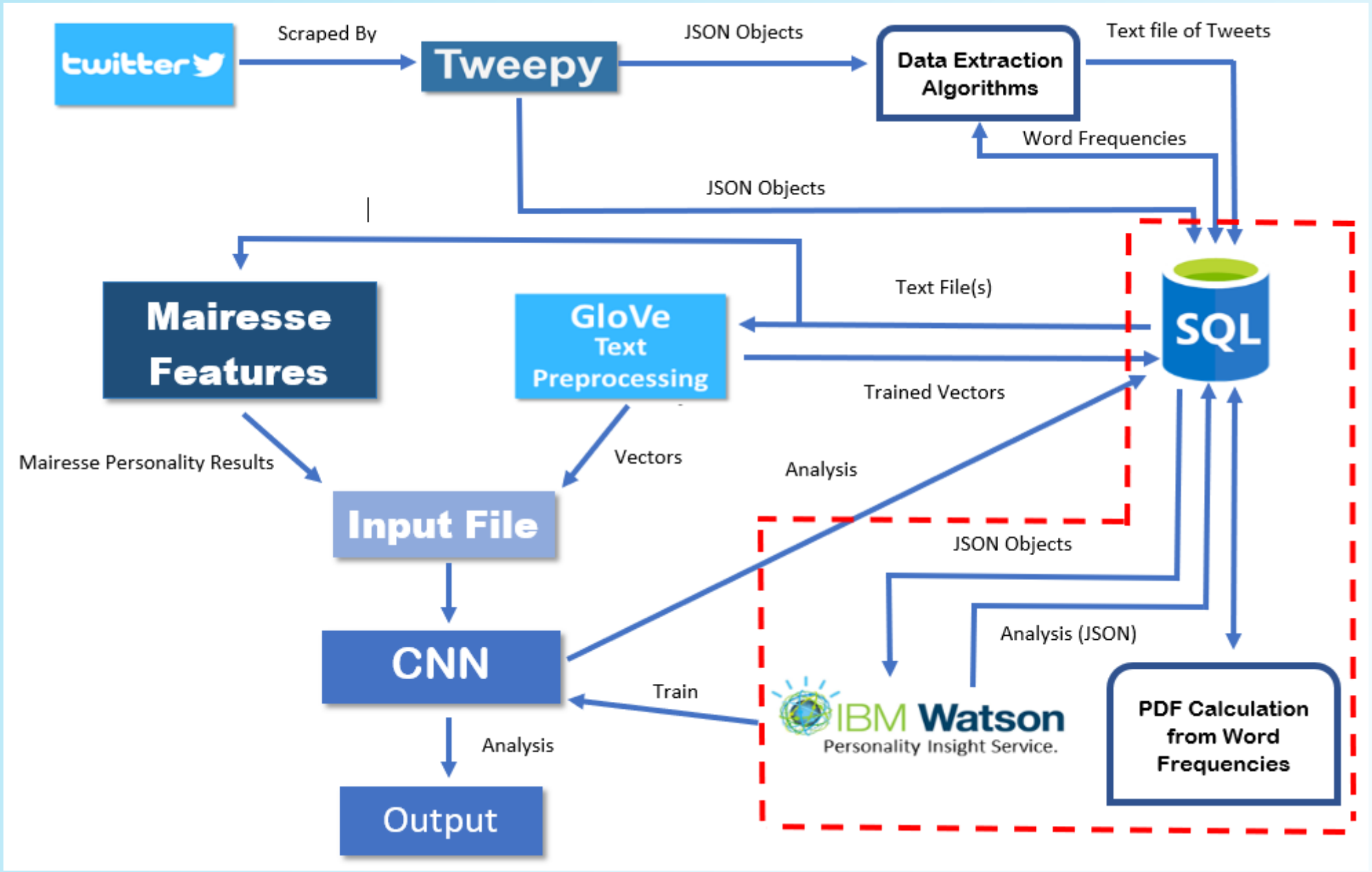

twitter
Scraped By
Tweepy
JSON Objects
Data Extraction Algorithms
Text file of Tweets
Word Frequencies
JSON Objects
Mairesse Features
GloVe Text Preprocessing
Text File(s)
SQL
Trained Vectors
Mairesse Personality Results
Vectors
Analysis
Input File
JSON Objects
CNN
Analysis (JSON)
Train
IBM Watson
Personality Insight Service.
PDF Calculation from Word Frequencies
Analysis
Output


twitter
Scraped By
Tweepy
JSON Objects
Data Extraction Algorithms
Text file of Tweets
Word Frequencies
JSON Objects
Mairesse Features
GloVe Text Preprocessing
Text File(s)
SQL
Trained Vectors
Mairesse Personality Results
Vectors
Analysis
Input File
JSON Objects
CNN
Analysis (JSON)
Train
IBM Watson
Personality Insight Service.
PDF Calculation from Word Frequencies
Analysis
Output

Data Mining

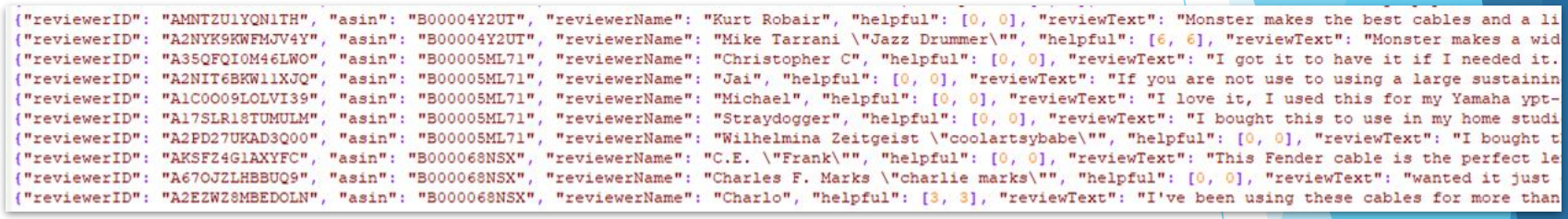
- Extract Data From Twitter
  - Twitter API Limits (450 requests / 15 minutes)
  - Efficiently uses API calls
- Can run multiple copies of script
  - Very low resource cost
- Store user files for analysis (`usr_id.txt`)

```
        Getting Followers
                Skipping: List Full

Accessing user: IridiumBoss
        Getting Tweets
                Writing Tweets
        Getting Followers
                Skipping: List Full

Accessing user: martysolomon
        Getting Tweets
                Writing Tweets
        Getting Followers
                Skipping: List Full

Accessing user: StationCDRKelly
        Getting Tweets
                Writing Tweets
        Getting Followers
                Skipping: List Full

Accessing user: DJSnM
        Getting Tweets
                Writing Tweets
        Getting Followers
                Skipping: List Full

Accessing user: rudyagovic
        Getting Tweets
                Writing Tweets
        Getting Followers
                Skipping: List Full

Accessing user: AstroAllie5
        Getting Tweets
```


twitter
Scraped By
Tweepy
JSON Objects
Data Extraction Algorithms
Text file of Tweets
Word Frequencies
JSON Objects
Mairesse Features
GloVe Text Preprocessing
Text File(s)
SQL
Trained Vectors
Mairesse Personality Results
Vectors
Analysis
Input File
JSON Objects
CNN
Analysis (JSON)
Train
IBM Watson
Personality Insight Service.
PDF Calculation from Word Frequencies
Analysis
Output

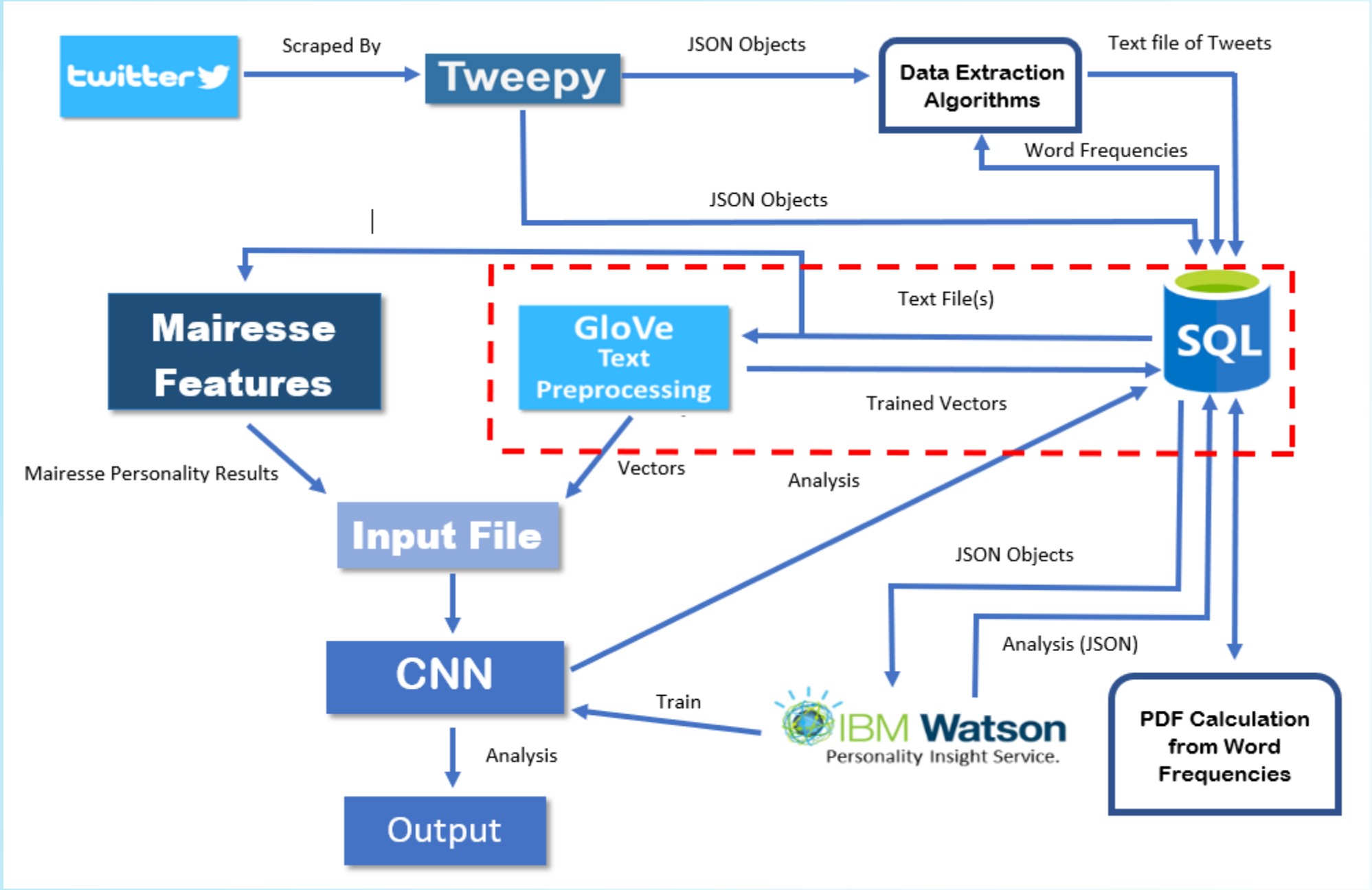
Database Administration

- Data
  - Amazon Reviews
    - Parsing
    - Processing
- Issues
  - Platforms




```
{"reviewerID": "AMNTZU1YQN1TH", "asin": "B00004Y2UT", "reviewerName": "Kurt Robair", "helpful": [0, 0], "reviewText": "Monster makes the best cables and a li
{"reviewerID": "A2NYK9KWFMJV4Y", "asin": "B00004Y2UT", "reviewerName": "Mike Tarrani \"Jazz Drummer\"", "helpful": [6, 6], "reviewText": "Monster makes a wid
{"reviewerID": "A35QFQI0M46LWO", "asin": "B00005ML71", "reviewerName": "Christopher C", "helpful": [0, 0], "reviewText": "I got it to have it if I needed it.
{"reviewerID": "A2NIT6BKW11XJQ", "asin": "B00005ML71", "reviewerName": "Jai", "helpful": [0, 0], "reviewText": "If you are not use to using a large sustainin
{"reviewerID": "A1C0009LOLVI39", "asin": "B00005ML71", "reviewerName": "Michael", "helpful": [0, 0], "reviewText": "I love it, I used this for my Yamaha ypt-
{"reviewerID": "A17SLR18TUMULM", "asin": "B00005ML71", "reviewerName": "Straydogger", "helpful": [0, 0], "reviewText": "I bought this to use in my home studi
{"reviewerID": "A2PD27UKAD3Q00", "asin": "B00005ML71", "reviewerName": "Wilhelmina Zeitgeist \"coolartsybabe\"", "helpful": [0, 0], "reviewText": "I bought t
{"reviewerID": "AKSFZ4G1AXYFC", "asin": "B000068NSX", "reviewerName": "C.E. \"Frank\"", "helpful": [0, 0], "reviewText": "This Fender cable is the perfect le
{"reviewerID": "A67OJZLHBBUQ9", "asin": "B000068NSX", "reviewerName": "Charles F. Marks \"charlie marks\"", "helpful": [0, 0], "reviewText": "wanted it just
{"reviewerID": "A2EZWZ8MBEDOLN", "asin": "B000068NSX", "reviewerName": "Charlo", "helpful": [3, 3], "reviewText": "I've been using these cables for more than
```


File	Text	Extraversion	Neuroticism	Agrreableness	Conscientiousness	Openness
A1AT0GONN4A9NA.txt	John P. Morgan "Light Coach" In 1971 I lived in a very	0.060363772	0.443129891	0.515596807	0.176387207	0.972399822
A1AT95C4D1Q3OF.txt	Zengrrl I love this album. LOVE IT! Even after 8 years	0.247043907	0.300341761	0.175839181	0.134450121	0.997602571
A1ATHDILELXROF.txt	Renee L Williams It was interesting how the roles be	0.008417838	0.406130934	0.192962259	0.248230044	0.99402586
A1ATORINH5WYDP.txt	Kerlyn and from there it just wow'd me. The way the	0.105078233	0.615783295	0.240232864	0.614326242	0.99827
A1AU4WG25D1M1I.txt	Some Guy I've always enjoyed Mythbusters. Glad thi	0.002385607	0.115268139	0.269953258	0.26570745	0.902075539
A1AU5K3JMPS3UQ.txt	Eric Pop filter works great. This really eliminated mai	0.619240138	0.583629563	0.022213322	0.024218327	0.992107126
A1AU80I28289RC.txt	kayla I love this show! I have watched alpha house b	0.008813745	0.261039442	0.117723425	0.033514629	0.964467159
A1AUDW7X02EEPJ.txt	Otto Wow these cables are made really sturdy. And t	0.773440161	0.759180967	0.040941533	0.311659725	0.957478829
A1AUOTPLQBK0RB.txt	virgil walker one great record one bad track on hold r	0.394294586	0.656362027	0.102803819	0.027916653	0.861778121
A1AVF90QEZHCYV.txt	Anastasia M. Policicchio easy to set up but they got tl	0.031511613	0.802466688	0.055517561	0.900339762	0.713746842
A1AVOKBIPDODKY.txt	G. Childers They should've created feature films for r	0.289188474	0.577520422	0.371925709	0.283292419	0.998573715
A1AWXXBQRQN13I.txt	Lytle3891 I ordered a couple sets of this strings so I w	0.336496145	0.694507142	0.076012391	0.378376715	0.935596821


twitter
Scraped By
Tweepy
JSON Objects
Data Extraction Algorithms
Text file of Tweets
Word Frequencies
JSON Objects
Mairesse Features
GloVe Text Preprocessing
Text File(s)
SQL
Trained Vectors
Mairesse Personality Results
Vectors
Analysis
Input File
JSON Objects
CNN
Analysis (JSON)
Train
IBM Watson
Personality Insight Service.
PDF Calculation from Word Frequencies
Analysis
Output

GloVe

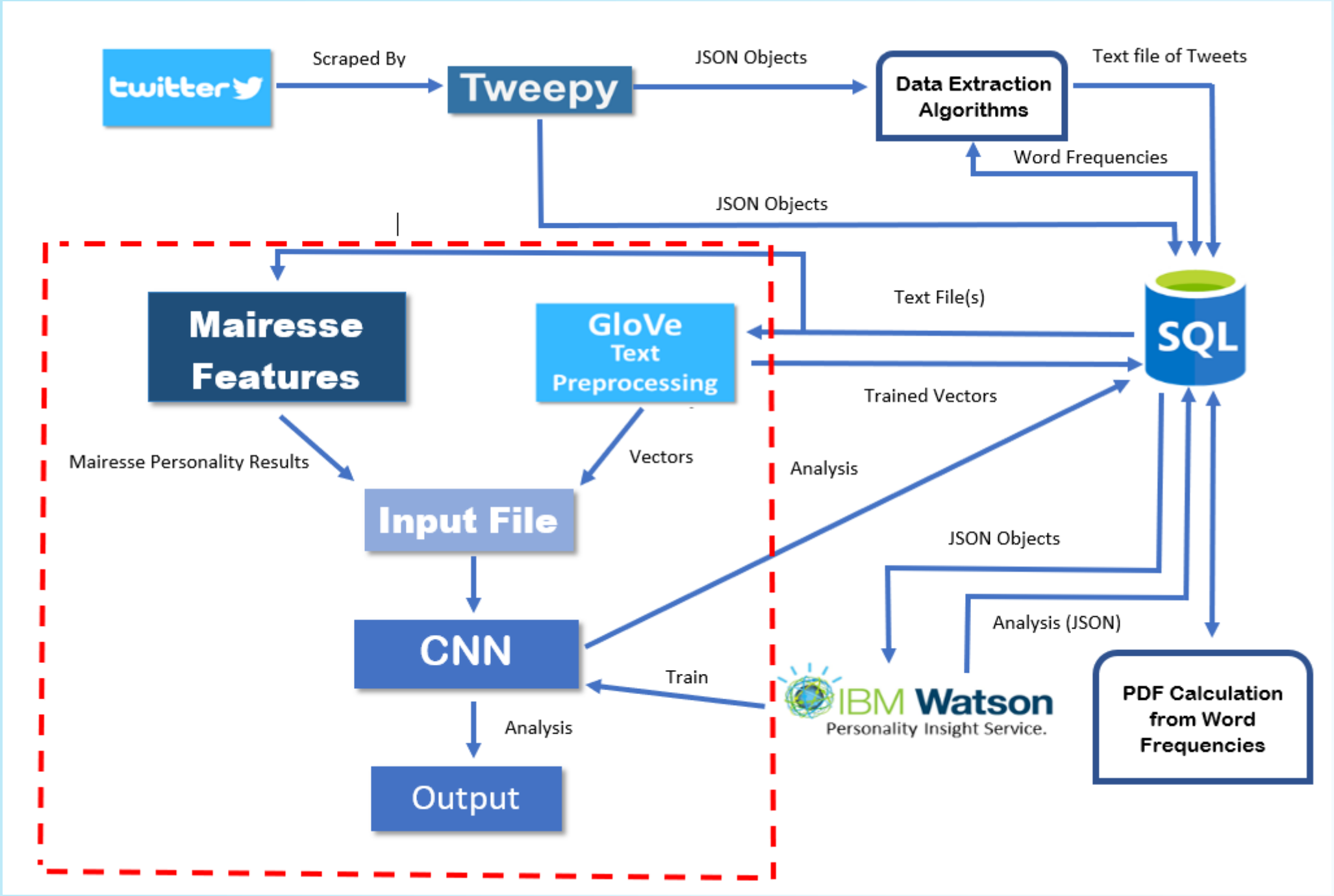
Global Vectors for Word Representation (GloVe)

Word-word co-occurrence probabilities are used to encode semantic meaning into feature vectors:

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Global Vectors for Word Representation (GloVe)

- Possibility of replacing randomly generated vectors in machine to achieve better results
- Machine can also train vectors directly -> Consider using GloVe as an optional module


twitter
Scraped By
Tweepy
JSON Objects
Data Extraction Algorithms
Text file of Tweets
Word Frequencies
JSON Objects
Mairesse Features
GloVe Text Preprocessing
Text File(s)
SQL
Trained Vectors
Mairesse Personality Results
Vectors
Analysis
Input File
JSON Objects
CNN
Analysis (JSON)
Train
IBM Watson
Personality Insight Service.
PDF Calculation from Word Frequencies
Analysis
Output

Machine Learning

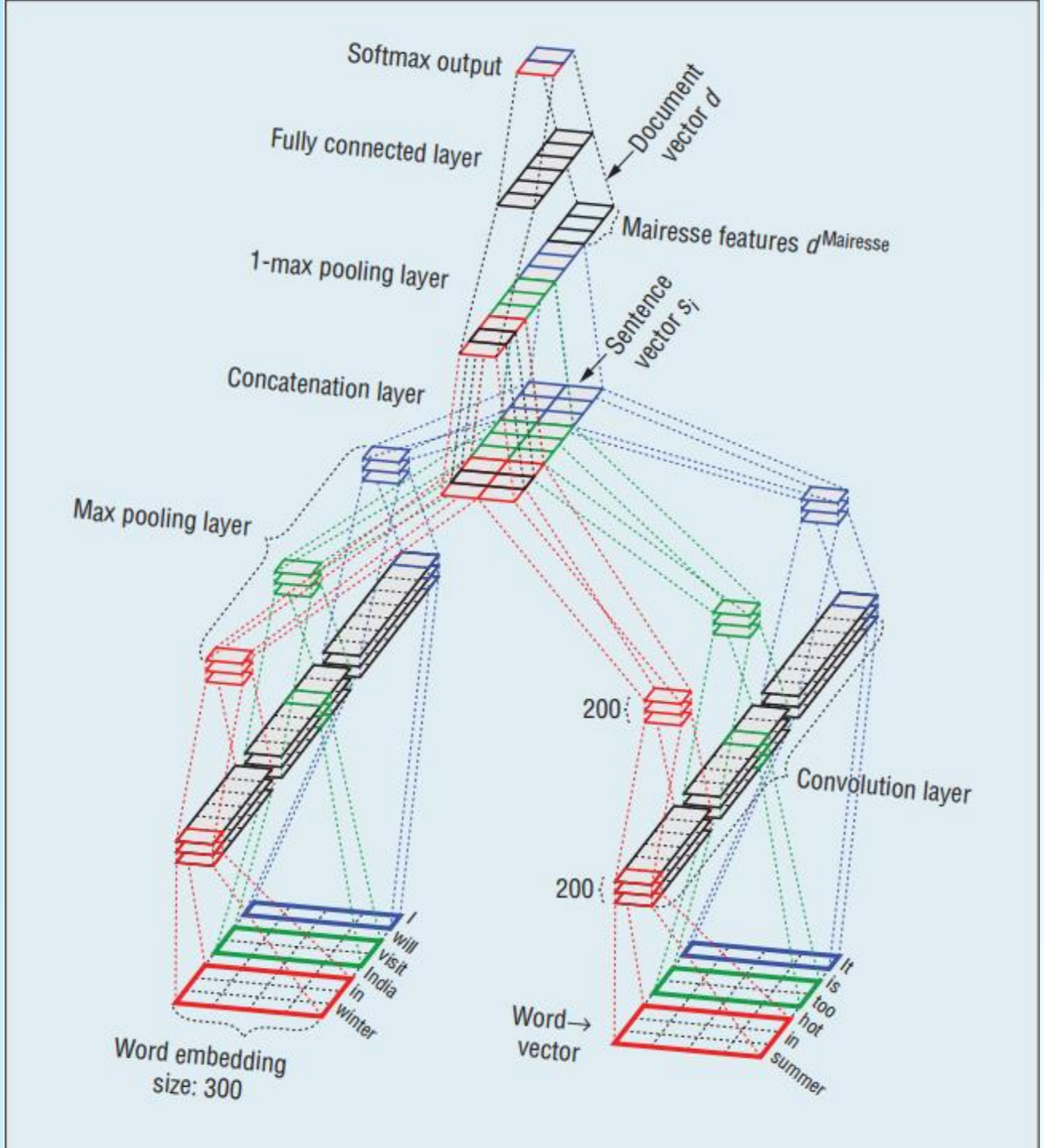
Choice - Convolutional Neural Network

- Ultimately decided that the CNN was the best use tool to fulfill our requirements
- Based on approach described in “Deep Learning-Based Document Modeling for Detection from Text” by Navonil Majumder, et al.
- Employs *word2vec* technology same as GP solution

Word vectorization -> Sentence Vectorization -> Document Vectorization -> Sigmoid Activation -> Classify

Mairesse Features

- Acts a secondary input to the CNN (appended to sentence vectors)
- Utilizes Linguistic Inquiry Word Count and MRC (Medical Research Council) Psycholinguistic database to score phrasing in text
- Then processes it through a WEKA model based on four different modes
  - 1 = Linear Regression
  - 2 = M5' Model Tree
  - 3 = M5' Regression Tree
  - 4 = Support Vector Machine with Linear Kernel
- Output for each trait is based on a score from 1 to 7


Softmax output
Fully connected layer
Document vector d
Mairesse features d^Mairesse
1-max pooling layer
Sentence vector s_i
Concatenation layer
Max pooling layer
200
200
Convolution layer
I will visit India in winter
It is too hot in summer
Word embedding size: 300
Word→ vector

Problems Overcame

- CNN implementation only usable on NVIDIA graphics card
- Theano (framework CNN written in) utilizes unspoken mechanics
- Rewrite CNN to do prediction easily (use classification as a test)

Remaining Issues

- 104 hours to train network on all five traits using 2400 essays
- Classification ----> Regression

Outcome #1: Convergence for Regression

- Gradient for loss function converges and CNN approximates Watsons output
- PDF's can be calculated on CNN data. Regression data can be generated for new samples and added to final compilation

Outcome #2: Non-Convergence for Regression

- Model never converges on loss function and Watson's output is not approximated (unlikely that we will know the absolute cause).
- PDF's can still be calculated from Watson data that is generated
- Mairesse Data, CNN classification, Watson data is what we have

Spring Schedule

Section	Monthly Goal (February)
Text Scraper	Continue to collect data for Machine Learning algorithms while implementing a cleaning algorithm for current data
PDF's	Write script to calculate PDF's from generated data
Machine Learning	Finish generating classification output from Amazon/Watson samples (initial results)

Spring Schedule

Section	Monthly Goal (March - Mid-April)
Text Scraper	Continue to collect data for Machine Learning algorithms while implementing a cleaning algorithm for current data
PDF's	Test PDF on Watson data to show that it works (use on CNN data if applicable)
Machine Learning	Determine whether CNN can be altered to regression technique. (if so apply to data)
Documentation	Finish design document with full CNN implementation/description of any remaining issues

Spring Schedule

Section	Monthly Goal (Mid-April - Mid-May)
Text Scraper	Finish cleaning data
Documentation	Compile document that records scores for Watson score, CNN classification score, Mairesse Score and CNN regression (if applicable)

References

- [1] Adedoyin-Olowe, Mariam, et al. “A Survey of Data Mining Techniques for Social Media Analysis.” 2013, Journal of Data Mining & Digital Humanities, 2014 (June 24, 2014) jdmdh:18.
- [2] Arnoux, Pierre-Hadrien, et al. “25 Tweets to Know You: A New Model to Predict Personality with Social Media.” 2017. https://arxiv.org/ftp/arxiv/papers/1704/1704.05513.pdf
- [3] Digman, John M. ”Personality structure: Emergence of the five-factor model. Annual Review of Psychology” 1990. 41.1: 417–440.
- [4] Kalghatgi, M. P., et al. “A Neural Network Approach to Personality Prediction based on the Big-Five Model.” International Journal of Innovative Research in Advanced Engineering, vol. 2, no. 8, 2015, pp. 56–63.
- [5] Kaur, Arvinder, and Deepti Chopra. “Comparison of Text Mining Tools.” Institute of Electrical and Electronics Engineers, 7 Sept. 2016, pp. 186–192. ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7784950.
- [6] Majumder, Navonil, et al. “Deep Learning-Based Document Modeling for Personality Detection from Text.” IEEE Intelligent Systems, vol. 32, no. 2, 2017, pp. 74–79., doi:10.1109/mis.2017.23.
- [7] Mnih, Andriy, Hinton, Geoffrey. ”Three New Graphical Models for Statistical Language Modelling.” 24th International Conference on Machine Learning, 2007
- [8] Pennington, Jeffrey, Socher, Richard, Manning D., Christoper. ”GloVe: Global Vectors for Word Representation.” 2014. https://nlp.stanford.edu/pubs/glove.pdf.
- [9] Rasmussen E, Carl, Williams K. I., Christoper. ”Gaussian Processes for Machine Learning.” 2006. http://www.gaussianprocess.org/gpml/.

Questions?