Introduction:

Classifying restaurant reviews into positive or negative using natural language processing, big data analytics and machine learning classification algorithms.

Requirements

1. PYSPARK
2. Jupiter Notebook
3. NLTK
4. Pandas Data frame
5. Bokeh

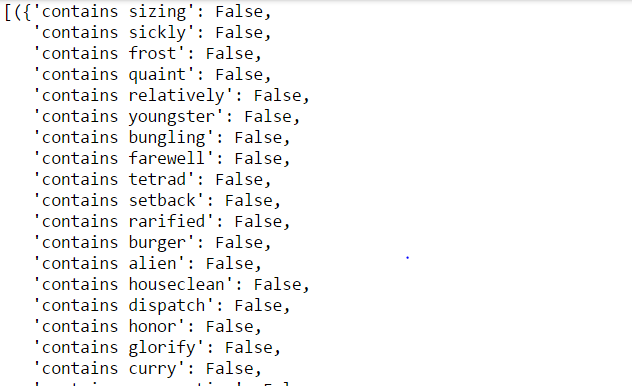
Dataset:

YelpRestaurant reviews: [yelp-dataset](https://yelp-dataset.s3.amazonaws.com/yelp_dataset_challenge_academic_dataset.tgz?Signature=HKewuzljqukP3R7bEJ6Hi7P519Q%3D&Expires=1481679957&AWSAccessKeyId=AKIAJ3CYHOIAD6T2PGKA)

Implementation:

1. Load restaurant reviews json dataset on hdfs
2. Load Data in pyspark rdd
3. Split reviews (sentences) into words
4. Using NLTK (Natural language processing Python module) get synonyms of each word.
5. Create feature vector which contains review data and synonyms of each word
6. Split certified data into train and test data.
7. Train NLTK and skit learn classifiers using train Data
8. Custom classifier combining all classifiers
9. Find accuracy of models using labeled test data
10. Classify data using classification models generated using.
11. Visualize classification of data based on different models

Feature Vector:



Sample Output

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **text** | **NaiveBayes** | **RF** | **BNB** | **Logistic** | **MNB** | **SGD** | **LinearSV** | **Combined** |
| My wifey & I decided to venture | negative | negative | negative | negative | negative | negative | negative | negative |
| Beer is awesome! | negative | negative | negative | negative | negative | negative | negative | negative |

Visualization of Output

Conclusion:

Text data has number of records compared to number of elements in feature vector. Learning algorithm converges slowly.

Future Scope:

1. Apply classification based on sub categories by considering the words in the text.