Sentiment Analysis of Reviews on Yelp Dataset

By - Prashant Chhabra, Nathan Mots, Melina Victoria Sparks

Why?

None of us did NLP Before.

Curious to know how efficient NLP is.

Academic dataset challenge is going on.

Dataset Description

- Dataset available as dump of json files.
- Relevant fields of relevant Json files

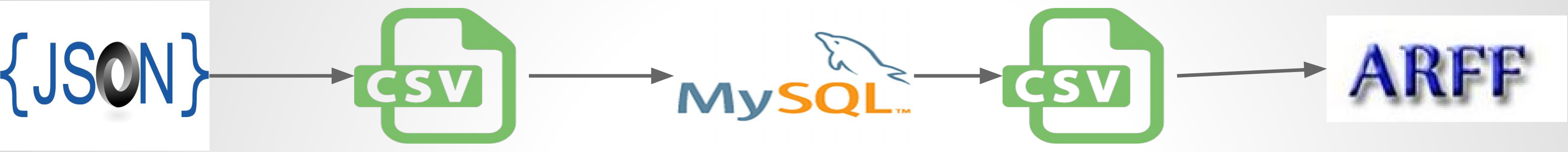
Business.json

```
{
'type': 'business',
'business_id': (encrypted business id),
'attributes': {
(attribute_name): (attribute_value),
...
}
```

Review.json

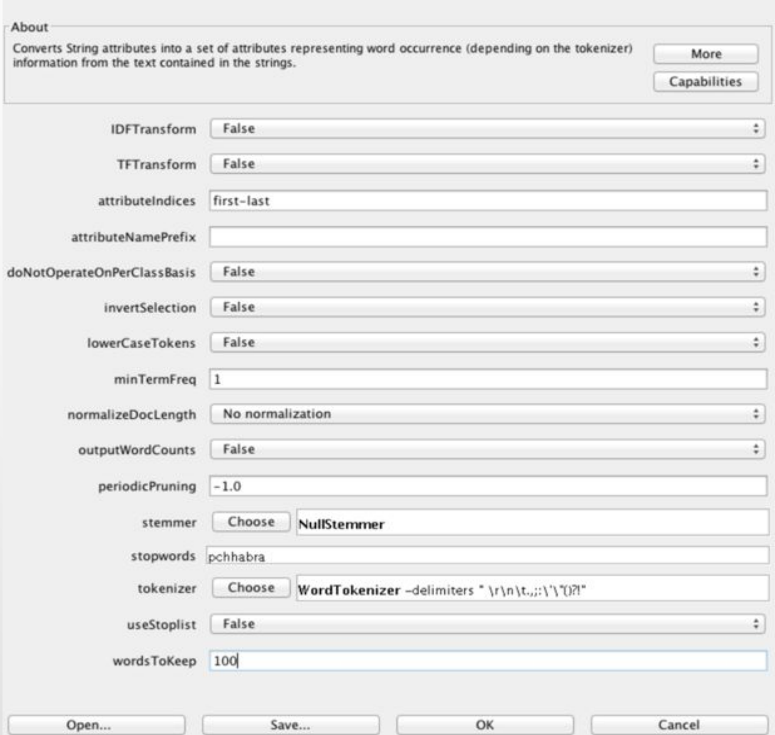
```
{
'business_id': (encrypted
business id),
'stars': (star rating,
rounded to half-stars),
'text': (review text),
}
```

Data Preprocessing



Data Preprocessing contd.(Weka)

String to word vector(Converts String to attributes representing word occurrence). Tried with default settings.



About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

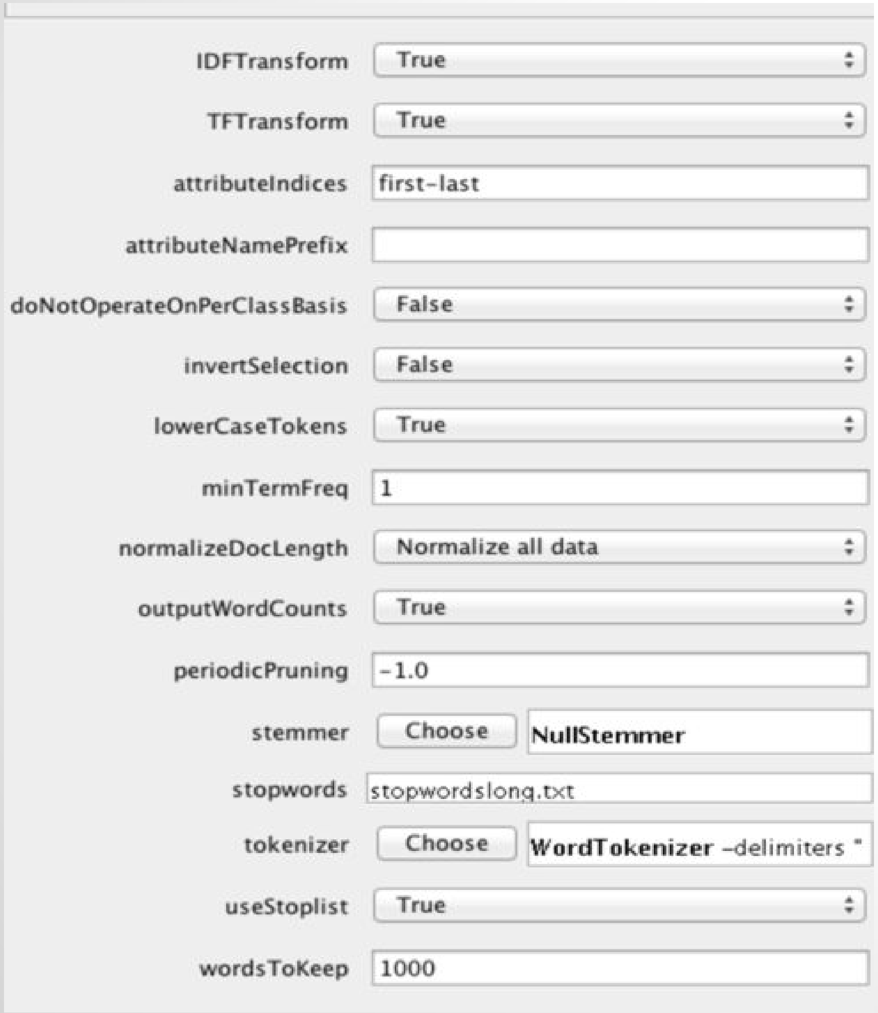
IDFTransform	False
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	False
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
stemmer	NullStemmer
stopwords	pchhabra
tokenizer	WordTokenizer –delimiters " \r\n\t.,;:\'\"()?!"
useStoplist	False
wordsToKeep	100

Classification

Try Algorithms with 5 fold cross Validation.

1. Naive Bayes - 72.32%
2. Multinomial Naive Bayes - 80.52%
3. K Nearest Neighbor(5) - 73.16%
4. J48 (Decision Tree) - 76.86%
5. Random Forest - 82.89%
6. Support Vector Machine - 83.79%

Play around with String to word Vector

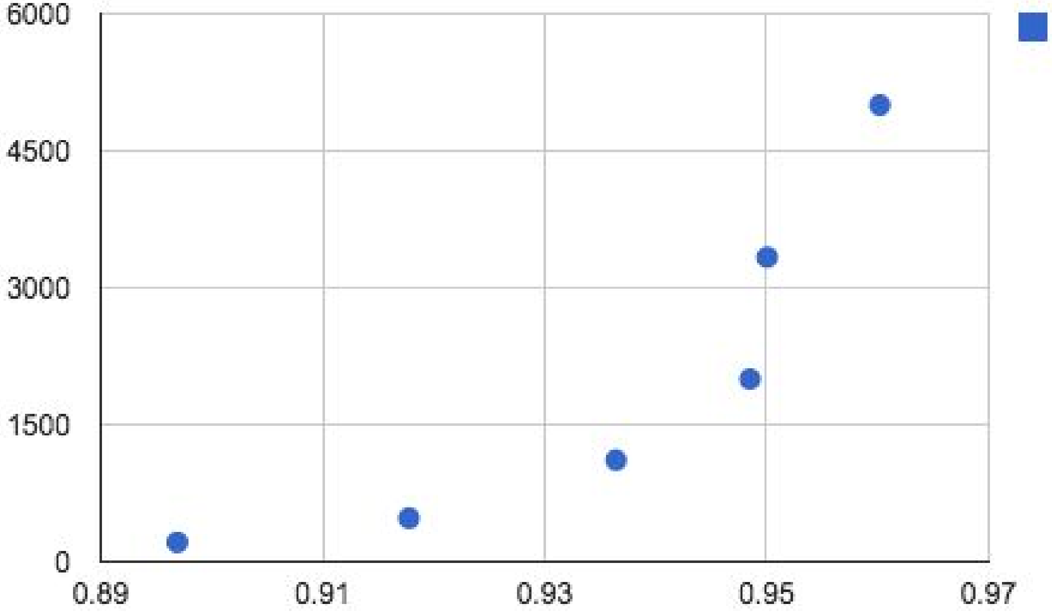


IDFTransform	True
TFTransform	True
attributeIndices	first-last
attributeNamePrefix	
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	1
normalizeDocLength	Normalize all data
outputWordCounts	True
periodicPruning	-1.0
stemmer	Choose NullStemmer
stopwords	stopwordslong.txt
tokenizer	Choose WordTokenizer –delimiters "
useStoplist	True
wordsToKeep	1000

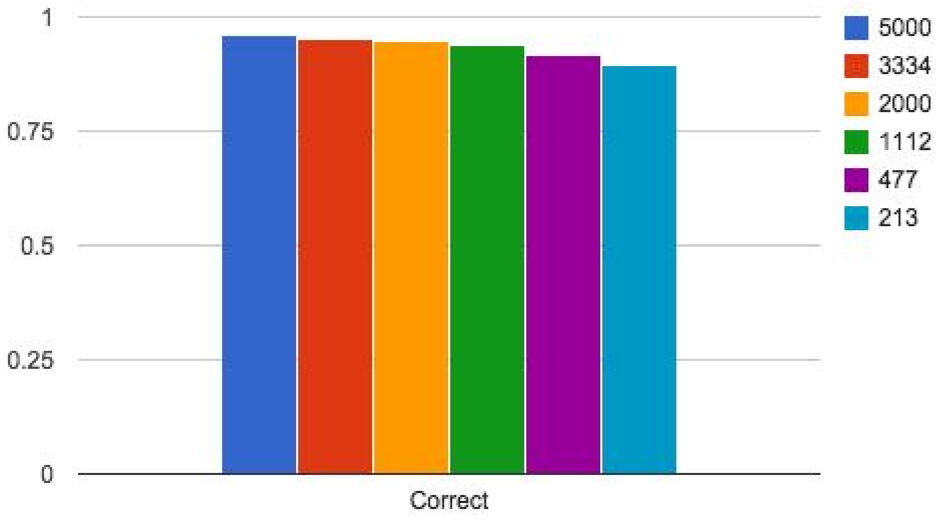
94.64%
Accuracy

LingPipe Java Library

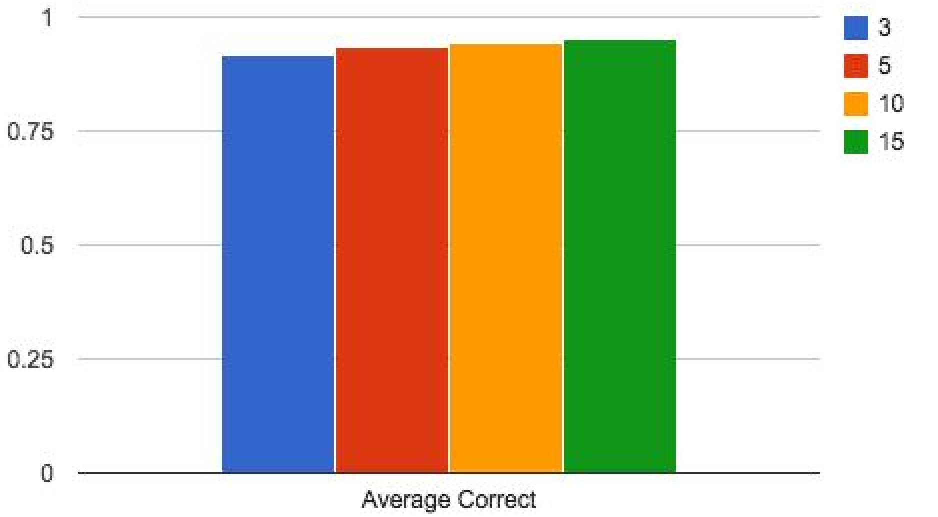
Percentage correct versus number of test cases.



Percent success compared to number of test cases



Average Correct Evaluated per Fold Count



Questions?