Comp 550 - Assignment 1

Anand Kamat
ID# 260773313

Question 1:

1. **Every student read a book**
    - The most salient type of ambiguity is 'Semantic Ambiguity (Lexical)' due to multiple meanings of the word 'read'. It can be argued that 'Phonological ambiguity' is also a possibility as the sound of the word causes the difference in interpretation.
    - The phonological ambiguity arises due to the word 'read' which can be perceived both in present tense as well as the past tense. Due to this the sentence has two possible interpretations. The first one can mean that every student has performed the action of 'reading' the book indicating past tense whereas the other meaning can be inferred as an instruction of suggestion to every student to currently perform the act of 'reading' the book indicating present tense.
    - I believe the correct interpretation can be derived by two methods. The pronunciation of the word if the statement is verbally communicated can help the machine or human assume the context (helps resolve phonological ambiguity). If the statement was communicated via text a more deeper analysis in the context is required. For example, understanding who said it (person saying it to students implies present tense whereas a person just talking about students who read books would imply past tense).

2. **The lion is a majestic animal**
    - The most salient type of ambiguity observed here would be the Semantic Ambiguity due to multiple meanings of the word 'the' which changes the context of the sentence.
    - The ambiguity arises from the fact that 'the' can refer to one particular lion in mind or lions in general. Depending on the situation multiple meanings of the phrase is generated.
    - To disambiguate this sentence it would help the machine/human to know the situation/location which prompted such a statement. If this was in a situation where the person is talking about one lion we would consider the first meaning, else the second. Another way to disambiguate would be to consider the most probable instance. We know that the phrase is majorly used to indicate the lion as the species

(all lion) and not just one particular lion. Knowing this can disambiguate the statement significantly.

3. **Use of sidewalk is prohibited by police officers**
    - The most salient type of ambiguity is the Syntactic Ambiguity caused by two different interpretations depending on the structure of the sentence.
    - The ambiguity arises due to the change in context of the sentence caused by the structure of the sentence around the word ‘by’. The two possible interpretations of this sentence can be:
        i) The using of sidewalk is prohibited according to the police officers.
        ii) The police officers are prohibited to walk on this sidewalk.
    - A deeper understanding of the context would help the machine/person disambiguate the sentence. One way to do this is to perhaps just use plain logic. In majority of such cases we would observe that it would hold the first interpretation instead of the second. If the machine is realized the common logical notions it can better understand the context. Another way might be to use additional punctuations or formatting. For example the statement below diminishes the ambiguity significantly:-

      The use of sidewalk is prohibited. ~By police officers.

4. **My English teacher recently recovered from a bowel cancer operation... and he tried to show me a semi colon. (Source: The 2016 UK Pun Championship)**
    - The most salient type of ambiguity is the ‘Orthographic’ ambiguity as the two possible interpretations are due to the way ‘semi colon’ is spelled. However, semantic ambiguity can be assumed by observing two possible interpretations of the word ‘semi colon’.
    - The orthographic ambiguity can be observed as the spelling of the word semi-colon and semi colon differs just by a hyphen but mean completely different things. Semi-colon is the punctuation tool (;) whereas semi colon means part of a colon or half a colon. I have stated semantic ambiguity here as well as the words semi-colon and semi colon seem like similar words having vastly different meanings.
    - Disambiguating this can be carried out by making the machine/person understand the pretext to this word. Given that the sentence mentioned a medical procedure, it would make more sense to understand the medical meaning of the word instead of the punctuation mark. If the sentence talked about grammar and punctuation, the word would be written as semicolon. Hence orthographic details can help differentiate between the two interpretations also.

5. **She is my ex-mother-in-law-to-be**
    - The most apt category to place this ambiguity would be ‘Morphological’. This is due to the fact the formation of the word from morphemes can imply different

interpretations. This is also an Semantic (Lexical) ambiguity as multiple meaning of the morpheme ‘ex’ is observed.

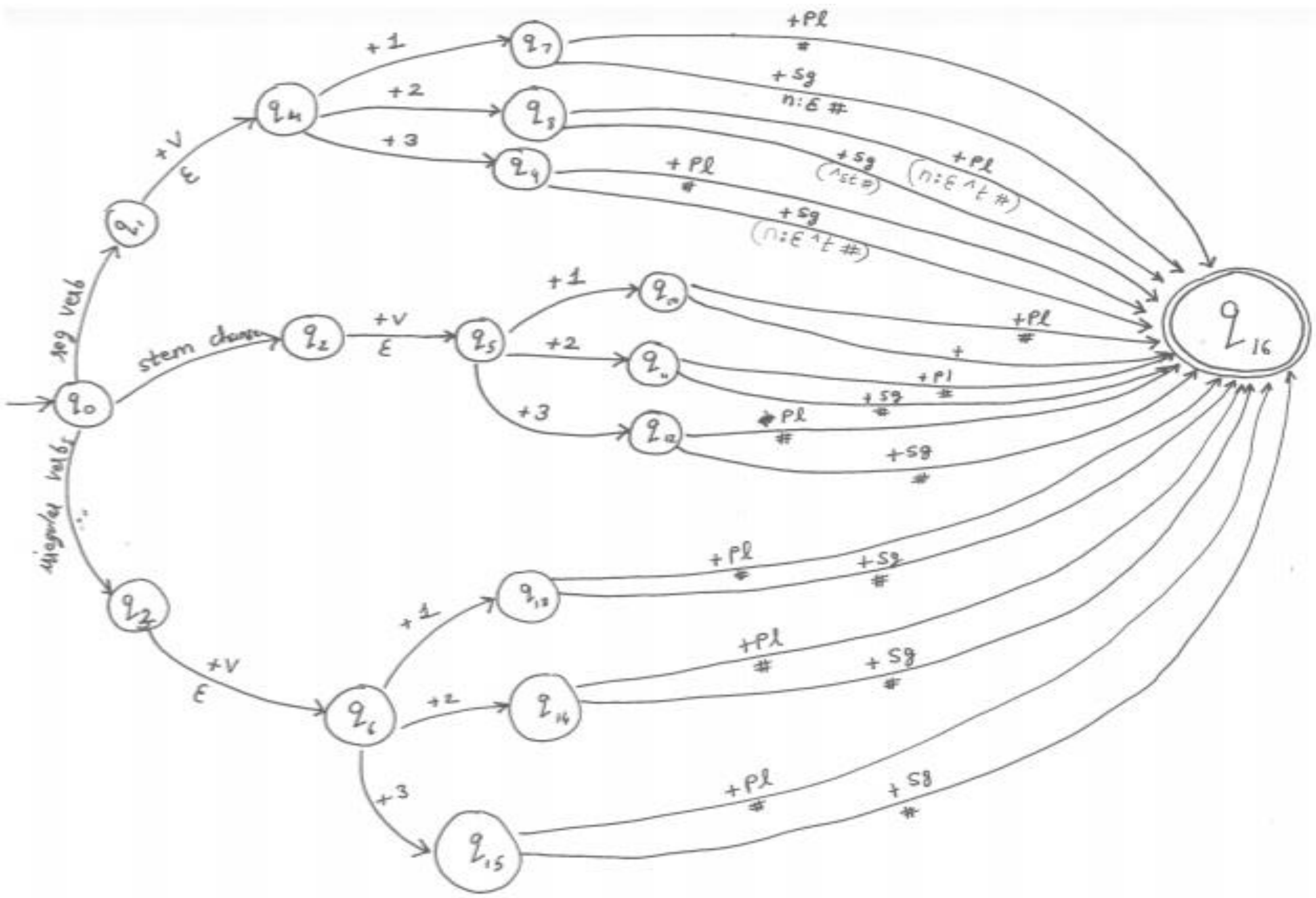
- The morphological ambiguity is seen by the way we analyze the morphemes. The following are the interpretations of ex-mother-in-law-to-be derived from the ambiguity:
  - **i.** Deceased future mother in law
  - **ii.** Future mother in law of a failed engagement (she would be his mother in law if he and his fiancé would have gotten married). Here ‘ex’ means someone from the past, like an ex-girlfriend
  - **iii.** Another possible interpretation (a really disturbing one) can be the person’s mother in law is about ‘to be’ dead. (This one is too far fetched)
- The way to disambiguate this would be to get a look at the person’s data. The context of the sentence can be understood by understanding the situation. If this statement was delivered at a wedding it would mean the first option. Otherwise, it most likely would meant the second option. The third option is too unlikely to be taken into account.

Question 2

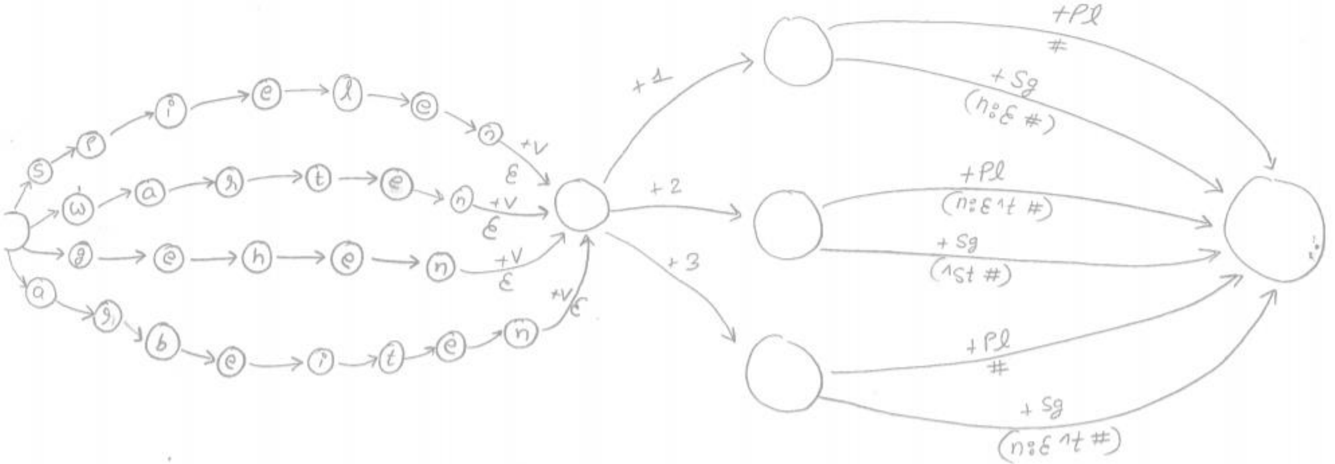
The following screenshots represent the FSTs in the following order:

1) Schematic transducer
2) Fleshed out transducer
3) Lexicon table

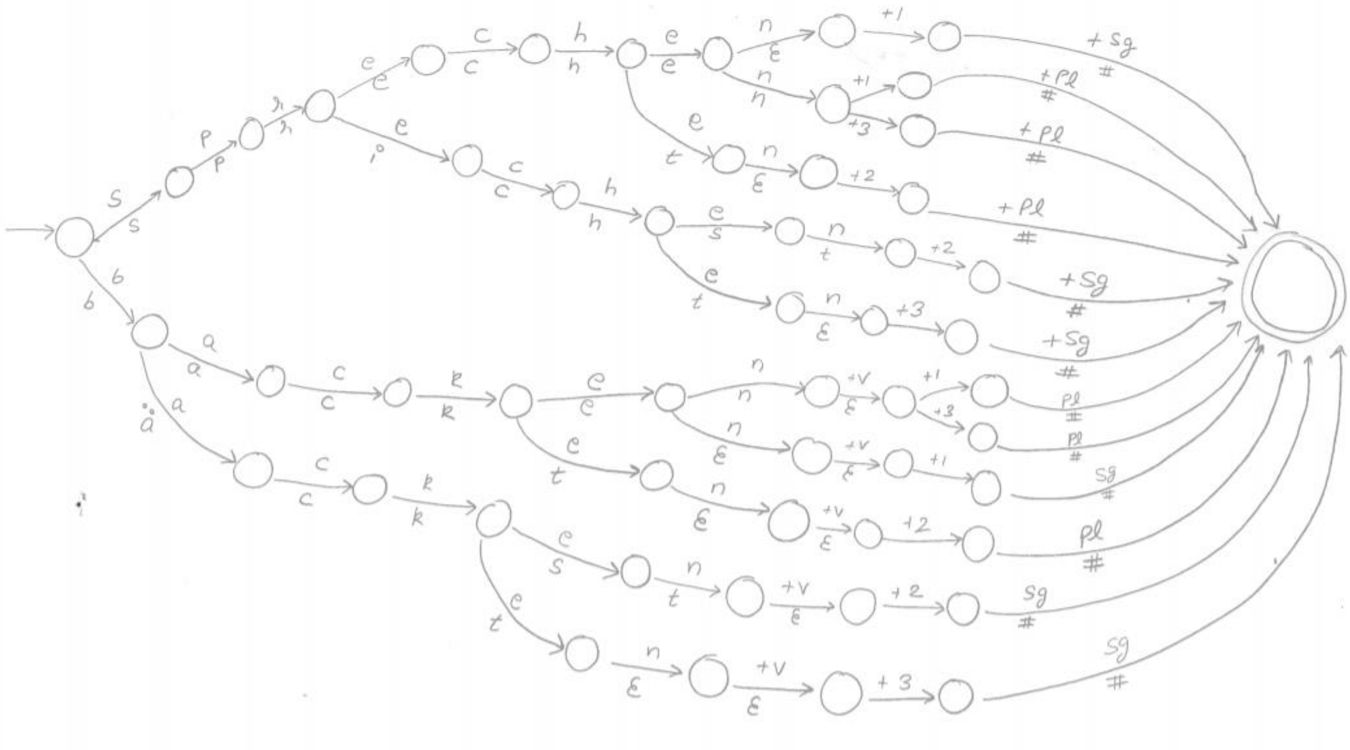
Note: The transducer was designed similar to Figure 3.13. The input is first channel to the type of word it is (regular/irregular/stem change) and then it is processed. The conventions such as ‘en:**Ɛ^#’** mean ‘en’ is replaced with ‘**Ɛ’** and followed by end of word.



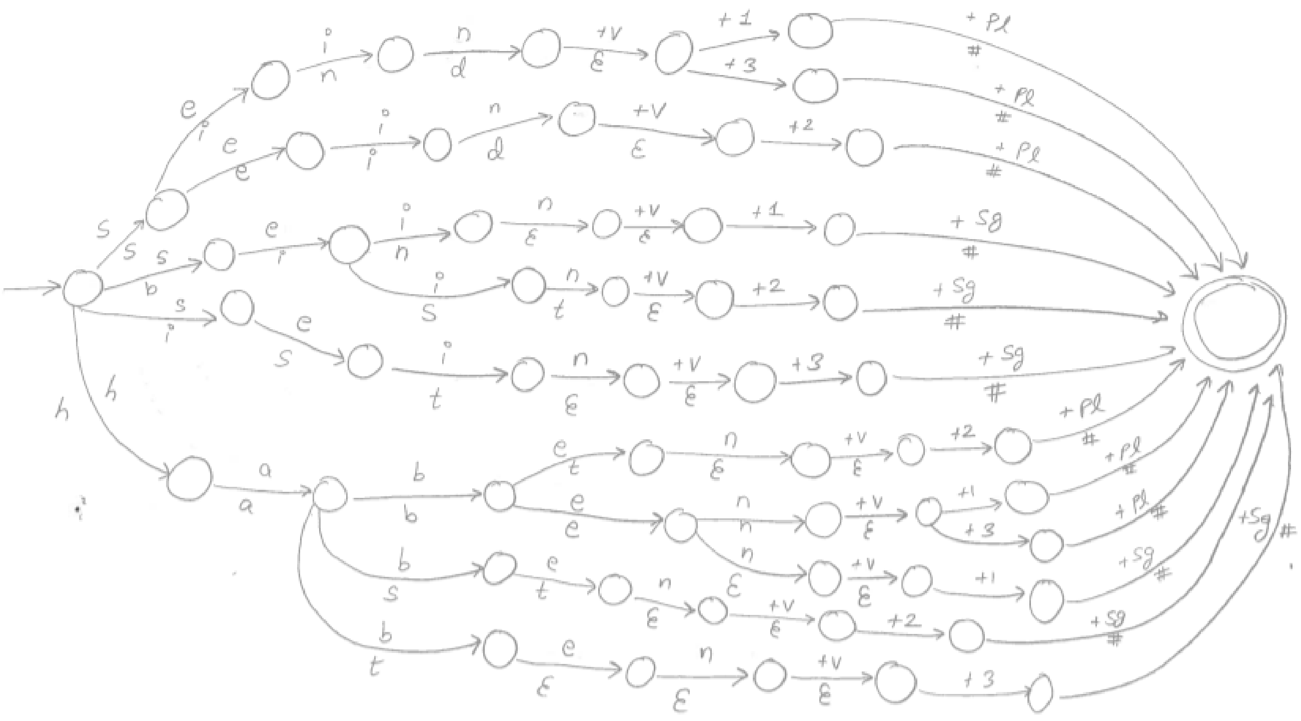
Regular Verb



Verbs with stem change



Irregular Verbs



- **Lexicon Table**

Regular Verb
spielen
wartan
gehen
arbeiten

Verbs with stem changes					
1Sg	**2Sg**	**3Sg**	**1Pl**	**2Pl**	**3Pl**
spreche n:Ɛ	spr e:i ch e:s n:t	spr e:i ch e:Ɛ n:t	sprechen	sprech e:t n:**Ɛ**	sprechen
backe n:Ɛ	b a:ä ck e:s n:t	b a:ä ck e:t n:**Ɛ**	backen	back e:t n:**Ɛ**	backen

Irregular Verbs					
1Sg	**2Sg**	**3Sg**	**1Pl**	**2Pl**	**3Pl**
s:b e:I i:n n:Ɛ	s:b e:I i:s n:t	s:i e:s i:t n:**Ɛ**	s e:i i:n n:d	sei n:d	s e:i i:n n:d
habe n:Ɛ	ha b:s e:t n:**Ɛ**	ha b:t e:Ɛ n:Ɛ	haben	hab e:t n :Ɛ	haben

Question 3

The problem defined is to perform sentiment analysis to analyse the polarity of text given as movie reviews. The algorithms/classifiers used performed supervised learning to predict the polarity of data. The movie reviews were provided as labelled text files (positive sentiments and negative sentiments)

Before the data was passed through the classifier, a the data was preprocessed to optimize it. This was done by removing stop words and removing infrequent words by adjusting the parameters of the CountVectorizer. The CountVectorizer was used to extract features from raw text files and ready the data for the classifiers. The classifiers used were: Naïve Bayes, SVM and Logistic Regression. The first iteration was proceeded with only unigrams. This was also performed during the feature extraction phase by the CountVectorizer. The algorithm was tweaked several times to achieve the best possible results. The following were the parameters chosen to achieve the current performance for the classifiers:

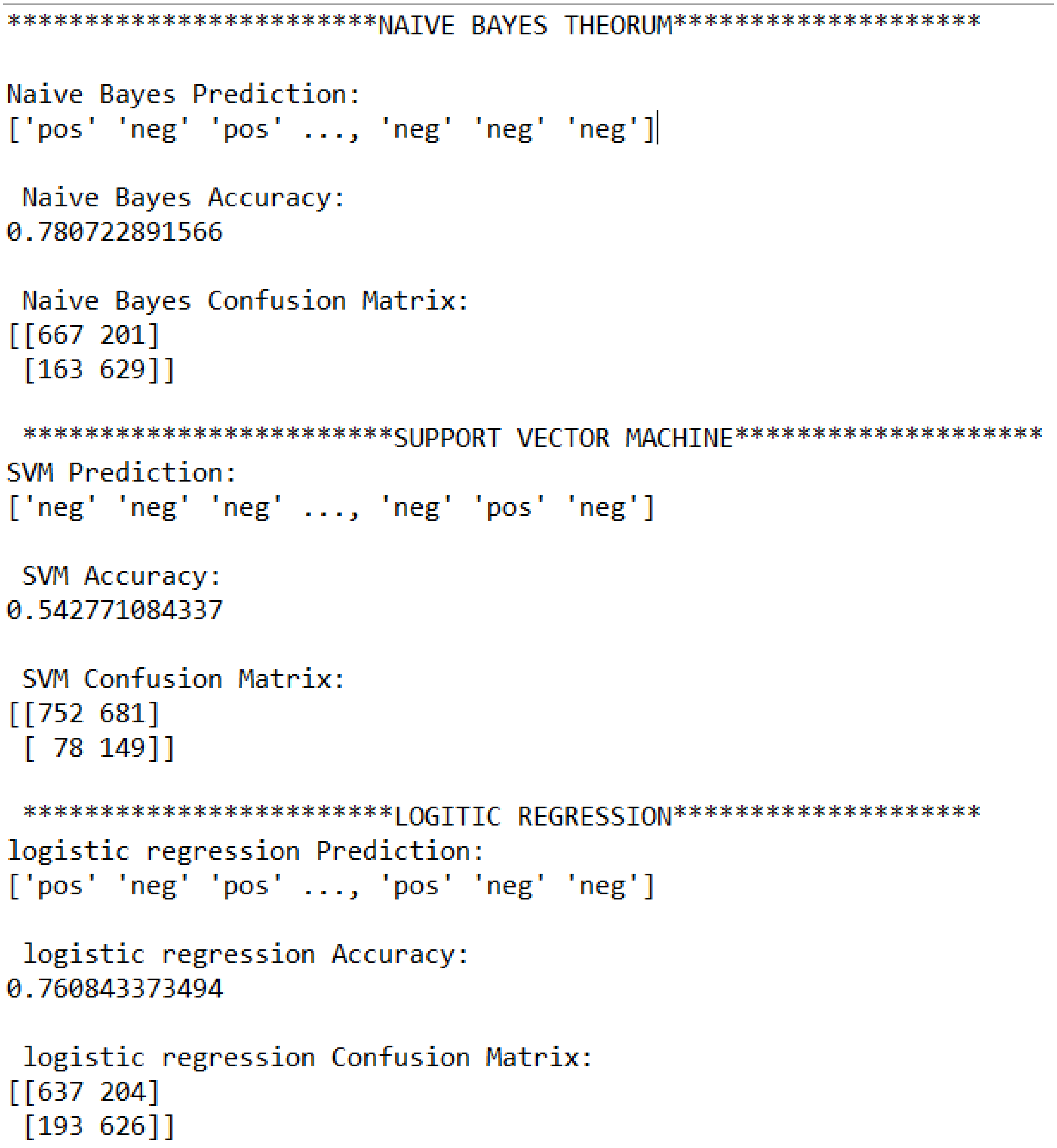
Naïve Bayes: Setting the smoothing parameter to 1 to include smoothing

SVM: Changing parameters such as setting the penalty parameter to achieve the best possible output

Logistic Regression: Changing parameters such as setting the penalty parameter, C, to! to achieve the best possible output.

The training data consisted of 4500 positive and 4500 negative movie reviews and the remaining were used as test data.

The output for the unigram analyzers are provided in the following screenshot along with the confusion matrix:

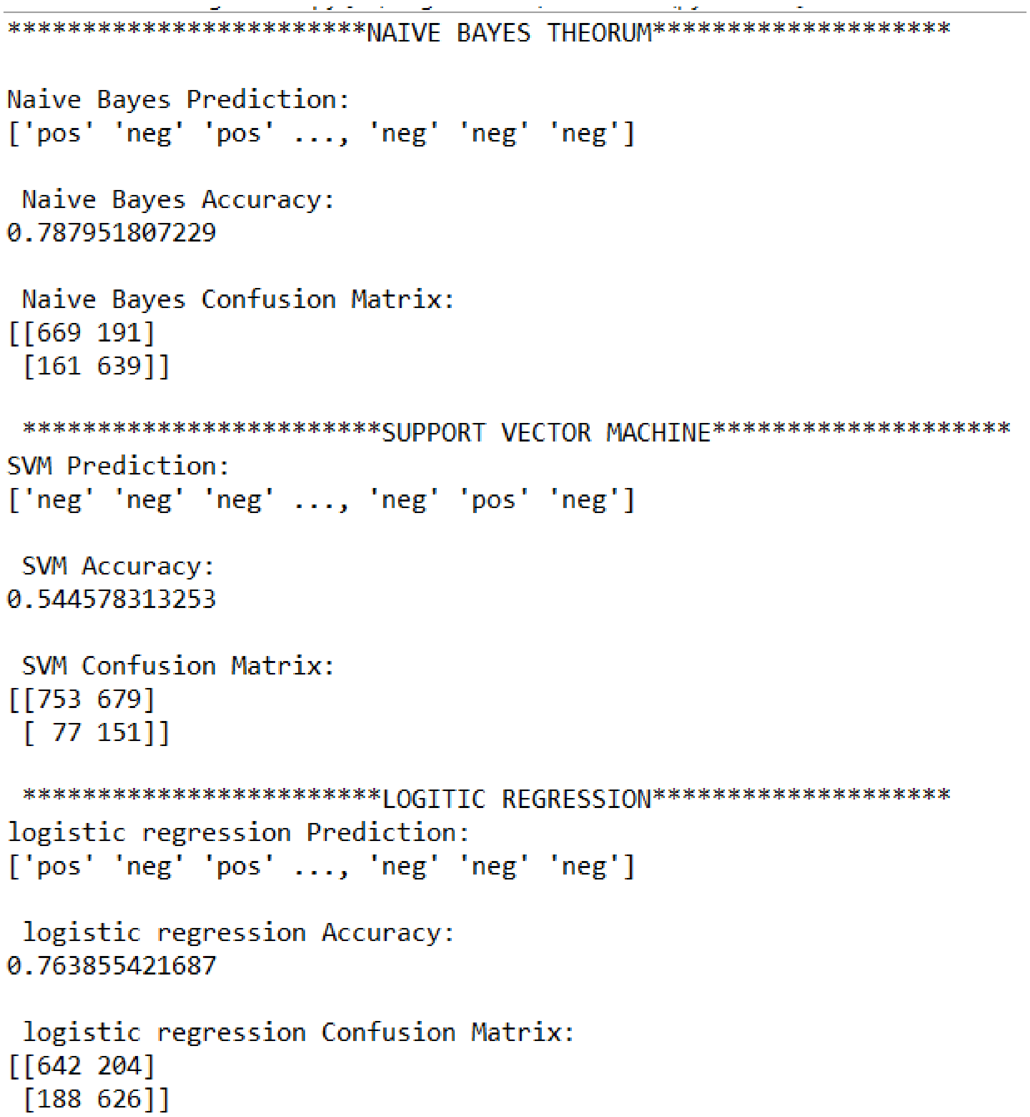


```
************************NAIVE BAYES THEORUM********************

Naive Bayes Prediction:
['pos' 'neg' 'pos' ..., 'neg' 'neg' 'neg']

 Naive Bayes Accuracy:
0.780722891566

 Naive Bayes Confusion Matrix:
[[667 201]
 [163 629]]

 ************************SUPPORT VECTOR MACHINE********************
SVM Prediction:
['neg' 'neg' 'neg' ..., 'neg' 'pos' 'neg']

 SVM Accuracy:
0.542771084337

 SVM Confusion Matrix:
[[752 681]
 [ 78 149]]

 ************************LOGITIC REGRESSION********************
logistic regression Prediction:
['pos' 'neg' 'pos' ..., 'pos' 'neg' 'neg']

 logistic regression Accuracy:
0.760843373494

 logistic regression Confusion Matrix:
[[637 204]
 [193 626]]
```

The confusion matrix has the format with model predictions in **red** and true values in **black**:

	POS	NEG
POS		
NEG		

The output for the bigram analyzers are provided in the following screenshot along with the confusion matrix (Same format):



```
************************NAIVE BAYES THEORUM********************

Naive Bayes Prediction:
['pos' 'neg' 'pos' ..., 'neg' 'neg' 'neg']

 Naive Bayes Accuracy:
0.787951807229

 Naive Bayes Confusion Matrix:
[[669 191]
 [161 639]]

 ************************SUPPORT VECTOR MACHINE********************
SVM Prediction:
['neg' 'neg' 'neg' ..., 'neg' 'pos' 'neg']

 SVM Accuracy:
0.544578313253

 SVM Confusion Matrix:
[[753 679]
 [ 77 151]]

 ************************LOGITIC REGRESSION********************
logistic regression Prediction:
['pos' 'neg' 'pos' ..., 'neg' 'neg' 'neg']

 logistic regression Accuracy:
0.763855421687

 logistic regression Confusion Matrix:
[[642 204]
 [188 626]]
```

We can observe that the bigram method yields better results as compared to the unigram method. As for performance the Naïve Bayes classifier gave the best accuracy followed closely by the logistic regression algorithm.

Below is the confusion matrix for the baseline test with randomly assigns the polarity having probability of 0.5.

	POS	NEG
POS	415	415
NEG	415	415

The accuracy for the baseline algorithm is 50%. Every algorithm evaluated outperforms the baseline algorithm by far. The Naïve Bayes algorithm proved to be 46% more accurate.