Emily Mulhall
COMP 550
Assignment 1
September 30, 2017

1 Question 1

1. Every student read a book.

   A The ambiguity is that either every student read the same book or every student read their own choice of book.

   B 'A' causes the ambiguity.

   C The ambiguity operates over the semantic domain.

   D The machine or human must know whether there is a particular book of importance which the students were all reading. If not, it likely means that the choice of book didn't matter and that not all children are reading the same book.

2. The lion is a majestic animal.

   A The ambiguity is that either there is a particular lion that is the topic of discussion or that lion is majestic, and that lions as a species are majestic.

   B The article 'the' causes the ambiguity.

   C The ambiguity is over the semantic domain.

   D The human or machine must know contextually whether a particular lion is being discussed. Otherwise, the statement is likely about the species as a whole.

3. Use of this sidewalk is prohibited by police officers.

   A The ambiguity is that either police officers cannot use this sidewalk or the police officers prohibit everyone from using the sidewalk.

   B 'By' causes the ambiguity

   C This ambiguity is over the syntactic domain.

   D The human or machine must know whether the police officers are doing the action of prohibiting or not

4. My english teacher recently recovered from a bowel cancer operation...and he tried to show me a semi colon.

   A The ambiguity is that either the teacher showed the student ';' or that he showed the student his colon (the body part).

B The ambiguity comes from 'semi colon.'

C This ambiguity is over the lexical domain.

D The machine or human could use the reaction following this statement or previous context of the relationship of the student and teacher in order to disambiguate the sentence and know whether a body part was shown or a punctuation mark.

5. She is my ex-mother-in-law-to-be

A The ambiguity is that either the speaker is in the process of getting a divorce or that he or she broke up with his or her fianc before they were married.

B The ambiguity comes from the prefix 'ex' and the suffix 'to-be.'

C This ambiguity is over the orthographic domain.

D To disambiguate the sentence the human or machine should use context to discover whether the speaker had gotten married yet or not.

2 Question 2

Regular verbs	**Verbs with stem change**	**Irregular**
Spiel-	spri:ech-	sein 1sg, 2sg= b:s i:e ϵ:i ϵ:n
Wart-	bä:ack-	sein 3sg= i:s ϵ:e ϵ:i ϵ:n
Geh-		sein 1pl, 3pl= s:s i:e ϵ:i ϵ:n
Arbeit-		sein 2pl = s:s e:e i:i ϵ:n
		haben 1sg, 1pl, 2pl, 3pl= hab-
		haben 2sg, 3sg = h:h a:a ϵ:b-

Please see the last two pages for the rest of question 2

3 Question 3

In order to approach the problem of building a classifier, I followed a few simple steps. First, I had to preprocess the data. This included opening both the positive and negative review text files and tokenizing them into sentences. These sentences were added to the 'text' section of the data frame, and they were labelled either 'pos' or 'neg' in the 'class' section of the data frame.

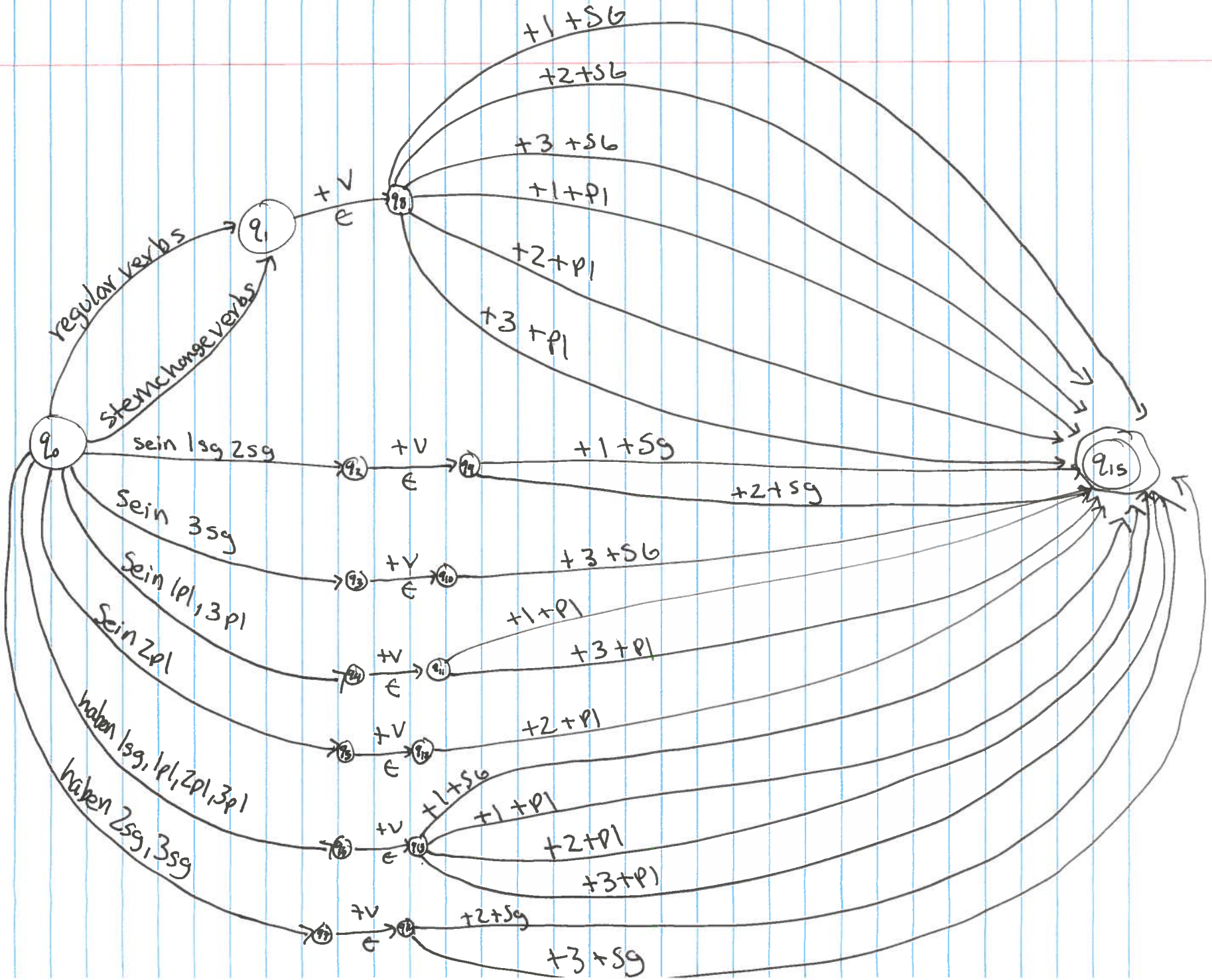
Once the preprocessing was complete, the training and testing began. I used a pipeline from sklearn to do this. The first step was to transform the data with a count vectorizer, which extracted the features from the data. Then, a classifier was trained with these features and the training data. To divide the data frame into training and testing data, I used a k fold with 6 folds. Thus, 5/6 of the data in the data frame was used for training, and 1/6 for testing. The text was extracted from the 'text' field of the data frame as was the labels in 'class.' The pipeline was then fitted to the training text and corresponding labels. Once

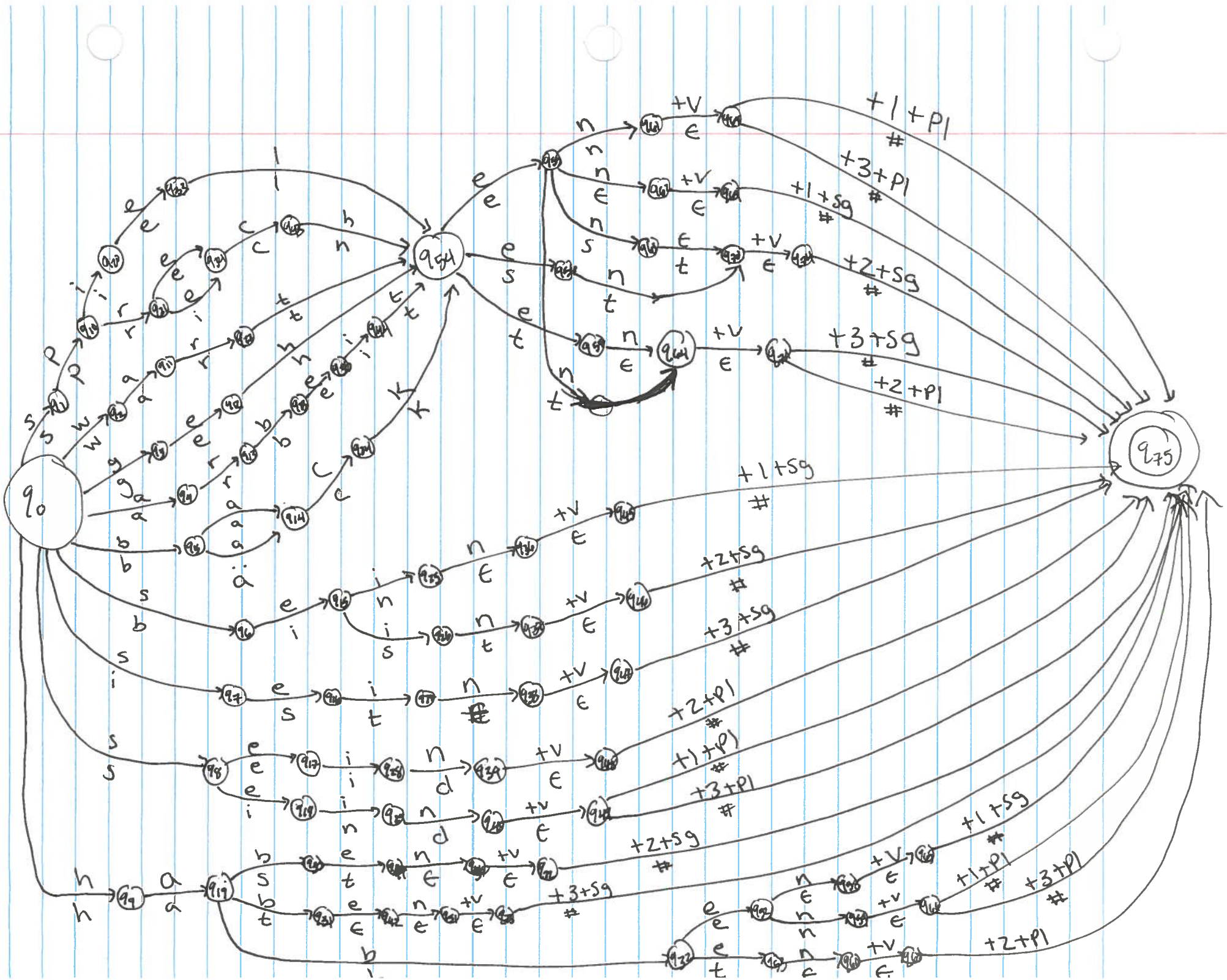
the classifier had been trained, predictions were made on the test data. The accuracy of these predictions were then scored, and the results printed.

Overall I tried quite a decent range of parameters. I altered the n gram range between (1,1) and (1,2), I experimented with adding stop words, and I tried a few different max df and min df values. I tried all possible combinations of these parameters in order to get the best possible results.

Overall, the multinomial naive Bayes algorithm seemed to perform with the greatest accuracy as compared to the logistic regression and the support vector classifiers. However, it still performed below chance, which is surprising. One of the reasons the accuracy could be so low is due to an encoding issue I had. When attempting to read the documents, I consistently faced a unicode error. I tried decoding and encoding, and quite a few other tactics, but I ended up using repr() which provides an alternate representation of the strings. This ended up adding quite a lot of backslashes to the text, which could have affected how it was tokenized, and it could have come up as a very frequent feature in both positive and negative data, leading to poor accuracy. It's also possible that some punctuation was misinterpreted and the sentences were not properly tokenized which again could lead to poorer accuracy.

Overall, the most successful model had the n gram range (1,2), no stop words, a max df of 0.9 and a min df of 0.01. It had a multinomial naive Bayes classifier, and an accuracy of about 36 percent. It's confusion matrix was: [[2606 4662] [4457 2635]]. Looking at the confusion matrix, it seems that the classifier had it almost entirely backwards, expecting way too few positives and too many negatives.


regular verbs
stemchangeverbs
+V
ε
+1 +Sg
+2+Sg
+3 +Sg
+1+Pl
+2+Pl
+3 +Pl
sein 1sg 2sg
Sein 3sg
Sein 1pl, 3pl
Sein 2pl
haben 1sg, 1pl, 2pl, 3pl
haben 2sg, 3sg


q0
q75
+1+Pl
+3+Pl
+1+Sg
+2+Sg
+3+Sg
+2+Pl
+V
#