CMPE 561 – Project 2 Report
Ayşe Müge Kurtipek
2011400159
Repo: https://github.com/mugekurtipek/Cmpe561

HMM Part-of-Speech Tagger:

To train my POS Tagger, there were two things that I needed. First one is the observation likelihood values for the words for each tag and the second one is the transition matrix of the tags. After obtaining these two matrices, the only thing left is to implement the Viterbi algorithm.

In order to obtain the observation likelihoods of the words, I scanned the training data. I created a Hashmap, which keeps the tokens as keys and an array of number of tag size as a value for each key. In the array I recorded number of occurrences of the token for each tag. While scanning the training data I was incrementing the related index of array to record the occurrence of the token as that tag. At the end of the scanning, I was able to say for which token to be seen with which tag. After that, I switched the number of occurrences with the probabilities.

In order to obtain the transition matrix of the tags, I scanned the training data again. I created a 2D array, which has dimensions of [number of tags + 1] x [number of tags]. (+1 is for start tag) The value in the intersection of row i and column j means: the number of occurrence of index of jth tag after index of ith tag. After calculating all the training data I switched the number of occurrences with the probabilities.

After obtaining these two data structures, I implemented the Viterbi algorithm. For each word in the test data, I labeled the words with the tags, which has the highest probability values came from the multiplication of the observation likelihood of the word for each tag and transition probability from the previous tag.

List of Tags:

- Cpostag
    1. Adv
    2. Punc
    3. Noun
    4. Adj
    5. Verb
    6. Det
    7. Pron
    8. Postp
    9. Conj
    10. Ques
    11. Num
    12. Interj
    13. Dup
    14. Zero

- Postag
    1. Adv
    2. Punc
    3. Noun
    4. Adj
    5. Zero
    6. Verb
    7. Prop
    8. Det
    9. PersP
    10. Postp
    11. Conj
    12. NPastPart
    13. APastPart
    14. NInf
    15. Ques
    16. DemonsP
    17. Card
    18. APresPart
    19. QuesP
    20. AFutPart
    21. Pron
    22. ReflexP
    23. NFutPart
    24. Interj
    25. Dup
    26. Ord
    27. Distrib
    28. Real
    29. Num
    30. Range

Steps for Task 1:

In this step I created observation likelihood and transition matrix data structures. Then I stored these two data structures in two different txt files in order to use them in Step 2.

In order to obtain the observation likelihoods of the words, I scanned the training data. I created a Hashmap, which keeps the tokens as keys and an array of number of tag size as a value for each key. In the array I recorded number of occurrences of the token for each tag. While scanning the training data I was incrementing the related index of array to record the occurrence of the token as that tag. At the end of the scanning, I was able to say for which token to be seen with which tag. After that I switched the number of occurrences with the probabilities.

In order to obtain the transition matrix of the tags, I scanned the training data again. I created a 2D array, which has dimensions of [number of tags + 1] x [number of tags]. (+1 is for start tag) The value in the intersection of row i and column j means: the number of occurrence of index of j'th tag after index of i'th tag. After calculating all the training data I switched the number of occurrences with the probabilities.

Steps for Task 2:

In order to test a given file I started my tagger with loading the observation likelihoods and transition matrix data structures from the txt files that I recorded in Step 1.

After obtaining these two data structures, I implemented the Viterbi algorithm. For each word in the test data, I labeled the words with the tags, which has the highest probability values came from the multiplication of the observation likelihood of the word for each tag and transition probability from the previous tag.

Accuracies (over all words):

- **Cpostag**
  Accuracy = 78%
  Confusion Matrix:
  (The intersection of Adv and Noun [C(Adv, Noun)] indicates the number of times (in percentage) an item with actual tag Adv has been assign the tag Noun by the model with respect to the total number of errors.)

	Adv	Punc	Noun	Adj	Verb	Det	Pron	Postp	Conj	Ques	Num	Interj	Dup	Zero
Adv	-	0.00	3.17	1.65	8.38	0.63	0.51	0.63	0.00	0.00	0.00	0.00	0.00	0.00
Punc	0.00	-	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Noun	0.00	0.00	-	5.08	16.88	0.63	0.00	0.25	0.25	0.00	0.38	0.00	0.00	0.00
Adj	0.25	0.00	18.27	-	11.80	0.38	0.25	0.25	0.13	0.00	0.00	0.00	0.00	0.00
Verb	1.02	0.00	13.96	4.31	-	0.13	0.00	0.25	0.89	0.00	0.00	0.00	0.00	0.00
Det	0.25	0.00	0.25	0.51	0.00	-	1.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pron	0.00	0.00	0.00	0.00	0.00	1.65	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Postp	0.38	0.00	0.76	0.63	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00
Conj	0.00	0.00	0.00	0.00	0.76	0.00	0.38	0.00	-	0.00	0.00	0.00	0.00	0.00
Ques	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00
Num	0.00	0.00	1.27	0.00	0.00	0.63	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00
Interj	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00
Dup	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
Zero	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-

- **Postag**
  Accuracy = 75%
  Confusion Matrix:
  (The intersection of Adv and Noun [C(Adv, Noun)] indicates the number of times (in percentage) an item with actual tag Adv has been assign the tag Noun by the model with respect to the total number of errors.)

	Adv	Punc	Noun	Adj	Zero	Verb	Prop	Det	PersP	Postp	Conj	NPastPart	APastPart	Ninf	Ques	DemonsP	Card	APresPart	QuesP	AFutPart	Pron	ReflexP	NFutPart	Interj	Dup	Ord	Distrib	Real	Num	Range
Adv	-	0.00	1.84	1.49	0.11	8.05	0.00	0.57	0.00	0.57	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.11	0.46	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Punc	0.00	-	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Noun	0.00	0.00	-	3.22	0.23	1.61	0.00	0.00	0.00	0.23	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Adj	0.23	0.00	12.87	-	0.57	0.34	0.00	0.34	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero	0.23	0.00	3.56	3.10	-	2.64	0.00	0.69	0.00	0.23	0.11	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Verb	0.69	0.00	4.60	1.03	0.11	-	0.00	0.00	0.00	0.00	0.69	0.00	0.00	0.46	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Prop	0.00	0.00	3.68	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Det	0.23	0.00	0.00	0.46	0.00	0.00	0.00	-	1.72	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PersP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Postp	0.46	0.00	0.57	0.46	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Conj	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NPastPart	0.00	0.00	0.57	0.11	0.00	2.41	0.00	0.00	0.00	0.00	0.11	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APastPart	0.00	0.00	0.00	0.00	0.00	2.87	0.00	0.00	0.00	0.00	0.11	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ninf	0.00	0.00	5.17	0.46	0.00	11.84	0.00	0.00	0.00	0.00	0.11	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ques	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DemonsP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.26	0.46	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Card	0.00	0.00	1.03	0.11	0.11	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APresPart	0.00	0.00	0.57	0.11	0.00	7.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QuesP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFutPart	0.00	0.00	0.11	0.00	0.00	1.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pron	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ReflexP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NFutPart	0.00	0.00	0.34	0.00	0.00	1.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Interj	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00
Dup	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
Ord	0.00	0.00	0.11	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00
Distrib	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00
Real	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00
Num	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
Range	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-

Accuracy for Known and Unknown Words:

- **Cpostag**
  Known words accuracy = 78,4%
  Unknown words accuracy = 73,6%
- **Postag**
  Known words accuracy = 78,9%
  Unknown words accuracy = 54,8%

Unknown Words:

For the unknown words I assigned the most common tag in the training data set, which was Noun tag. During the training task I also calculated the most common tag from the given training data.

Performance of my tagger:

Overall performance of my tagger is quite well. It is above 70% for both tag sets. With improving the training data size, it can be improved.

However for the unknown words especially with for Postag, the accuracy is relatively low. The main reason is the simplicity of the handling of the unknown words. With an improved technique for tagging the unknown words, this accuracy also can be improved.