2.1 ⟨s⟩ →

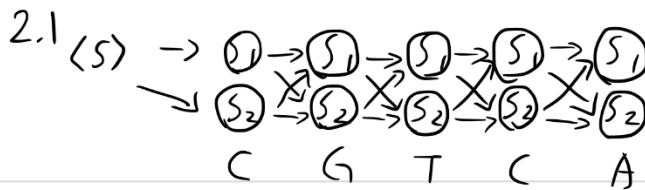C G T C A

Jihyung Kim

$V_1(S_1) = P(S_1|⟨S⟩) \, P(C|S_1) = 0.5 \times 0.2 = 0.1$

$V_1(S_2) = P(S_1|⟨S⟩) \, P(C|S_2) = 0.5 \times 0.4 = 0.2$

$V_2(S_1|S_1) = V_1(S_1) \, P(S_1|S_1) \, P(G|S_1) = 0.1 \times 0.8 \times 0.3 = 0.024$

$V_2(S_1|S_2) = V_1(S_2) \, P(S_1|S_2) \, P(G|S_1) = 0.2 \times 0.3 \times 0.3 = 0.018$

$V_2(S_2|S_1) = V_1(S_1) \, P(S_2|S_1) \, P(G|S_2) = 0.1 \times 0.2 \times 0.1 = 0.002$

$V_2(S_2|S_2) = V_1(S_2) \, P(S_2|S_2) \, P(G|S_2) = 0.2 \times 0.7 \times 0.1 = 0.014$

$V_3(S_1|S_1) = V_2(S_1) \, P(S_1|S_1) \, P(T|S_1) = 0.024 \times 0.8 \times 0.2 = 0.00384$

$V_3(S_1|S_2) = V_2(S_2) \, P(S_1|S_2) \, P(T|S_1) = 0.014 \times 0.3 \times 0.2 = 0.00084$

$V_3(S_2|S_1) = V_2(S_1) \, P(S_2|S_1) \, P(T|S_2) = 0.024 \times 0.2 \times 0.4 = 0.00192$

$V_3(S_2|S_2) = V_2(S_2) \, P(S_2|S_2) \, P(T|S_2) = 0.014 \times 0.7 \times 0.4 = 0.00392$

$V_4(S_1|S_1) = V_3(S_1) \, P(S_1|S_1) \, P(C|S_1) = 0.00384 \times 0.8 \times 0.2 = 0.0006144$

$V_4(S_1|S_2) = V_3(S_2) \, P(S_1|S_2) \, P(C|S_1) = 0.00392 \times 0.3 \times 0.2 = 0.0002352$

$V_4(S_2|S_1) = V_3(S_1) \, P(S_2|S_1) \, P(C|S_2) = 0.00384 \times 0.2 \times 0.4 = 0.0003072$

$V_4(S_2|S_2) = V_3(S_2) \, P(S_2|S_2) \, P(C|S_2) = 0.00392 \times 0.7 \times 0.4 = 0.0010976$

$V_5(S_1|S_1) = V_4(S_1) \, P(S_1|S_1) \, P(A|S_1) = 0.0006144 \times 0.8 \times 0.3 = 0.000147456$

$V_5(S_1|S_2) = V_4(S_2) \, P(S_1|S_2) \, P(A|S_1) = 0.0010976 \times 0.3 \times 0.3 = 0.000098784$

$V_5(S_2|S_1) = V_4(S_1) \, P(S_2|S_1) \, P(A|S_2) = 0.0006144 \times 0.2 \times 0.1 = 0.000012288$

$V_5(S_2|S_2) = V_4(S_2) \, P(S_2|S_2) \, P(A|S_2) = 0.0010976 \times 0.7 \times 0.1 = 0.000076832$

	T=1	T=2	T=3	T=4	T=5	
S_1	0.1	0.024	0.00384	0.0006144	0.000147456	V-matrix
S_2	0.2	0.014	0.00392	0.0010976	0.000076832	

	T=1	T=2	T=3	T=4	T=5	
S_1	⟨s⟩	S_1	S_1	S_1	S_1	Backtracking matrix
S_2	⟨s⟩	S_2	S_2	S_2	S_2	Decoded sequence

⟨s⟩ → S_1 → S_1 → S_1 → S_1 → S_1

2.2

```
(base) C:\Users\Jihyung\Downloads\HW2>python preprocessing.py

Preprocessing autos training set...
Completed processing of autos training set. Time elapsed: 18.37s

Preprocessing hardware training set...
Completed processing of hardware training set. Time elapsed: 14.04s

Preprocessing autos test set...
Completed processing of autos test set. Time elapsed: 11.59s

Preprocessing hardware test set...
Completed processing of hardware test set. Time elapsed: 10.78s
X_train shape:  (1172, 15579)
y_train shape:  (1172,)
X_test shape:  (781, 15579)
y_test shape:  (781,)
Feature set size:  15579
```

Training dataset is comprised of 1172 examples for both autos and hardware set. Also, it is comprised of unique 15579 lowercase words. Test set has the size of 781 grounded on the trained model.

```
(base) C:\Users\Jihyung\Downloads\HW2>python text_classifier.py
Loading dataset...

Preprocessing autos training set...
Completed processing of autos training set. Time elapsed: 18.64s

Preprocessing hardware training set...
Completed processing of hardware training set. Time elapsed: 14.09s

Preprocessing autos test set...
Completed processing of autos test set. Time elapsed: 10.90s

Preprocessing hardware test set...
Completed processing of hardware test set. Time elapsed: 9.56s

Dataset loaded. Total time elapsed: 55.00s

Number of training examples:  1172
Number of test examples:  781
Number of unique words in feature map:  15579

Fitting Multinomial Naive Bayes model...

Training time elapsed: 0.11s

F1 score for positive class (hardware):  0.9856209150326798
F1 score for negative class (autos):  0.9861982434127979
Mean Accuracy: 98.59%

Total time elapsed: 55.19s
```

After training the dataset, naive bayes model's F1 score is about 0.98562 on mac.hardware which is positive class. For negative class (autos), F1 score is about 0.98619.