Name: Songqi Chen

	S_1	S_2
S_1	0.8	0.2
S_2	0.3	0.7

	A	C	G	T
S_1	0.3	0.2	0.3	0.2
S_2	0.1	0.4	0.1	0.4

V- Matrix

	C	G	T	C	A
S_1	0.1	0.024	0.00384	0.000614	0.000147
S_2	0.2	0.014	0.00392	0.001098	0.00077

$$V_1(S_1) = P(S_1 | *) P(C | S_1)$$

$$= 0.5 \cdot 0.2 = 0.1$$

$$V_1(S_2) = P(S_2 | *) P(C | S_2)$$

$$= 0.5 \cdot 0.4 = 0.2$$

$$V_2(S_1) = \max \begin{cases} V_1(S_1) \cdot P(S_1 | S_1) \cdot P(G | S_1) = 0.024 \\ V_1(S_2) \cdot P(S_1 | S_2) \cdot P(G | S_1) = 0.018 \end{cases}$$

$$V_2(S_2) = \max \begin{cases} V_1(S_1) \cdot P(S_2 | S_1) \cdot P(G | S_2) = 0.002 \\ V_1(S_2) \cdot P(S_2 | S_2) \cdot P(G | S_2) = 0.014 \end{cases}$$

$$V_3(S_1) = \max \begin{cases} V_2(S1) \cdot P(S_1|S_1) \cdot P(T|S_1) = 0.00384 \\ V_2(S_2) \cdot P(S_1|S_2) \cdot P(T|S1) = 0.00084 \end{cases}$$

$$V_3(S_2) = \max \begin{cases} V_2(S_1) \cdot P(S_2|S_1) \cdot P(T|S_2) = 0.00192 \\ V_2(S_2) \cdot P(S_2|S_2) \cdot P(T|S_2) = 0.00392 \end{cases}$$

$$V_4(S_1) = \max \begin{cases} V_3(S_1) \cdot P(S_1|S_1) \cdot P(C|S_1) = 0.000614 \\ V_3(S_2) \cdot P(S_1|S_2) \cdot P(C|S_1) = 0.000235 \end{cases}$$

$$V_4(S_2) = \max \begin{cases} V_3(S_1) \cdot P(S_2|S_1) \cdot P(C|S_2) = 0.000307 \\ V_3(S_2) \cdot P(S_2|S_2) \cdot P(C|S_2) = 0.001098 \end{cases}$$

$$V_5(S_1) = \max \begin{cases} V_4(S_1) \cdot P(S_1|S_1) \cdot P(A|S_1) = 0.000147 \\ V_4(S_2) \cdot P(S_1|S_2) \cdot P(A|S_1) = 0.000099 \end{cases}$$

$$V_5(S_2) = \max \begin{cases} V_4(S_1) \cdot P(S_2|S_1) \cdot P(A|S_2) = 0.000012 \\ V_4(S_2) \cdot P(S_2|S_2) \cdot P(A|S_2) = 0.000077 \end{cases}$$

Based on the V matrix above,

$$\begin{array}{c|c|c|c|c|c|} S_1 & * & \boxed{S_1} & \boxed{S_1} & \boxed{S_1} & \boxed{S_1} \\ S_2 & * & S_2 & S_2 & S_2 & S_2 \end{array}$$

The backtracking will be

$*, S_1, S_1, S_1, S_1, S_1$

Question 2.2

Number of Files: Positive: 594
Negative: 578

Total vocabulary: 15357

F1-score: 0.99483117785447164

The first step of this homework is to use spaCy to extract words from files. The code for this step can be find in extractFile.py. The stopword, punctuations are removed during this process. Then using the sentiment_reader.py to generate train and test dataset from the result of previous step. Then using multinomial_naive_bayes.py to make prediction. The runtime for file extraction is 68s and the runtime for multinomial naive bayes is 0.21s.