Yuwen Sang
COMP 3705: Text Mining - Assignment 3

Harry Potter Character Analysis Report

Introduction

This is an analysis report for Harry Potter I (Harry Potter and the Philosopher's Stone) characters. This report focuses on comparing two characters: Albus Dumbledore (Dumbledore) and Severus Snape (Snape). In Harry Potter I, Dumbledore is a typical positive fictional character, while Snape is the boss in the Harry Potter I. Therefore, we choose these two characters and try to use different ways to predict the sentences quoted from the book is about who.

In this text mining project uses the quoted text that mentioned the characters, or spoke by the characters to test and training. The project first modified the provided code, write each character's related text into a file. In this process, the character's name has been removed from the text. Then, the project uses *txt_processor.py* randomly choose 20 quotes as the test quotes, and the last of the quotes were used as training text.

Analysis

Part I

In this part, by modifying the project provided code, we can make a very general analysis by counting the positive and negative words from the sentences that include the character's name. According to the analysis, there are about 20.3% sentences about Dumbledore are positive, 16.54% are negative, and 63.16 are neutral. There are 10.64% sentences about Snape have more positive words than the negative words, and 26.24% have more negative words.

	Positive Sentences	Negative Sentences	Netural Sentences
Dumbledore	20.30%	16.54%	63.16%
Snape	10.64%	26.24%	63.12%

Figure 1

Part II

a. Naive Bayes Method

Referring *Naïve Bayes.py*, the code uses the processed files to predict which character the tested quotes belong to and evaluates the model. Here we use the bag of words with the Multinomial Naïve Bayes Method. Figure 2 shows the result of the NB training model to test randomly 20 quotes and predict whether they belong to the specific character. The column *whether the character* equals 1 if the model thinks the *content* belongs to the character, and 0 if it does not think so. There are 55% of 20 quotes from Dumbledore are predicted correctly, and there are 75% of 20 quotes from Snape are predicted correctly.

NB Result Test of Dumbledore			
content	prob_pos	prob_neg	whether the character
forget dog forget guardin professor nicolas flamel a	2.06E-18	9.83E-18	1
said first one ever gasped	4.53E-14	6.65E-14	1
sir quick calm dear boy little behind times said	1.16E-21	2.33E-20	1
full minute three stood looked little bundle hagrid sh	3.35E-28	1.73E-28	0
took several purple firecrackers exploding end profe	8.98E-09	8.91E-09	0
professor informed	0.00327447	0.008494623	1
wondered trusted enough help apart	4.49E-10	1.98E-10	0
plain whatever everyone saying going believe told t	4.92E-15	2.77E-15	0
man name albus	2.24E-10	6.34E-09	1
said thinks	0.006276067	0.011326163	1
great man	1.43E-07	6.70E-07	1
harry relax hermione right stone safe long around	7.91E-19	5.81E-20	0
threw sharp sideways glance though hoping going t	6.76E-18	1.97E-18	0
idea told might trouble gettin hold yeh much yeh kr	1.33E-23	3.44E-24	0
even blimey come watch	3.37E-09	3.96E-10	0
albus gotten feet	1.50E-10	1.58E-09	1
c c stand lily james dead poor little harry ter live mu	2.43E-44	3.41E-43	1
good afternoon harry said	1.10E-07	7.92E-08	0
said slowly shows us want whatever want yes said q	1.28E-18	7.44E-18	1
see professor	1.03E-05	4.02E-05	1
		Correct Percent	55.00%

NB Result test of Snape			
content	prob_pos	prob_neg	whether the character
wherever filch must know shortcut soft greasy	4.06E-17	9.83E-19	1
harry see could yet sometimes horrible feeling	6.60E-17	9.63E-16	0
cheer said ron always taking points fred geor	6.79E-14	1.66E-14	1
harry know whether imagining seemed keep	3.46E-16	1.97E-17	1
start term banquet harry gotten idea profess	9.02E-11	3.51E-11	1
looked books dursleys expect remember eve	9.84E-16	2.49E-15	0
times even wondered whether following tryin	2.24E-09	3.96E-10	1
quirrell said professor harry	3.46E-09	2.81E-09	1
friend miss granger accidentally knocked rush	5.87E-12	1.17E-13	1
always seemed hate much	2.02E-09	5.94E-10	1
idea voldemort certainly scared keep visiting	6.13E-32	1.42E-34	1
take hospital wing spat seamus	5.71E-07	1.67E-07	1
father something could never forgive	8.77E-12	1.31E-11	0
made rule harry muttered angrily limped awa	5.90E-15	4.16E-15	1
looked books dursleys expect remember eve	9.84E-16	2.49E-15	0
moment later professor mcgonagall come bu	9.71E-20	9.30E-20	1
harry know whether imagining seemed keep	3.46E-16	1.97E-17	1
quietly possible crept along next corridor fad	1.48E-15	1.39E-16	1
peering around however saw percy	2.24E-09	3.17E-09	0
trying steal	3.43E-06	3.35E-07	1
		Correct Percent	75.00%

Figure 2

According to the evaluation (figure 3), the AKA predictive positive precision is higher than the negative precision. However, the true pos rate is lower than the neg one.

NB Model Evaluation	Percent
AKA predictive pos (Dumbledore) precision	75.00%
AKA true pos (Dumbledore) rate	63.83%
AKA predictive neg (Snape) precision	69.09%
AKA true neg (Snape) rate	79.17%

Figure 3

b. Logistic Regression Method

In this section, we use word bag with count vectorizer training logistic regression (LR) model (referring *Logistic Regression.py*). According to use several different weight to test, the weight of [3, 1] for Dumbledore, [1, 2] for Snape as [pos, neg] training weight has the best performance of the test result.

Figure 4 shows the prediction of the quote samples. We can see that the prediction of Dumbledore's quotes has the accuracy of 85% in the 20 samples, and the accuracy of predicting Snape reaches 100%. According to the evaluation result, LR model's precision is 75%.

LR Test Result for Dumbledore		
content	p1	whether the character
You forget that dog, an' you forget	0.047425873	0
" again," he said, "He was the first o	0.99966465	1
Sir, quick --" "Calm yourself, dear b	0.999993856	1
For a full minute the three of them	0.993307149	1
It took several purple firecrackers e	0.731058579	1
Professor will be informed of this.	0.5	0
"We wondered who had trusted e	0.731058579	1
It was plain that whatever "everyon	0.99966465	1
This man's name was Albus .	0.993307149	1
"That's what I said, but thinks that	0.880797078	1
Great man, ."	0.952574127	1
"Harry, relax, Hermione's right, the	0.993307149	1
She threw a sharp, sideways glance	0.999983299	1
"I had no idea, when told me there	0.999876605	1
"Even -- blimey -- 's come to wat	0.731058579	1
Albus had gotten to his feet.	0.880797078	1
"But I c-c-can't stand it -- Lily an' J	0.999954602	1
"Good afternoon, Harry," said .	0.731058579	1
Then he said slowly, "It shows us wh	0.999993856	1
"See Professor ?"	0.5	0
	Correct Percent	85%

LR Test Result for Snape		
content	p1	whether the character
Wherever he was, Filch must know a sho	0.000911051	1
Harry didn't see how he could -- yet he	0.01798621	1
"Cheer up," said Ron, "'s always taking p	0.01798621	1
Harry didn't know whether he was imag	0.000911051	1
At the start-of-term banquet, Harry had	0.01798621	1
He had looked through his books at the	0.268941421	1
At times, he even wondered whether w	0.119202922	1
"Quirrell said --" "Professor , Harry."	0.268941421	1
Your friend Miss Granger accidentally kn	0.002472623	1
"But always seemed to hate me so muc	0.01798621	1
The idea of Voldemort certainly scared t	2.26E-06	1
"Take him up to the hospital wing," spat	0.119202922	1
And then, your father did something co	0.006692851	1
"He's just made that rule up," Harry mutt	0.119202922	1
He had looked through his books at the	0.268941421	1
A moment later, Professor McGonagall l	0.119202922	1
Harry didn't know whether he was imag	0.000911051	1
Quietly as possible, they crept along the	0.000911051	1
Peering around it, however, they saw no	0.119202922	1
"But 's trying to steal it."	0.047425873	1
	Correct Percent	100%

Figure 4

Conclusion

According to the analysis, we see that the Logistic Regression (LR) model has better performance on predicting the quotes after adjusting the weight, though the evaluation shows the precisions of the two methods do not have a big difference.

In both methods, the precision of predicting Snape's quotes is higher than predicting Dumbledore's. There are many reasons can cause this.

First, the randomly chosen quotes might not "truly" related with the character. For example, the text might be quoted from the plot that only mention the character and state some facts without strong sentiments. Additionally, as we see in the part 1 analysis, there are much more negative words for the character Snape, while for Dumbledore, the difference of the number of the positive words and the number of negative words related with him is not as big as Snape. This might also the reason that why the prediction of Dumbledore's quotes has lower accuracy.