6,a

Top ten Non-Hate Word

1. Thanks
2. sf
3. html
4. Sports
5. 15
6. information
7. Check
8. email
9. facebook
10. [

Top Ten Hate Word

1. non
2. mud
3. asian
4. ape
5. Jews
6. dumb
7. liberal
8. non-white
9. filth
10. Scum

2c. $P(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x,y)$

and since X and y are independent. We have

$$-\sum_{x \in X} \sum_{y \in Y} P(x) P(y) \log P(x) \log P(y)$$

$$= -\sum_{x \in X} P(x) \log P(x) \sum_{y \in Y} P(y) \log P(y)$$

$$= -H(X) * H(Y)$$

where $H(x)$, $H(Y)$ are the marginal entropy that gives us average information when observing X and Y.

6

B) I noticed that logistic regression took longer to train since it's an iterative process. I noticed a slight improvement when adding a regurolization term to the loss function which prevents overfitting. the Optimal $\lambda = 0.0001$. while the other Values Quickly underflow that's because the Weight, were approching 0. I achived %74 accuracy on traing set and ~%70 on the test set. So it's safe to assume I overfitted on the training set.

5.

c) We can try to tune Parameters for n-gram, and Come up with a ~~dd~~ different model and Observe it's accuracy. Also, the word "look" and "others" could be added as additional features. Also adding Punctuation into our bag-of-word. Adding these as ~~feature~~ feature could give better context for future document to make more accurate Prediction

5B

Prior = $P(+) = P(-) = 0.5$

5 B)

Review	great	Amazing	epic	Boring	terrible	disappointing	Sum V
doc = S	2	1	0	0	1	1	5

Sum colums

							Sum V
"+"	7	3	6	1	1	2	20
"−"	3	1	2	6	4	3	19

$$P(t|c) = \frac{T_{ct} + 1}{\sum_{v \in V}(T_{ct}) + |V|}$$ Laplace smoothing

Joint "+": $\frac{7+1}{20+5}$ $\frac{3+1}{20+5}$ $\frac{6+1}{20+5}$ $\frac{1+1}{20+5}$ $\frac{1+1}{20+5}$ $\frac{2+1}{20+5}$

Joint "−": $\frac{3+1}{19+5}$ $\frac{1+1}{19+5}$ $\frac{2+1}{19+5}$ $\frac{6+1}{19+5}$ $\frac{4+1}{19+5}$ $\frac{3+1}{19+5}$

take the log and sum each term and we get ~~we get~~.

$P(+|s) = -9.451$

$P(-|s) = -10.122$

$-0.941 > 10.122$

So our model with Laplace model still predicts "+".

Prior = $P(+) = P(-) = 0.5$

5a)

Review	great	amazing	epic	Boring	terrible	disappointing	Sum of word
S	2	1	0	0	1	1	5
"Sum" colum							
+	7	3	6	1	1	2	20
−	3	1	2	6	4	3	19
Joint +	7/20	3/20	6/20	1/20	1/20	2/20	
Joint '−'	3/19	1/19	2/19	6/19	4/19	3/19	

$$P(+|s) = \underbrace{-0.693}_{\text{prior}} + 2(-1.050) + 1(-1.897) + 0 + 0 + 1(-2.996) + 1(-2.303) = \underline{-9.988}$$

$$P(-|s) = -0.693 + 2(\cancel{-2.944}\ -1.846) + 1(-2.944) + 0 + 0 + 1(-1.558) + 1(-1.846) = \underline{-10.733}$$

$-9.988 > -10.733$, hence our Naive Bayes model predicts "+".

3 c)

100 = total article

- 15 fake
  - 3 Came from NYT
  - 12 from Buzzfeed
- 85 real
  - 51 from NYT
  - 34 from Buzzfeed

$$P(f \mid NYT) = \frac{\text{fake article from NYT}}{\text{Total article from NYT}}$$

$$= \frac{3}{3+51}$$

$$= 0.056$$

$$= \% 5.6$$

3.B $P(A,B,C) =$

$$Pr(A|B,C) * P(B|C) * P(C)$$

$$P(fake|read) = 0.5$$

$$P(NYT|fake, read) = 0.2$$

$$P(read) = 40/1000$$

$$P(NYT, fake, Read) =$$

$$P(NYT|fake, read) * P(fake|read) * P(read) =$$

$$0.5 * 0.2 * 0.04 = 0.004$$

$$= \% 4$$

3.a) Bayes Rule $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

F = fake

BF = Buzz feed

$$P(F|BF) = \frac{P(BF|F)\,P(F)}{P(BF)} =$$

$$\frac{0.7 \times 0.05}{0.25} = 0.14$$

$$\boxed{14\%}$$

2 c. Let X and Y be independent.

then the joint Entropy is

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x,y)$$

Marginal Entropy is

$$H(X) = \sum_{x \in X} P(x) \log P(x)$$

Hence Marginal Entropy is the average total information given X, Y.

2b Uniforn dist. , Continous function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \in [a,b] \\ 0 & \text{else} \end{cases}$$

$$P(x) = \frac{1}{m} \quad \forall x \in [1, M]$$

then $a = 1,\ b = m$

$$H(x) = \int_{a}^{b} \frac{1}{m-1} \log(m-1)\, dx$$

$$= \log(m-1)$$

2a $H(x) = -\sum p(x) \log p(x)$

$-\left[0.15 \log(0.15) + 0.4 \log(0.4) + 0.45 \log(0.45)\right]$

$-\left[-0.285 + (-0.3665) + (-0.359)\right]$

$= 1.0104$

4) $\theta^* = \underset{\theta}{\operatorname{argmax}} \left[\left[\sum_{i=1}^{m} \log P(y^i|x^i)\right] - \alpha \sum_{j=1}^{n} |\theta_j| \right]$

$\cancel{\theta}$ $\cancel{\geq} 0$, where $\alpha \geq 0$

$$\sum_{i=1}^{m} \log P(y^i|x^i) - \alpha \sum_{j=1}^{n} |\theta_j|$$

$$\leq \sum_{i=1}^{m} \log P(y^i|x^i)$$

$$= \hat{\theta}$$

and since $\theta^* \leq \hat{\theta}$

note $\|\theta\|_2 = \sqrt{\sum_{i=1}^{n} \theta_i^2}$ and $\|\theta\|_2^2 = \sum_{i=1}^{n} \theta_i^2$

$$\text{So } \sum_{i=1}^{n} \theta^{*2} = \|\theta^*\|_2^2$$

$$\leq \sum_{i=1}^{n} \theta^* \hat{\theta}$$

$$\leq \sum_{i=1}^{n} \hat{\theta}^2 = \|\hat{\theta}\|_2^2$$

Hence $\|\theta^*\|_2^2 \leq \|\hat{\theta}\|_2^2$