Part A:

1) Team Jarlsberg
   Siddhesh Madeshwar
   Zachary Noel
   Erin Dolson

2)
   a) Bernoulli - similar to Multinomial but no count per document, binary per document (does that term exists there or it does not)

      i) $P(X_{peony} = True \mid Class = 2)$
         Occurs both times in both documents, laplace = 1, 2 documents in class 2, 2 binary
         (SUM(does it occur in doc in class) + 1)/((Num of docs in class) * (2 b/c binary)
         = (2+1)/(2+2) = ¾ = .75

      ii) $P(X_{crocus} = True \mid Class = 2)$
      Occurs in one document from class 2, laplace = 1, 2 documents in class 2, 2 for binary
         = (1+1)/(2+2) = 2/4 = .500

      iii) $P(X_{peony} = True \mid Class = 1)$
      Occurs in one document from class 1, laplace = 1, 1 document1 in class 1, 2 for binary
         = (1+1)/(1+2) = 2/3 = .667

   b) Multinomial NB model with LaPlace

$$\frac{T_{ct}+1}{(\sum_{t'\in V} T_{ct'})+|V|}$$

      i) P(X = peony | Class = 2) = .1786
         T_{ct} = 4 (number of times peony shows up in the class)
         V = 14 (unique words in all)
         $\Sigma_{t' \text{ in } V} T_{ct'}$ = 14 (number of times t' (from V) shows up in training docs in C)
         $(T_{ct} + 1)/((\Sigma_{t' \text{ in } V} T_{ct'}) + |V|) = (4+1)/(14+14) = 5/28 = .1786$

      ii) P(X = crocus| Class = 2) = .0714
         T_{ct} = 1
         V = 14
         $\Sigma_{t' \text{ in } V} T_{ct'}$ = 14
         $(T_{ct} + 1)/((\Sigma_{t' \text{ in } V} T_{ct'}) + |V|) = (1+1)/(14+14) = 2/28 = .0714$

      iii) P(X = peony| Class = 1) = 0.0909
         T_{ct} = 1
         V = 14
         $\Sigma_{t' \text{ in } V} T_{ct'}$ = 8
         $(T_{ct} + 1)/((\Sigma_{t' \text{ in } V} T_{ct'}) + |V|) = (1+1)/(8+14) = 2/22 = 0.0909$

   c)

$P(Class = 1) = ¼$
$P(Class = 2) = ½$
$P(Class = 3) = ¼$

$P(X_{daffodil} = True \mid Class = 1) = (0 +1)/(1+2) = 1/3 = .333$
$P(X_{daffodil} = True \mid Class = 2) = (1 +1)/(2+2) = 2/4 = .500$
$P(X_{daffodil} = True \mid Class = 3) = (0 +1)/(1+2) = 1/3 = .333$

$P(X_{crocus} = True \mid Class = 1) = (0 +1)/(1+2) = ⅓ = .333$
$P(X_{crocus} = True \mid Class = 2) = (1 +1)/(2+2) = 2/4 = .500$
$P(X_{crocus} = True \mid Class = 3) = (0 +1)/(1+2) = ⅓ = .333$

$P(X_{daisy} = True \mid Class = 1) = (0 +1)/(1+2) = ⅓ = .333$
$P(X_{daisy} = True \mid Class = 2) = (0 +1)/(2+2) = ¼ = .25$
$P(X_{daisy} = True \mid Class = 3) = (1 +1)/(1+2) = ⅔ = .667$

$P(X_{tulip} = True \mid Class = 1) = (1 +1)/(1+2) = ⅔ = .667$
$P(X_{tulip} = True \mid Class = 2) = (0 +1)/(2+2) = ¼ = .25$
$P(X_{tulip} = True \mid Class = 3) = (1 +1)/(1+2) = ⅔ = .667$

$P(X_{clemantis} = True \mid Class = 1) = (1 +1)/(1+2) = ⅔ = .667$
$P(X_{clemantis} = True \mid Class = 2) = (2 +1)/(2+2) = ¾ = .75$
$P(X_{clemantis} = True \mid Class = 3) = (0 +1)/(1+2) = ⅓ = .333$

$P(X_{peony} = True \mid Class = 1) = (1 +1)/(1+2) = ⅔ = .667$
$P(X_{peony} = True \mid Class = 2) = (2 +1)/(2+2) = ¾ = .75$
$P(X_{peony} = True \mid Class = 3) = (0 +1)/(1+2) = ⅓ = .333$

P(c1 | "daffodil crocus daisy tulip clematis peony")
= ¼ * .333 * .333 * .333 * .667 * .667 * .667 = 0.002739
P(c2 | "daffodil crocus daisy tulip clematis peony")
= 1/2 * .500 * .500 * .25 * .25 * .75 * .75 = 0.0043945
P(c3 | "daffodil crocus daisy tulip clematis peony")
= ¼ * .333* .333 * .667 * .667 * .333 * .333 = 0.0013676
Predicted class: Class 2

d)
P(Class = 1) = ¼
P(Class = 2) = ½
P(Class = 3) = ¼

P(X = daffodil | Class = 1) = (0 +1)/(8+14) = 1/22 = 0.0455
P(X = daffodil | Class = 2) = (1 +1)/(14+14) = 2/28 = 0.0714
P(X = daffodil | Class = 3) = (0 +1)/(7+14) = 1/21 = 0.0476

P(X = crocus| Class = 1) = (0 +1)/(8+14) = 1/22 = 0.0455

P(X = crocus| Class = 2) = (1 +1)/(14+14) = 2/28 = 0.0714
P(X = crocus| Class = 3) = (0 +1)/(7+14) = 1/21 = 0.0476

P(X = daisy | Class = 1) = (0 +1)/(8+14) = 1/22 = 0.0455
P(X = daisy | Class = 2) = (0 +1)/(14+14)= 1/28 = 0.0357
P(X = daisy | Class = 3) = (1 +1)/(7+14) = 2/21 = 0.0952

P(X = tulip | Class = 1) = (1 +1)/(8+14) = 2/22 =0.0455
P(X = tulip | Class = 2) = (0 +1)/(14+14) = 1/28 = 0.0357
P(X = tulip | Class = 3) = (2 +1)/(7+14) = 3/21 = 0.14286

P(X = clematis | Class = 1) = (1 +1)/(8+14) = 2/22 = 0.0909
P(X = clematis | Class = 2) = (4 +1)/(14+14) = 5/28 = 0.17857
P(X = clematis | Class = 3) = (0 +1)/(7+14) = 1/21 = 0.0476

P(X = peony | Class = 1) = (1 +1)/(8+14) = 2/22 = 0.0909
P(X = peony | Class = 2) = (4 +1)/(14+14) = 5/28 = 0.17857
P(X = peony | Class = 3) = (0 +1)/(7+14) = 1/21 = 0.0476

P(c1 | "daffodil crocus daisy tulip clematis peony")
= ¼ * 0.0455 * 0.0455 * 0.0455 * 0.0455 * 0.0455 * 0.0455 = 2.218e-9
P(c2 | "daffodil crocus daisy tulip clematis peony")
= 1/2 * 0.0714 * 0.0714 * 0.0357 * 0.0357 * 0.17857 * 0.17857 = 1.036e-7
P(c3 | "daffodil crocus daisy tulip clematis peony")
= ¼ * 0.0476 * 0.0476 * 0.0952 * 0.14286 * 0.0476 * 0.0476 = 1.745e-8
Predicted class: Class 2

3)
a)

	cat	bat	rat	fat	mat	pat	sat
Doc 1	3	1	1	1	0	0	0
Doc 2	3	3	1	0	1	1	0
Doc 3	1	0	1	1	1	1	1

b)

The tf-idf weight of a term is the product of its tf weight and its idf weight.

$$w_{t,d} = \log(1 + tf_{t,d}) * \log_{10}(\frac{N}{df_t})$$

$tf_{t,d}$: Document frequencies of t (num of docs that contain t)
N: Num of documents
df_t: Frequency of term t in doc d (num of times t occurs in d)

Term	d_1f_t	d_2f_t	d_3f_t	N	$tf_{t,d}$	$w_{t,d1}$	$w_{t,d2}$	$w_{t,d3}$
bat	1	3	0	3	2	.2276	0	-
cat	3	0	1	3	2	0	-	.2276
fat	1	0	1	3	2	.2276	-	.2276
mat	0	1	1	3	2	-	.2276	.2276
pat	0	1	1	3	2	-	.2276	.2276
rat	1	1	1	3	3	.2873	.2873	.2873
sat	0	0	1	3	1	-	-	.1436

c) The term-document pairs with the highest TF-IDF values are (Rat, Doc1), (Rat, Doc2), and (Rat, Doc3).