Edward Plesa

CS421

Assignment 2

1b.

Similarity 1: The most common bigrams were “of_the” and “in_the”.

Similarity 2: The most common unigrams were “the”, “to”, “of”, “and”.

Similarity 3: Subject matter of the words seems to most revolve around people and their interactions, e.g. bigrams with “said” followed by a name.

Difference 1: The most common bigrams in “A room with a view” were character names e.g. “miss_bartlett” and for “Middlemarch” were mainly bigrams including a pronoun and some other word.

Difference 2: The word choice in “A room with a view” seems more varied and wider based on the unigrams in the book.

Difference 3: The title of “Middlemarch” appears quite often in the text.

Based on the observations made above, it seems that the two books are very similar in terms of their subject matters, based on the common unigrams and bigrams. It talks about the experiences and interactions between characters. The writing style is also very similar but “A room with a view” is a little bit better with vocabulary choice probably denoting the the characters have more dialogue compared to “Middlemarch”. Seems like there are also more characters in “A room with a view” since there are more names appearing in the ngrams. Overall I would say that both authors have a nearly identical writing style and that the contents of their books are also very similar in nature.

2.

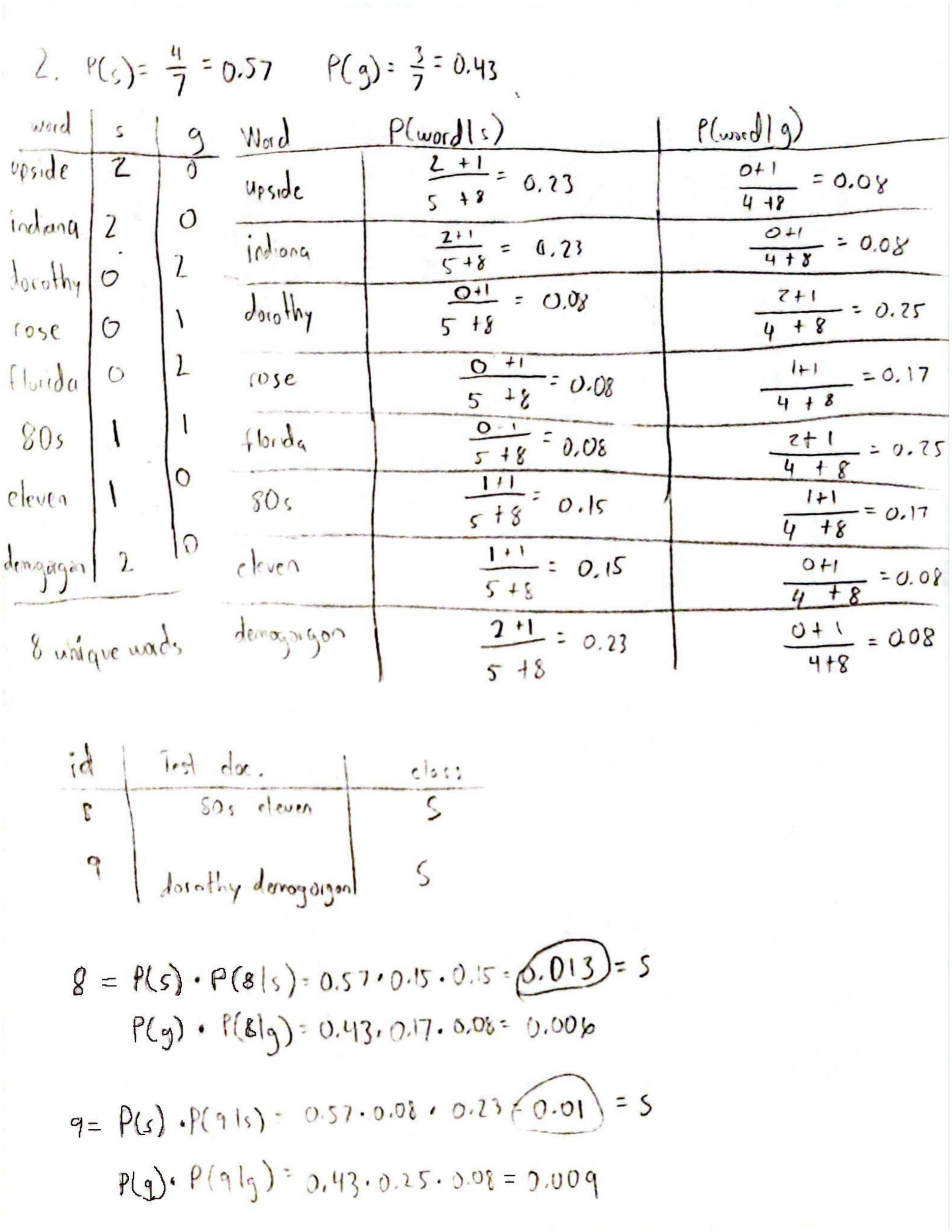
1. Test document “80s eleven” will get class “s” for stranger things.

   Test document “dorothy demogorgon” will get class “s” for stranger things.

2. The more words belonging to a certain class in the test documents made it more likely that the test document belonged to that class.

3. In test document “dorothy demogorgon” the class predictions were closer because the values for the probability of the words belonging to a class were closer for both “s” and “g”. “S” won out only because of the higher prior probability.

WORK FOR 2:



2. $P(s) = \frac{4}{7} = 0.57$ $P(g) = \frac{3}{7} = 0.43$

word	s	g
upside	2	0
indiana	2	0
dorothy	0	2
rose	0	1
florida	0	2
80s	1	1
eleven	1	0
demogorgon	2	0

8 unique words

Word	P(word|s)	P(word|g)
upside	$\frac{2+1}{5+8} = 0.23$	$\frac{0+1}{4+8} = 0.08$
indiana	$\frac{2+1}{5+8} = 0.23$	$\frac{0+1}{4+8} = 0.08$
dorothy	$\frac{0+1}{5+8} = 0.08$	$\frac{2+1}{4+8} = 0.25$
rose	$\frac{0+1}{5+8} = 0.08$	$\frac{1+1}{4+8} = 0.17$
florida	$\frac{0+1}{5+8} = 0.08$	$\frac{2+1}{4+8} = 0.25$
80s	$\frac{1+1}{5+8} = 0.15$	$\frac{1+1}{4+8} = 0.17$
eleven	$\frac{1+1}{5+8} = 0.15$	$\frac{0+1}{4+8} = 0.08$
demogorgon	$\frac{2+1}{5+8} = 0.23$	$\frac{0+1}{4+8} = 0.08$

id	Test doc.	class
8	80s eleven	S
9	dorothy demogorgon	S

$8 = P(s) \cdot P(8|s) = 0.57 \cdot 0.15 \cdot 0.15 = \boxed{0.013} = S$

$P(g) \cdot P(8|g) = 0.43 \cdot 0.17 \cdot 0.08 = 0.006$

$9 = P(s) \cdot P(9|s) = 0.57 \cdot 0.08 \cdot 0.23 = \boxed{0.01} = S$

$P(g) \cdot P(9|g) = 0.43 \cdot 0.25 \cdot 0.08 = 0.009$

3.

3. $b = 1 \times 0.5 = 0.5$ input: and the haters gonna hate hatehate hate hate

-> contains 2 stop words (and, the)

-> contains 4 repetitions

-> contains 9 total words

-> contains 1 "haters"

$$P(\text{beyonce}) = \sigma(0.5 + (9 \cdot 0.26 + 2 \cdot 0.04 + 4 \cdot 0.16 + 1 \cdot 0.05)$$

$$= \sigma(5.57) = \frac{1}{1+e^{-5.57}} = 0.9962 \ast = Y_{\text{Beyonce}}$$

$$P(\text{Taylor Swift}) = \sigma(0.5 + (9 \cdot 0.25 + 2 \cdot 0.02 + 4 \cdot 0.36 + 1 \cdot 0.04)$$

$$= \sigma(4.27) = \frac{1}{1+e^{-4.27}} = 0.9862 = Y_{\text{Taylor Swift}}$$

Predicted Class = $\underline{\underline{\text{Beyonce}}}$