Problem 1: Laplace Smoothing

Let $|V| = 5$ where V is the vocabulary and $N = \sum_{i}^{|V|} freq(w_i) = 1,200$ where N is the number of words in the corpus. Without smoothing, the probabilities of each noun in the corpus are:

$P(maple) = \frac{freq(maple)}{N} = \frac{600}{1,200} = 0.5$

$P(oak) = \frac{freq(oak)}{N} = \frac{400}{1,200} = 0.333$

$P(pine) = \frac{freq(pine)}{N} = \frac{180}{1,200} = 0.15$

$P(spruce) = \frac{freq(spruce)}{N} = \frac{20}{1,200} = 0.167$

$P(aspen) = \frac{freq(aspen)}{N} = \frac{0}{1,200} = 0$

Using Laplace smoothing (aka: Add-One Smoothing), the new frequencies for each noun:

$new\_freq(maple) = (freq(maple) + 1) * \frac{N}{N+V} = (600 + 1) * \frac{1,200}{1,200+5} = \frac{721,200}{1,205} = 598.506$

$new\_freq(oak) = (freq(oak) + 1) * \frac{N}{N+V} = (400 + 1) * \frac{1,200}{1,200+5} = \frac{481,200}{1,205} = 399.336$

$new\_freq(pine) = (freq(pine) + 1) * \frac{N}{N+V} = (180 + 1) * \frac{1,200}{1,200+5} = \frac{217,200}{1,205} = 180.249$

$new\_freq(spruce) = (freq(spruce) + 1) * \frac{N}{N+V} = (20 + 1) * \frac{1,200}{1,200+5} = \frac{25,200}{1,205} = 20.913$

$new\_freq(aspen) = (freq(aspen) + 1) * \frac{N}{N+V} = (0 + 1) * \frac{1,200}{1,200+5} = \frac{1,200}{1,205} = 0.966$

Given these new frequencies, the new probabilities are calculated:

$new\_P(maple) = \frac{new\_freq(maple)}{N} = \frac{598.506}{1,200} = 0.499$

$new\_P(oak) = \frac{new\_freq(oak)}{N} = \frac{399.336}{1,200} = 0.333$

$new\_P(pine) = \frac{new\_freq(pine)}{N} = \frac{180.249}{1,200} = 0.15$

$new\_P(spruce) = \frac{new\_freq(spruce)}{N} = \frac{20.913}{1,200} = 0.017$

$new\_P(aspen) = \frac{new\_freq(aspen)}{N} = \frac{0.996}{1,200} = 0.001$

Completing the table:

Noun	**Freq.**	**Unsmoothed Prob.**	**Smoothed Freq.**	**Smoothed Prob.**
maple	600	0.5	598.506	0.499
oak	400	0.333	399.336	0.333
pine	180	0.15	180.249	0.15
spruce	20	0.017	20.913	0.017
aspen	0	0	0.996	0.001

Table 1: Unsmoothed and smoothed probabilities and frequencies.

2: Grammars and Recursive Transition Networks

(a) Grammar A and Grammar B - **DIFFERENT**

Grammar A requires the NP to begin with an article but Grammar B does not.

An example POS tag sequence **accepted by Grammar B** and not by Grammar A is: **noun**.

(b) Grammar A and Grammar C - **DIFFERENT**

Grammar C requires one or more adjectives after the article but Grammar A requires zero or more adjectives after the article.

An example POS tag sequence **accepted by Grammar A** and not by Grammar C is: **art noun**.

(c) Grammar A and RTN-2 - **DIFFERENT**

Grammar A requires the NP to begin with an article but RTN-2 does not.

An example POS tag sequence **accepted by RTN-2** and not by Grammar A is: **noun**.

(d) Grammar A and RTN-3 - **DIFFERENT**

RTN-3 requires one or more adjectives after the article but Grammar A requires zero or more adjectives after the article.

An example POS tag sequence **accepted by Grammar A** and not by RTN-3 is: **art noun**.

(e) Grammar B and RTN-2 - **SAME**

(f) Grammar C and RTN-1 - **DIFFERENT**

Grammar C requires the NP to end with one or more nouns but RTN-1 does not.

An example POS tag sequence **accepted by RTN-1** and not by Grammar C is: **art adj**.

(g) Grammar C and RTN-3 - **SAME**

(h) RTN-1 and RTN-3 - **DIFFERENT**

RTN-3 requires the NP to end with one or more nouns but RTN-1 does not.

An example POS tag sequence **accepted by RTN-1** and not by Grammar C is: **art adj**.

3: N-Gram Probabilities

Let $|V| = 18$ where V is the vocabulary and $N = \sum_{i}^{|V|} freq(w_i) = 34$ where N is the number of words in the tiny text corpus.

(a) $P(the) = \frac{5}{34}$

(b) $P(VERB) = \frac{6}{34}$

(c) $P(young \mid girl)$

(d) $P(girl \mid young)$

(e) $P(and \mid women)$

(f) $P(thanked \mid young\ girl)$

(g) $P(five \mid gave\ her)$

(h) $P(the \mid ART)$

(i) $P(cross \mid NOUN)$

(j) $P(thanked \mid VERB)$

(k) $P(NUM \mid PRO)$

(l) $P(ART \mid VERB)$