Estimating Frequencies Predicted by Zipf's Law

Assume you downloaded Shakespeare's *Hamlet* via NLTK and obtained the following counts (after tokenization and lowercasing):

$N = 36421$ (Number of tokens), $V = 4789$ (Vocabulary size, i.e. number of types)

Answer the following questions after writing code:

1. (2 points) According to Herdan's Law (or Heaps' Law), the relationship between the number of types $|V|$ and number of tokens N in a corpus is:

$$|V| = kN^{\beta}$$

For $\beta = 0.7$ (a standard value typically found for English), compute the value of k for your lowercased, tokenized *Hamlet* corpus.

2. (5 points) We can describe Zipf's law more generally, using a formula whose parameters vary slightly across languages. Formally, let:

- $|V|$: the number of types (hyperparameter),
- s : the exponent characterizing the distribution (hyperparameter),
- k : the rank of a type.

Zipf's law predicts that out of a population of N elements, the normalized frequency of the element of rank k, $f(k; s, |V|)$, is:

$$f(k; s, |V|) = \frac{1/k^s}{\sum_{n=1}^{|V|} 1/n^s}$$

Fill the following table based on the (lowercased, tokenized) *Hamlet* corpus. Use $s = 1$. Report relative frequencies up to 5 decimal digits.

Type	**Rank**	**Predicted relative frequency**
the	3	
hamlet	54	
royally	3411	
rose	778	
honourable	1003	

Extra credit (5 points) Do the predicted relative frequency estimates match the observed relative frequency of each word in the NLTK *Hamlet* corpus? Extend the table by adding an extra column for the observed frequency, and write a short note on the fit obtained.

Type	**Rank**	**Predicted frequency**	**Observed frequency**
the	3		
hamlet	54		
royally	3411		
rose	778		
honourable	1003		