CS446
P1
Michael Chen

Question 1:
The top terms don't really seem related to the story at all. It's mostly words like her, I, she, not, you, his, and other words that seem pretty generic. Elinor does appear a whopping 685 times, and since that's a pretty unique name I'd say it seems pretty relevant to the story. Other names like Marianne and Veri appear a lot too, which also seem pretty unique to this story. Overall though, most of the words seem to just be everyday words.

2. Honestly yeah. Most of these words should probably be stop words. Like not, you, his, had, but, have, all, so. These are all just commonly used everyday words. I don't think any changes need to be made to the stemming or tokenization though, overall it looks pretty clean.

3. It seems to follow a square root kind of pattern. As the number of words goes up, the number of unique words increases sharply in the beginning, but stops increasing as much later on. I imagine as the text gets larger and larger, it'll eventually flatten out substantially.

4. I think so. The graph I generated from the unique words to total words data seems to follow heap's law pretty closely. And I don't see why Sense and Sensibility shouldn't follow Heaps Law, it's not like some strange text written by aliens or something. It's like an old English classic.