Question 1

In text4, the *Inaugural Addresses by US Presidents*, the plural form of the word 'father', i.e, 'fathers' appears a total of 32 times, in most cases of which, it follows plural possessive determinators like 'our' and 'their', and is majorly followed by verbs. In comparison, the singular form, i.e 'father' follows the definite article 'the' in the context where it's a proper noun, and mostly follows singular possessive determinators like 'his' or 'my', in the context where it's a common noun. In text2, *Sense and Sensibility*, the word 'affection' mainly follows possessive pronouns, and in a few cases, follows adjectives/adverbs which modify the noun 'affection'.

Question 2

As the attestation count increases, it can be observed that the number of words having that larger attestation count value decreases. This can be attributed to the fact that there will be a lot of words in a text which will be used fewer number of times, but more general words (such as, stop words) tend to appear more number of times (corresponding to a higher attestation count), and the number of such words will be lower.

Question 3

```
>>> from nltk.book import *
>>> austen = FreqDist(text2)
>>> [word for word in austen if austen[word]>10 and word[-2:]=='ly']
```

The output of the above code snippet is:

```
['family', 'merely', 'really', 'only', 'highly', 'earnestly', 'frequently',
'exactly', 'equally', 'hardly', 'Certainly', 'perfectly', 'entirely',
'certainly', 'melancholy', 'early', 'scarcely', 'warmly', 'easily',
'exceedingly', 'lively', 'greatly', 'folly', 'likely', 'instantly',
'particularly', 'especially', 'immediately', 'hastily', 'directly',
'thoroughly', 'extremely', 'probably', 'suddenly', 'eagerly',
'heartily', 'naturally', 'openly', 'lately', 'reply']
```

Question 4

By looking at the top-50 most common words in *Sense and Sensibility* that're not entirely punctuation characters, the proper names that occur are: *elinor* and *marianne*.

The words similar to *elinor* are: *marianne, she, he, it, edward, him, lucy, i, her, you, which, that, willoughby, me, they, herself, them, be, all, what.*

- Nouns: *marianne, edward, lucy, willoughby*
- Pronouns: *she, he, it, him, i, her, you, that, me, they, herself, them*

The words similar to *marianne* are: *elinor, it, she, he, edward, her, lucy, i, me, you, willoughby, them, him, they, herself, that, all, which, what, there.*

- Nouns: *elinor, edward, lucy, willoughby*
- Pronouns: *it, she, he, her, i, me, you, them, him, they, herself, that, there*

In general, the words returned similar to either *elinor* or *marianne* are *proper nouns*, and *pronouns* generally used to describe these nouns.

Question 5

The words similar to *husband* are: *sister, mother, heart, opinion, own, situation, mind, daughter, brother, eyes, letter, ladyship, and, family, life, companion, children, way, time, elinor.*

- Nouns: *sister, mother, heart, opinion, situation, mind, daughter, eyes, letter, ladyship, family, life, companion, children, time, elinor*

In general, most of the words returned signify relations, like 'husband', and a general class that they belong to is *noun.*

Question 6

```
>>> from nltk.book import *
>>> bd = FreqDist((bigram for bigram in bigrams(text1) if bigram[1] == 'whale'))
>>> bd.max()
```

The output on running the above code snippet is:

```
('the', 'whale')
```

From the above result, it can be concluded that 'the' is the most popular word preceding 'whale' in Moby Dick.

Question 7

Words are generated using the *generate()* function in nltk in the following way: a pseudo-random number generated corresponds to a particular bin between 0 & 1, which in turn corresponds to an n-gram frequency in the frequency distribution of n-grams for a text. In particular, in a trigram model, each consequent word is generated randomly in this fashion, from among the trigrams containing the previous two words in the generated sequence. In my opinion, the trigram model, or in general ngram modeling is linguistically naive, wherein, it's not capturing any grammatical rules or semantics of the sentence, but is more of a statistical combination of words.