SLP_3e_chp3

September 25, 2021

1 N-grams

[ ]:
```
import nltk
nltk.download()
# download nltk.book
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

[1]:
```
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, …, text9 and sent1, …, sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

[5]:
```
text1.concordance("""its water is so transparent that""")
```

```
no matches
```

This simple example (which can be extended to other texts) suggests that many valid sentences in the English language are novel and we might never find them used earlier. Finding probabilities for the word which should appear next becomes challenging.

The heuristics we'll use are the chain rule to calculate probabilites, and not going too deep, we'll also use the Markov assumption. For an n-gram, this is -

$$P(w_n|w_{1:n-1}) = P(w_n|w_{n-N+1:n-1})$$

[ ]:
```

```