Arisa Chue
AI Period 2
Mr. Eckel

NLTK OW

For 1 OW credit I did:

- Chapter 2, questions 25, 26, and 28
- Chapter 5, questions 34 and 35 (a new chapter)

Chapter 2

25. Define a function find_language() that takes a string as its argument, and returns a list of languages that have that string as a word. Use the udhr corpus and limit your searches to files in the Latin-1 encoding.

With word “solo”: ['Asturian_Bable-Latin1', 'Bikol_Bicolano-Latin1', 'Chamorro-Latin1', 'Kimbundu_Mbundu-Latin1', 'Lingala-Latin1', 'Samoan-Latin1', 'Wolof-Latin1']

26. What is the branching factor of the noun hypernym hierarchy? I.e. for every noun synset that has hyponyms — or children in the hypernym hierarchy — how many do they have on average? You can get all noun synsets using wn.all_synsets('n').

average/branching factor: 4.543820763194153

28. Use one of the predefined similarity measures to score the similarity of each of the following pairs of words. Rank the pairs in order of decreasing similarity. How close is your ranking to the order given here, an order that was established experimentally by (Miller & Charles, 1998): car-automobile, gem-jewel, journey-voyage, boy-lad, coast-shore, asylum-madhouse, magician-wizard, midday-noon, furnace-stove, food-fruit, bird-cock, bird-crane, tool-implement, brother-monk, lad-brother, crane-implement, journey-car, monk-oracle, cemetery-woodland, food-rooster, coast-hill, forest-graveyard, shore-woodland, monk-slave, coast-forest, lad-wizard, chord-smile, glass-magician, rooster-voyage, noon-string.

[('car', 'automobile', 1.0), ('midday', 'noon', 1.0), ('coast', 'shore', 0.5), ('tool', 'implement', 0.5), ('boy', 'lad', 0.3333333333333333), ('journey', 'voyage', 0.25), ('coast', 'hill', 0.2), ('shore', 'woodland', 0.2), ('monk', 'slave', 0.2), ('lad', 'wizard', 0.2), ('magician', 'wizard', 0.16666666666666666), ('lad', 'brother', 0.14285714285714285), ('gem', 'jewel', 0.125), ('asylum', 'madhouse', 0.125), ('brother', 'monk', 0.125), ('monk', 'oracle', 0.125), ('bird', 'crane', 0.1111111111111111), ('cemetery', 'woodland', 0.1111111111111111), ('glass', 'magician', 0.1111111111111111), ('crane', 'implement', 0.1), ('food', 'fruit', 0.09090909090909091), ('coast', 'forest', 0.09090909090909091), ('chord', 'smile', 0.09090909090909091), ('furnace', 'stove', 0.07692307692307693), ('forest', 'graveyard', 0.07142857142857142), ('bird', 'cock', 0.0625),

('food', 'rooster', 0.0625), ('noon', 'string', 0.058823529411764705), ('journey', 'car', 0.05), ('rooster', 'voyage', 0.041666666666666664)]
Using path_similarity, the ranking is a bit different but car-automobile is still at the top and rooster-voyage and noon-string is near the bottom.

Chapter 5

34. There are 264 distinct words in the Brown Corpus having exactly three possible tags.

1. **Print a table with the integers 1..10 in one column, and the number of distinct words in the corpus having 1..10 distinct tags in the other column.**
2. **For the word with the greatest number of distinct tags, print out sentences from the corpus containing the word, one for each possible tag.**

0 0
1 47328
2 7186
3 1146
4 265
5 87
6 27
7 12
8 1
9 1
10 2
[('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD'), ('Friday', 'NR'), ('an', 'AT'), ('investigation', 'NN'), ('of', 'IN'), ("Atlanta's", 'NP$'), ('recent', 'JJ'), ('primary', 'NN'), ('election', 'NN'), ('produced', 'VBD'), ('``', '``'), ('no', 'AT'), ('evidence', 'NN'), ("''", "''"), ('that', 'CS'), ('any', 'DTI'), ('irregularities', 'NNS'), ('took', 'VBD'), ('place', 'NN'), ('.', '.')]
[('Regarding', 'IN'), ("Atlanta's", 'NP$'), ('new', 'JJ'), ('multi-million-dollar', 'JJ'), ('airport', 'NN'), (',', ','), ('the', 'AT'), ('jury', 'NN'), ('recommended', 'VBD'), ('``', '``'), ('that', 'CS'), ('when', 'WRB'), ('the', 'AT'), ('new', 'JJ'), ('management', 'NN'), ('takes', 'VBZ'), ('charge', 'NN'), ('Jan.', 'NP'), ('1', 'CD'), ('the', 'AT'), ('airport', 'NN'), ('be', 'BE'), ('operated', 'VBN'), ('in', 'IN'), ('a', 'AT'), ('manner', 'NN'), ('that', 'WPS'), ('will', 'MD'), ('eliminate', 'VB'), ('political', 'JJ'), ('influences', 'NNS'), ("''", "''"), ('.', '.')]
…
[('But', 'CC'), ('when', 'WRB'), ('to', 'TO-NC'), ('represents', 'VBZ'), ('to', 'IN-NC'), ('consciousness', 'NN-NC'), ('in', 'IN'), ('that', 'WPS-NC'), ('was', 'BEDZ-NC'), ('the', 'AT-NC'), ('moment', 'NN-NC'), ('that', 'CS-NC'), ('I', 'PPSS-NC'), ('came', 'VBD-NC'), ('to', 'IN-NC'), (',', ','), ('and', 'CC'), ('similarly', 'RB'), ('in', 'IN'), ('that', 'WPS-NC'), ('was', 'BEDZ-NC'), ('the', 'AT-NC'), ('moment', 'NN-NC'), ('I', 'PPSS-NC'), ('came', 'VBD-NC'), ('to', 'IN-NC'), (',', ','),

('there', 'EX'), ('is', 'BEZ'), ('much', 'QL'), ('stronger', 'JJR'), ('stress', 'NN'), ('on', 'IN'), ('to', 'IN-NC'), ('.', '.')]
[('Factors', 'NNS-HL'), ('that', 'WPS-HL'), ('inhibit', 'VB-HL'), ('learning', 'VBG-HL'), ('and', 'CC-HL'), ('lead', 'VB-HL'), ('to', 'IN-HL'), ('maladjustment', 'NN-HL')]

35. Write a program to classify contexts involving the word *must* according to the tag of the following word. Can this be used to discriminate between the epistemic and deontic uses of *must*?

must be = epistemic
must sign = deontic
must hold = deontic
…
must look = deontic
must be = epistemic
must have = deontic
Observation: most epistemic uses of must are “must be”, which is similar to what Google said was an example of an epistemic modal verb