Submission Guidelines

1. Any written portion of the assignment:
   (a) Should be printed out and submitted in class on November 28, 2007.
   (b) Should also be emailed electronically to the TAs.

2. Any code for the assignment should ONLY be submitted electronically to the TAs.

3. For electronic submissions:
   (a) Use the subject \textit{CMSC723: Assignment 5}.
   (b) Use tarballs and zip files instead of sending multiple attachments.

Introduction

For this assignment, you will primarily be working with the \texttt{wordnet} module that comes bundled with NLTK. Listings 1 and 2 show how to perform all the functions that you will need for this assignment. Please read these listings carefully before attempting the problems.

\textbf{Important}: Please note that some of these problems may be CPU- and/or memory-intensive. We recommend that you use the linux lab server to test and run your code if this is the case.
Listing 1: Using the wordnet module

```python
from nltk import *

# Look for bank in the Noun taxonomy (just a dictionary)
# The other taxonomies are V, ADJ and ADV
w = N['bank']
print w
dictionary

# Print first two senses (synsets) of the noun bank
for synset in w.synsets()[:2]:
    print synset

dictionary

# Inspect the second synset
synset = w.synsets()[1]
dictionary

# The list of words for this synset
print synset.words

# The gloss describing this synset
print synset.gloss

# Each synset has a unique offset (identifier)
print synset.offset
8420278

# Access the synset using the offset
N.getSynset(8420278)
```

*a financial institution that accepts deposits and channels the money into lending activities; ”he cashed a check at the bank”; ”that bank holds the mortgage on my home”*
Listing 2: Using the wordnet module (contd.)

```python
# List of all direct hyponyms (descendant synsets) of this synset
>>> print synset.relation(HYPONYM)
[[{noun: credit_union}, {noun: Federal_Reserve_Bank, reserve_bank}],
 {noun: agent_bank}, {noun: commercial_bank, full_service_bank},
 {noun: state_bank}, {noun: lead_bank, agent_bank},
 {noun: member_bank}, {noun: merchant_bank, acquirer},
 {noun: acquirer}, {noun: thrift_institution},
 {noun: Home_Loan_Bank}]

# A lazy, breadth-first iterator over the subhierarchy defined by the
# HYPONYM relation and rooted at this synset. If second parameter
# is specified, it restricts the iterator only to that depth.
# So, to get only the direct hyponyms as above
>>> g = synset.closure(HYPONYM, 1)

# Use next() method to get each element of the iterator
>>> print g.next()
{noun: credit_union}
>>> print g.next()
{noun: Federal_Reserve_Bank, reserve_bank}

# The iterator continues until it has nothing left, and then
# it raises an exception. We can use this to loop.
>>> while 1:
    try:
        print g.next()
    except StopIteration:
        break
{noun: agent_bank}
{noun: commercial_bank, full_service_bank}
{noun: state_bank}
{noun: lead_bank, agent_bank}
{noun: member_bank}
{noun: merchant_bank, acquirer}
{noun: acquirer}
{noun: thrift_institution}
{noun: Home_Loan_Bank}
```
Problem 1: WordNet Topology & Statistics

(a) Plot a graph with the number of senses of each verb in the Verb taxonomy on the vertical axis and its polysemy rank\(^1\) on the horizontal axis. What conclusion can you draw from this graph?

(b) A WordNet taxonomy (Noun, Verb etc.), may be simplified and viewed as a directed acyclic graph where the relation between the nodes is one of hypernymy/hyponymy. Figure 1 shows the first three levels of such a simplified Noun taxonomy (not all the nodes are shown). The only information provided to you is that there is a single node at the first (root) level of the taxonomy and that its offset is 1740. There are two pieces of information missing for each node:

- \(W\), the synset words (not the gloss)
- \(N\), the number of descendants (direct & indirect) in the taxonomy rooted at the node.

Complete this given subgraph for the Noun taxonomy with both \(W\) and \(N\) each of the nodes filled in.

(c) For each of the Verb, Adjective and Adverb taxonomies, compute the following statistics:

- The number of monosemous words.
- The number of polysemous words.
- The number of these polysemous senses, i.e., the number of senses that are apportioned to the polysemous words.
- Average polysemy including monosemous words.
- Average polysemy excluding monosemous words.

Hint: You may find a \texttt{FreqDist} very useful.

(d) Consider the above directed acyclic graph for the Noun taxonomy again. Let the \textbf{Branching Factor} (BF) for a node be defined as \(=\) number of direct descendants + 1. Compute the following statistics:

- The range of branching factors (minimum, maximum).
- The average branching factor.

---

\(^1\)its position in a list of these verbs sorted by number of senses, highest first.
Figure 1: A subgraph of the Noun hyponymy taxonomy showing the first three levels (not all nodes are shown). The offset O for the root node is given. W refers to the synset words and N is the number of total descendents rooted at the synset. Ellipsis at a level indicates possible existence of additional nodes.

- The average branching factor excluding leaf nodes.
- The percentage of nodes with BF < 5.
- The percentage of nodes with BF < 20.

What do these figures, combined with with the values of N from (b), suggest w.r.t the shallowness or depth of the taxonomy structure?

**Problem 2: Lexical Semantics**

(a) Write down, in English and without using WordNet or NLTK, between 5 and 10 different senses of the verb (not the noun) *break*. For example, here are two:

- **Sense**: break an object into pieces.
  **Example**: Edgar broke the vase.
- **Sense**: break a bone.
  **Example**: Mildred broke her wrist.
Try to do this without a dictionary if you can, but if you’re not a native speaker of English, use a dictionary if you need to.

(b) Use NLTK to look up the verb senses for *break*. Which WordNet senses do your senses from part (a) match, if any? (One of your senses might match more than one WordNet sense, of course.) For example, Sense 1: matches WN senses 2,3,4,5.

(c) Do any of your senses group naturally into a class with common elements of meaning? How would you group them? (Use a hierarchy if that makes more sense.) Hint: You should examine the list of 5 to 10 senses in the context of the WordNet structure and determine whether there is a way to group these 5 to 10 senses into a smaller number of “equivalence classes”.

(d) Use NLTK to inspect the subgraphs associated with the senses and manually identify a “core meaning” for the verb *break* that covers a reasonably large subset of the different senses.