Problem 1: Furniture

$$\boxed{PMI(w_1, w_2) = log_2\left(\frac{P(w_1 \ \& \ w_2)}{P(w_1) * P(w_2)}\right)} \tag{1}$$

$$\boxed{drift(t, n, m) = \frac{AvgSim(L_{1..n}, t)}{AvgSim(L_{(N-m)..N,t})}} \tag{2}$$

(a) To compute the *semantic drift* score for "futon", first we must compute the point-wise mutual information (PMI) for "futon" and the first 2 words and last 3 words added to the lexicon using eq. 1.

$$PMI(futon,\ chair) = log_2\left(\frac{P(futon,\ chair)}{P(futon) * P(chair)}\right) = log_2\left(\frac{40}{\frac{60}{2{,}000} * \frac{200}{2{,}000}}\right) = 13.703$$

$$PMI(futon,\ couch) = log_2\left(\frac{P(futon,\ couch)}{P(futon) * P(couch)}\right) = log_2\left(\frac{20}{\frac{60}{2{,}000} * \frac{50}{2{,}000}}\right) = 14.703$$

$$PMI(futon,\ board) = log_2\left(\frac{P(futon,\ board)}{P(futon) * P(board)}\right) = log_2\left(\frac{50}{\frac{60}{2{,}000} * \frac{300}{2{,}000}}\right) = 13.44$$

$$PMI(futon,\ closet) = log_2\left(\frac{P(futon,\ closet)}{P(futon) * P(closet)}\right) = log_2\left(\frac{25}{\frac{60}{2{,}000} * \frac{80}{2{,}000}}\right) = 14.347$$

$$PMI(futon,\ set) = log_2\left(\frac{P(futon,\ set)}{P(futon) * P(set)}\right) = log_2\left(\frac{60}{\frac{60}{2{,}000} * \frac{900}{2{,}000}}\right) = 12.118$$

With the PMI scores, we can calculate the semantic drift using eq. 2.

$$drift(futon, 2, 3) = \frac{AvgSim([chair, couch], futon)}{AvgSim([board, closet, set], futon)} = \frac{\frac{13.703+14.703}{2}}{\frac{13.44+14.347+12.118}{3}} = \boxed{1.0678}$$

(b) We can compute the *semantic drift* score for "hammock" as we did for "futon" is part a. This time we will consider the first 4 words and the last 2 words.

$$PMI(hammock,\ chair) = log_2\left(\frac{P(hammock,\ chair)}{P(hammock) * P(chair)}\right) = log_2\left(\frac{30}{\frac{10}{2{,}000} * \frac{200}{2{,}000}}\right) = 15.873$$
$$PMI(hammock,\ couch) = log_2\left(\frac{P(hammock,\ couch)}{P(hammock) * P(couch)}\right) = log_2\left(\frac{10}{\frac{10}{2{,}000} * \frac{50}{2{,}000}}\right) = 16.288$$
$$PMI(hammock,\ sofa) = log_2\left(\frac{P(hammock,\ sofa)}{P(hammock) * P(sofa)}\right) = log_2\left(\frac{8}{\frac{10}{2{,}000} * \frac{40}{2{,}000}}\right) = 16.288$$
$$PMI(hammock,\ bed) = log_2\left(\frac{P(hammock,\ bed)}{P(hammock) * P(bed)}\right) = log_2\left(\frac{34}{\frac{10}{2{,}000} * \frac{100}{2{,}000}}\right) = 17.053$$
$$PMI(hammock,\ closet) = log_2\left(\frac{P(hammock,\ closet)}{P(hammock) * P(closet)}\right) = log_2\left(\frac{15}{\frac{10}{2{,}000} * \frac{80}{2{,}000}}\right) = 16.195$$
$$PMI(hammock,\ set) = log_2\left(\frac{P(hammock,\ set)}{P(hammock) * P(set)}\right) = log_2\left(\frac{30}{\frac{10}{2{,}000} * \frac{900}{2{,}000}}\right) = 13.703$$

With the PMI scores, we can calculate the semantic drift using eq. 2.

$$\begin{aligned} drift(futon, 4, 2) &= \frac{AvgSim([chair, couch, sofa, bed], hammock)}{AvgSim([closet, set], hammock)} \\ &= \frac{\frac{15.873+16.288+16.288+17.053}{4}}{\frac{16.195+13.703}{2}} = \boxed{1.0954} \end{aligned}$$

(c) I think this similarity metric would be a **poor choice** for detecting semantic drift. A similarity metric that computes Jaccard distance on a vector of POS statistics would not express the semantic similarity between words in the lexicon and a candidate word.

For example, consider the semantic category FURNITURE we are working with and the 10 initial words in the lexicon. Most of these words would have a POS vector with a very high probability of being a NOUN. Lets take two new candidate words that would also have a very high probability of being a NOUN: *recliner* and *onion*. Both would score a similar semantic drift score under Jaccard distance but *onion* obviously drifts dramatically more than *recliner*.

Problem 2: Snowball Patterns

$$\boxed{Match(T_p, T_s) = \begin{cases} L_p \cdot L_s + M_p \cdot M_s + R_p \cdot R_s & if\ the\ tags\ match \\ 0 & otherwise \end{cases}} \tag{3}$$

(a) We use eq. 3 to compute the degree of similarity between P_1 and P_2. The tags match on P_1 and P_2, so we use case 1 of the piecewise function.

$$
\begin{aligned}
Match(P_1,\ P_2) &= L_1 \cdot L_2 + M_1 \cdot M_2 + R_1 \cdot R_2 \\
Match(P_1,\ P_2) &= ((3*5)+(4*2)+(1*2))+ \\
&\quad ((8*0)+(9*1)+(0*9))+ \\
&\quad ((5*1)+(0*7)+(6*4)) \\
Match(P_1, P_2) &= \boxed{63}
\end{aligned}
$$

(b) The Tag1 for P_1 (LOC) does not match the Tag1 for P_3 (PER) so the degree of similarity is zero.

$$Match(P_1,\ P_3) = \boxed{0}$$

(c) Similar to part a, we use case 1 of the piecewise function.

$$
\begin{aligned}
Match(P_1,\ P_4) &= L_1 \cdot L_4 + M_1 \cdot M_4 + R_1 \cdot R_4 \\
Match(P_1,\ P_4) &= ((3*0)+(4*9)+(1*5))+ \\
&\quad ((8*3)+(9*2)+(0*6))+ \\
&\quad ((5*4)+(0*2)+(6*1)) \\
Match(P_1, P_4) &= \boxed{109}
\end{aligned}
$$

Problem 3: Hypernyms

(a)

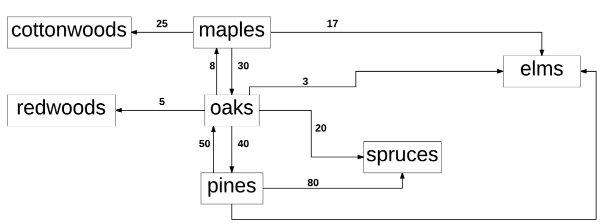

Figure 1: HPLG representing the web query table.

(b) The function $weight(u, v)$ returns the weight of a directed edge from node u to node v. If such an edge doesn't exist, 0 is returned.

$$
\begin{aligned}
Popularity(pines) &= \sum_{v \in V}^{|V|} weight(v, pines) \\
&= weight(oaks, pines) \\
&= \boxed{40}
\end{aligned}
$$

$$
\begin{aligned}
Popularity(oaks) &= \sum_{v \in V}^{|V|} weight(v, oaks) \\
&= weight(maples, oaks) + weight(pines, oaks) \\
&= 30 + 50 \\
&= \boxed{80}
\end{aligned}
$$

$$
\begin{aligned}
Popularity(spruces) &= \sum_{v \in V}^{|V|} weight(v, spruces) \\
&= weight(oaks, spruces) + weight(pines, spruces) \\
&= 20 + 80 \\
&= \boxed{100}
\end{aligned}
$$

(c)

$$
\begin{aligned}
Productivity(pines) &= \sum_{v \in V}^{|V|} weight(pines, v) \\
&= weight(pines, oaks) + weight(pines, spruces) + weight(pines, elms) \\
&= 50 + 80 + 60 \\
&= \boxed{190}
\end{aligned}
$$

$$\begin{aligned}
Productivity(oaks) = &\sum_{v \in V}^{|V|} weight(oaks, v) \\
=&weight(oaks, maples) + weight(oaks, pines)+ \\
&weight(oaks, elms) + weight(oaks, spruces) + weight(oaks, redwoods) \\
=&8 + 40 + 3 + 20 + 5 \\
=&\boxed{76}
\end{aligned}$$

$$\begin{aligned}
Productivity(spruces) = &\sum_{v \in V}^{|V|} weight(spruces, v) \\
= &\boxed{0}
\end{aligned}$$

(d) The Concept Positioning Test (CPT) is used to determine if a learned hypernym is more general than our Root Concept ”plants”. Given the two patterns:

(a) <Hypernym> such as plants and *

(b) plants such as <Hypernym> and *

We will consider a learned hypernym to be less general than ”plants” and pass the test if the following are true:

Pattern (b) produces at least 50 hits.

Pattern (b) returns at least 4 times as many hits as pattern (a).

(d.i) *ferns*: **would** - Ferns are a hyponym of plants so I would expect the pattern *plants such as ferns and ** to produce at least 50 hits and to appear much more often than *ferns such as plants and ** which sounds unnatural.

(d.ii) *things*: **would not** - I expect the pattern *plants such as things and ** to produce little to no hits as things are not a hyponym of plants. Additionally, *things such as plants and ** would produce substantially more hits as plants are a hyponym of things.

(d.iii) *vegetables*: **would not** - I don’t expect the pattern *plants such as vegetables and ** to yield many hits. The word vegetables is most commonly used to describe parts of a plant for consumption and sound unnatural in this pattern. Additionally, in biology the word vegetables is used to describe all plant matter so in this context it would not be a hyponym of plants.

(d.iv) *succulents*: **would** - Similar to ferns, succulents are a hyponym of plants so the phrase *plants such as succulents and ** sounds natural and would be common.

(d.v) *organisms*: **would not** - The root concept plants in a hyponym of organisms. Because of this I expect the test to fail as the phrase *plants such as organisms and ** is nonsense.

Problem 4: Identifying Monetary Amounts

(a) $P(O \mid E)$: **CAN** - This feature represents the probability of moving from state E to state O. This would be an entry in the transition probability matrix. This is a legal state transition as *other* could follow the *end* of a named entity.

(b) $P(IsNumber(w_i) \mid C)$: **CANNOT** - HMM's are a local generative model and cannot use arbitrary features. We would have to use a MEMM to use a richer feature such as this.

(c) $P(w_i \mid O)$: **CAN** - This is the state observation likelihood of the observation symbol w_i given the current state is O.

(d) $P(B \mid IsCapitalized(w_i))$: **CANNOT** - This is neither a transition probability or an emission probability and not allowed in the HMM model.

(e) $P(C \mid C)$: **CAN** - Represents the probability of moving from state C to state C and is an entry in the transition probability matrix. This would also be a legal state transition as *continue* would follow *continue* in a named entity that is at least four words long.

(f) $ContainsDollarSign(w_i)$: **CANNOT** - HMM's are a local generative model and cannot use global features such as this one. We could use a model such as a CRF or structured perceptron.

(g) $P(w_i \mid U)$: **CAN** - The state observation likelihood of the observation symbol w_i given the current state is U.