NAME:

CS-5340/6340, SOLUTIONS for Final Exam, Fall 2012

1. (12 pts) For each sentence below, name the conceptual dependency primitive ACT that would best represent the activity described in the sentence. (You do not need to draw a picture – just name the primitive ACT.)

   (a) The spider crawled up the wall.

   PTRANS

   (b) Tom smelled the roses in his garden.

   ATTEND

   (c) Charmaine solved the challenging riddle.

   MBUILD

   (d) Mustafa shared his secret with Amy.

   MTRANS

   (e) The movie star gave her parents a new house as a gift.

   ATRANS

   (f) Keith took cough medicine in the morning.

   INGEST

   (g) Charlie smiled broadly.

   MOVE

   (h) Eric often sings in the shower.

   SPEAK

(i) Bill began to drool from a tooth infection.

EXPEL

(j) Arun skied back to the lodge at sunset.

PTRANS

(k) Autumn read all the birthday greetings from her friends on Facebook.

MTRANS

(l) Jim felt the raindrops on his arm.

ATTEND

2. (23 pts) In each sentence below, circle each noun phrase and label it with the thematic role that it should be assigned with respect to the main verb. Choose from the following thematic roles: *agent, beneficiary, co-agent, co-theme, destination, instrument, recipient, source, theme*

   (a) The leaking faucet was fixed by Natalie with a wrench.

   The leaking faucet = Theme
   Natalie = Agent
   a wrench = Instrument

   (b) The waiter served the man a burnt steak.

   The waiter = Agent
   the man = Recipient
   a burnt steak = Theme

   (c) Barbara hiked to Idaho with her best friend.

   Barbara = Agent
   Idaho = Destination
   her best friend = Co-Agent

   (d) The old house was damaged by the hurricane.

   The old house = Theme
   the hurricane = Agent

   (e) Nancy organized a fund drive for her favorite charity.

   Nancy = Agent
   a fund drive = Theme
   her favorite charity = Beneficiary

   (f) Greg was kidnapped with his brother by three armed men.

   Greg = Theme
   his brother = Co-Theme
   three armed men = Agent

(g) Joe picked the lock with a credit card.

Joe = Agent
the lock = Theme
a credit card = Instrument

(h) The boy was injured by a punch from the school bully.
The boy = Theme
a punch = Instrument
the school bully = Agent

3. (8 pts) Suppose you are in the kitchen trying to get a cup that is on the top shelf of the cabinet, but you are not tall enough. Your very tall brother walks by, so you ask him: “Can you reach that cup?”

   (a) What is the locutionary act of this utterance?

   The act of saying the words “Can you reach that cup”.

   (b) What is the illocutionary act of this utterance?

   A request for your brother to get the cup for you.

   (c) What is the perlocutionary act of this utterance?

   That you will obtain the cup.

   (d) Is this utterance a *direct* or *indirect* speech act?

   It is an indirect speech act because the literal interpretation of the question is simply asking whether your brother is capable of reaching the cup. But the intent of the question is actually a request for your brother to get the cup for you.

4. (10 pts) The table below contains terms and their document frequency (DF) counts, where DF is the number of documents that mention the term in an imaginary text corpus that contains a total of 1,000 documents.

Term	analysis	chart	grammar	parsing	systems
DF	100	10	25	200	250

Consider the two documents D1 and D2 shown below:

> D1: Chart parsing is a common parsing technique used in natural language processing systems. Chart parsing uses a grammar and an agenda to perform syntactic analysis.
>
> D2: NLP systems use many techniques, including morphological analysis, syntactic analysis, semantic analysis, and discourse analysis. The user may need to define a grammar and decide on a parsing strategy, such as chart parsing.

Using the information above, compute the following TF-IDF values.

- TF-IDF("chart", D1)

  2 * log(1,000/10) = 2 * log(100)

- TF-IDF("chart", D2)

  1 * log(1,000/10) = log(100)

- TF-IDF("parsing", D1)

  3 * log(1,000/200) = 3 * log(5)

- TF-IDF("parsing", D2)

  2 * log(1,000/200) = 2 * log(5)

- TF-IDF("analysis", D2)

  4 * log(1,000/100) = 4 * log(10)

5. (7 pts) Consider the (tiny!) text corpus below, which contains 30 words:

Penguins are birds.
Most birds can fly.
Penguins can not fly.
Penguins eat fish.
Pelicans eat fish and can fly.
Penguins are cold weather birds.
Owls are common nocturnal birds.

Fill in the following co-occurrence matrix for the words: "birds", "fly", "pelicans", "penguins". Each cell should be filled with the number of times that the pair of words co-occur in the same sentence. A word only co-occurs with itself if it appears in the same sentence more than once. For example, "buffalo" co-occurs with itself once in"Utah buffalo buffalo", but does not co-occur with itself in "Utah buffalo snore".

	birds	**fly**	**pelicans**	**penguins**
birds	0	1	0	2
fly	1	0	1	1
pelicans	0	1	0	0
penguins	2	1	0	0

(a) Using the co-occurrence matrix, compute the similarity of "birds" and "penguins" with the Manhattan distance similarity metric.

Manhattan distance = 2 + 0 + 0 + 2 = 4

(b) Using the co-occurrence matrix, compute the similarity of "birds" and "penguins" with the Jaccard similarity metric.

Jaccard distance = (0 + 1 + 0 + 0) / (2 + 1 + 0 + 2) = 1/5

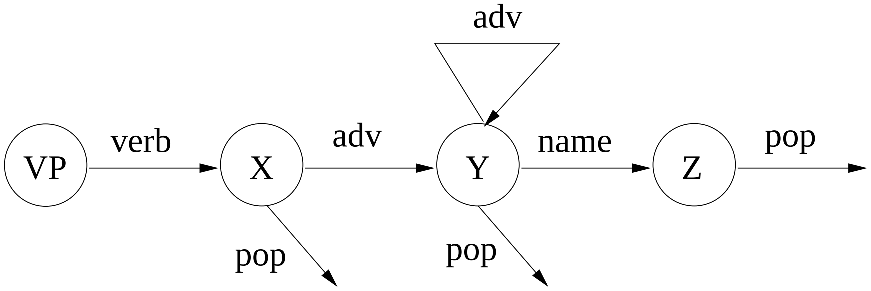
(c) Using the co-occurrence matrix, compute the similarity of "birds" and "penguins" with the cosine distance similarity metric.

cosine similarity = (0 + 1 + 0 + 0)/sqrt(0 + 1 + 0 + 4)*sqrt(4 + 1 + 0 + 0)
= 1 / sqrt(5)*sqrt(5) = 1/5

6. (10 pts) For each grammar below, draw a recursive transition network that recognizes exactly the same language as the grammar.

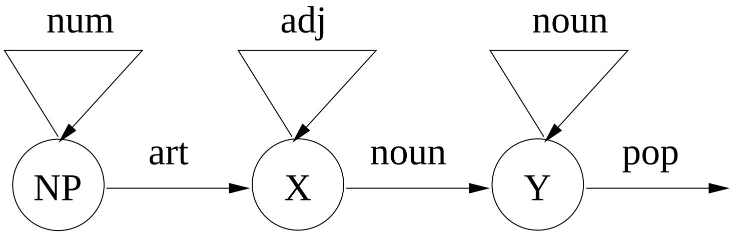
   (a) Grammar #1:

   VP → verb
   VP → verb VP1
   VP1 → adv
   VP1 → adv VP1
   VP1 → adv name



   (b) Grammar #2:

   NP → num NP
   NP → art NP1
   NP1 → adj NP1
   NP1 → noun
   NP1 → NP2
   NP2 → noun NP2
   NP2 → noun



7. (7 pts) For each natural language processing task described below, indicate whether the problem is best characterized as an *information extraction* problem, an *information retrieval* problem, or a *named entity recognition* problem. (Choose only one!) No explanation is necessary.

   - Find noun phrases that refer to specific countries.

     Named Entity Recognition

   - Find documents that discuss the history of Utah.

     Information Retrieval

   - Find noun phrases that refer to the victim of a shooting.

     Information Extraction

   - Find noun phrases that refer to a sports team.

     Named Entity Recognition

   - Find documents about Oprah Winfrey.

     Information Retrieval

   - Find noun phrases that refer to airplanes that crashed.

     Information Extraction

   - Find noun phrases that refer to buildings that were destroyed in a hurricane.

     Information Extraction

8. (6 pts) Suppose a machine translation system (MT) is trying to choose between the following two English sentences as a translation for a Spanish sentence about a dog barking:

(1) The dog barks a lot.
(2) A dog barks many times.

(a) Show the formula that the MT system would use to compute the probability of sentence (1) using a trigram language model. Be sure to instantiate the formula with the specific words in sentence (1).

$P(The \mid \phi_2\ \phi_1) * P(dog \mid \phi_1\ The) * P(barks \mid The\ dog) * P(a \mid dog\ barks) * P(lot \mid barks\ a)$

(b) Show the formula that the MT system would use to compute the probability of sentence (2) using a trigram language model. Be sure to instantiate the formula with the specific words in sentence (2).

$P(A \mid \phi_2\ \phi_1) * P(dog \mid \phi_1\ A) * P(barks \mid A\ dog) * P(many \mid dog\ barks) * P(times \mid barks\ many)$

9. (10 pts) Answer each question below as TRUE or FALSE. No explanation is necessary.

(a) Single-document question answering is easier than multiple-document question answering.

FALSE

(b) Named entity recognition is useful for factoid question answering.

TRUE

(c) Information retrieval is a component of most question answering systems.

TRUE

(d) If the Mean Reciprocal Rank (MRR) score for a question answering system is .50, then the question answering system ranks the correct answer among its top 5 hypotheses 50% of time.

FALSE

(e) If the Mean Reciprocal Rank (MRR) score for a question answering system is .75, then the question answering system often ranks the correct answer among its top 2 hypotheses.

TRUE

(f) Distributional similarity methods could be useful for query expansion.

TRUE

(g) Part-of-speech tagging is part of most statistical speech recognition models.

FALSE

(h) Part-of-speech tagging is part of most information retrieval systems.

FALSE

(i) Relevance feedback techniques use N-gram language models to improve information retrieval performance.

FALSE

(j) Statistical machine translation systems are trained with texts that have been manually translated into two or more languages.

TRUE

10. (7 pts) Consider the following part-of-speech tagged tongue twister:

> I/PRO thought/VB a/ART thought/NOUN but/CONJ the/ART thought/NOUN I/PRO thought/VB was/VB not/ADV the/ART thought/NOUN I/PRO thought/VB I/PRO thought/VB

Compute the following probabilities from the tongue twister. **Leave all your answers in fractional form!**

- $P(I)$

  4/17

- $P(NOUN)$

  3/17

- $P(thought \mid I)$

  4/4

- $P(thought \mid VB)$

  4/5

- $P(PRO \mid \phi)$

  1/1

- $P(VB \mid PRO)$

  4/4

- $P(VB \mid CONJ)$

  0/1

Question #11 is for CS-6340 students only!

11. (10 pts) Consider the following two stories:

> (1) A twister hit Kansas on Monday and it caused massive damage. Three people were killed and 20 people were injured. A tree fell on 3 of the injured men. A dog was also hit by flying debris and injured. A historic farm took a direct hit from the twister. The tornado occurred at midnight.
>
> (2) A group of children played twister on Monday. A bizarre accident occurred and 1 child was injured and another child was killed while they were playing the game. A large chandelier fell on them. The children often played in that room. The incident occurred at noon. 10 people have been killed while playing twister this year.

Assume that Story 1 is a *relevant* text, and Story 2 is an *irrelevant* text. For each of the information extraction patterns (p_i) below, compute $P(relevant \mid p_i)$. PassiveVP(verb) means the verb appears in a passive voice verb phrase construction. ActiveVP(verb) means that the verb appears in an active voice verb phrase construction.

(a) <subject> PassiveVP(killed)

1/3

(b) <subject> ActiveVP(hit)

1/1

(c) <subject> PassiveVP(injured)

2/3

(d) ActiveVP(occurred) at <np>

1/2

(e) ActiveVP(played) <direct-object>

0/1