CS 124 Winter 2020 Practice Final Exam

Instructions:

- The main exam consists of **38 questions**, worth a total of **54 points**.
  You must record your answers on the separately provided bubble sheet.
  You must turn in both the exam and the bubble sheet.
- **Extra Credit:** The exam also consists of **4 points** of extra credit questions at the very end.
- You have **3 hours** to complete this exam.
- In the exam you may use a computer on which anything can be downloaded, but you may use the web only to view the `cs124.stanford.edu` website for our class, to view the edX website for our class, and to view the Piazza website for our class. You may not use other information on the web.
- You are not allowed to write programs to check answers during the exam.
- You may use the calculator functions on your computer to compute values for functions such as cosine.

Regular Expressions and Edit Distance (6 points)

1. (1 point) What is the meaning of the regular expression character '+'?

   (a) Matches one or more of the preceding token

   (b) Matches exactly one of the preceding token

   (c) Matches zero or more of the preceding token

   (d) Matches zero or one of the preceding token

   A - this is just the definition of the Kleene plus +

2. (1 point) Given this list of spellings for the Jewish holiday:

   Hanukkah
   Chanukah
   Hanukah
   Hannukah
   Chanuka
   Chanukkah
   Hanuka
   Channukah

   Which of the following regular expressions successfully captures all of the variants?

   (a) `C?hann?uk?ah?`

   (b) `(Ch|H)an+uk+ah+`

   (c) `(Ch|H)an*uk*ah*`

   (d) `[CH]h?an?uk?ah?`

   The answer is C. A does not capture variants starting with capital H. B only captures variants that end with at least one h. D does not capture variants with double n's or double k's.

3. (2 points) Which of the following regular expressions captures the group `lists.stanford.edu` from the text `cs124-staff-win1819@lists.stanford.edu` and nothing else?

   (a) `(?:\w+-)*\w*\s*(?:@|at)\s*((?:\w+\.)+edu)`

   (b) `(?:\w+-)*\w*\s*(?:@|at)\s*(?:(?:\w+\.)+edu)`

   (c) `((?:\w*[.-])+\w+)`

   (d) `((?:\w+[\.-])+\w+)`

   The solution is A. B doesn't have any capturing groups at all. C captures two groups. D captures two groups (D is exactly the same as C).

4. (2 points) Suppose we weighted the edited distance so that the cost of each substitution was inversely proportional to the frequency that the mistake was made (see confusion matrix from lecture slides below). Which of the following has the closest edit distance to the word: “real”

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

(a) read

(b) reel

(c) heal

(d) raal

The answer is B.
We assume that the substitutions in the answer choices are X and the letters in "real" are Y. Then we get that the sub of d for l is 3, sub of e for a is 388, sub of h for r is 3, sub of a for e is 342, so since we are looking for inverse proportional, the largest frequency/smallest substitution cost is B.
If you read the table the opposite way, then you would get d for l is 4, sub of e for a is 342, sub of h for r is 8, sub of a for e is 388, so since we are looking for inverse proportional, the largest frequency/smallest substitution cost is D.

Language Modeling and Naive Bayes (6 points)

5. (1 point) We are interested in building a language model over a four-word corpus with the words - A, B, C, D. Consider the following training corpus: AABCDDABDACDBBA. Train a **bigram** language model on the above, including a <start> and <end> token when needed. What is $P(\text{B}|\text{A})$?

   (a) 0
   (b) 2/5
   (c) 1/5
   (d) 3/10

   The answer is B. To calculate P(B | A), we need (occurrences of AB) / (of occurrences of A) = 2 / 5 = 0.4.

6. (1 point) You evaluate your Naive Bayes classifier on a dataset with 100 positive $(y = 1)$ examples and 50 negative $(y = 0)$ examples. It correctly classifies 70 examples as positive and classifies the remaining 80 examples as negative. What is the precision?

   (a) 70/70
   (b) 70/100
   (c) 70/150
   (d) 30/100

   The answer is A. How many of the selected (positive) items are relevant (positive)

7. (2 points) Given the following confusion matrices, calculate the macroaveraged recall.

Class 1	True: yes	True: no
Classifier: yes	87	30
Classifier: no	88	295

Class 2	True: yes	True: no
Classifier: yes	10	25
Classifier: no	30	435

Class 3	True: yes	True: no
Classifier: yes	115	75
Classifier: no	50	260

   (a) 0.54
   (b) 0.48
   (c) 0.36
   (d) 0.14

   The answer is B. Calculate the recall for each matrix, sum them up, then divide by the number of matrices

8. (2 points) Let's use Naive Bayes for language identification!

   Currently, we have the following corpus of English and German documents that are labeled with their respective language. Unfortunately, the corpus was already converted to all lowercase, which makes this particular task even harder!

   das kind liebt pizza **(German)**
   das baby sagt mama **(German)**
   ein kind und ein baby essen pizza **(German)**
   that baby loves mama **(English)**
   mama is kind **(English)**

   Using Naive Bayes, determine whether the following sentence is most likely to be classified as English or German, and calculate the probability that Naive Bayes assigns to that most likely class. For this problem, use Laplace (add-one) smoothing, and ignore any words that do not appear in the corpus.

   $s =$ the kind baby loves pizza

   (a) English, with $P(s|\text{English}) = 5.00 \times 10^{-5}$
   (b) English, with $P(s|\text{English}) = 2.92 \times 10^{-5}$
   (c) German, with $P(s|\text{German}) = 8.64 \times 10^{-6}$
   (d) German, with $P(s|\text{German}) = 2.64 \times 10^{-5}$

   D. There are 15 German tokens and 7 English tokens. The German prior is 0.6 and the English prior is 0.4. The vocab size is 13. Thus: $p(s|English) = .4 * (.1)^3 * .05 = 2.00 * 10^-5$ and $p(s|German) = .6 * (1/28) * (3/28)^3 = 2.64 * 10^-5$.

Logistic Regression and Sentiment Analysis (4 points)

9. (2 points) Which of the following is true about logistic regression? **Select all that apply.**

   (a) It is a discriminative classifier

   (b) It is a generative classifier

   (c) It is a building block in a neural network

   (d) It is good for classification tasks

   A, C, D.
   (a): Logistic regression is a discriminative classifier, not a generative classifier. An example of a generative classifier would be Naive Bayes.
   (c): You can think of logistic regression as a one layer neural network.
   (d): As covered in class Logistic Regression performs well on classification tasks.

10. (1 point) Consider the statement: "*The holiday party was terrific! I was pleased with the atmosphere and food!*" What would be the best valence score / arousal score / ekman emotion for the statement above?

    (a) 7 / 2 / happiness

    (b) 6 / 2 / surprise

    (c) 7 / 6 / happiness

    (d) 2 / 6 / surprise

    C. Explanation:
    Valence: higher = more pleasant. We want something high.
    Arousal: higher = more intensity of emotion. Also want something high.
    Emotion: Happiness is more fitting.

11. (1 point) Suppose you built a rule-based sentiment analysis algorithm to classify a book review as positive or negative. The algorithm follows these steps in order:

    - For every word in the sentence, assign a score of $+1$ if it occurs in the positive lexicon, -1 if it occurs in the negative lexicon, and 0 otherwise.
    - If the immediately preceding word is "**very**" or "**really**", multiply the word's score by 2.
    - If you see the word "**not**", flip the sign of all subsequent words in the review.
    - Add the scores for each word in the sentence. If it is strictly greater than 0, the sentence is positive; otherwise, it is negative.

    Positive lexicon: interesting, great, cool

    Negative lexicon: stupid, boring, bad

    What score and class does your algorithm output for the following sentence?

It has cool characters and an interesting setting, but the plot was very stupid and not great.

(a) Negative, -1

(b) Negative, 0

(c) Positive, 1

(d) Positive, 2

A. At the end of the algorithm, *cool* and *interesting* both have scores of 1. *Stupid* has a score of -2, and *great* has a score of -1. That sums to -1.

Information Retrieval (10 points)

12. (1 point) Suppose we have the following three documents and we use them to create an inverted index.

    I. new and improved house robot
    II. robot breaks in to pet cat
    III. studies show cat improved owner happiness

    Which of the following should be in the index?

    (a) cat $\rightarrow$ 1 $\rightarrow$ 2

    (b) house $\rightarrow$ 1 $\rightarrow$ 2

    (c) robot $\rightarrow$ 1 $\rightarrow$ 2 $\rightarrow$ 3

    (d) improved $\rightarrow$ 1 $\rightarrow$ 3

    D.
    (a) is incorrect because "cat" doesn't show up in Document 1, but it does show up in Document 3 (which is also not reflected).
    (b) is incorrect because "house" doesn't show up in Document 2.
    (c) is incorrect because "robot" doesn't show up in Document 3.
    (d) is correct, as "improved" shows up in both Document 1 and Document 3 but not in Document 2.

13. (2 points) Say we have an inverted index for 600 documents with the following document frequencies for each term:

term	document frequency
language	200
robot	350
caramel	75
apple	450
puppy	550

    For the query — *NOT caramel AND NOT robot AND puppy AND (apple OR language)* — which of the following terms would we process first?

    (a) puppy AND (apple OR language)

    (b) NOT caramel AND (apple OR language)

    (c) NOT robot AND puppy

    (d) NOT caramel AND puppy

    C. Remember that we estimate (x OR y) by adding |x| + |y|, (x AND y) by taking min(|x|, |y|), and (NOT x) by subtracting N - |x| where N is the total number of documents.

    (a) (apple OR language) is 450 + 200 = 650, and then taking the AND of this with puppy is min(550, 650) = 550.
    (b) As above, (apple OR language) is 650. NOT caramel is 600 - 75 = 525. Taking the AND of these

two is min(525, 650) = 525.
(c) NOT robot is 600 - 350 = 250. Taking the AND of this with puppy is min(250, 550) = 250.
(d) As above, NOT caramel is 525. Taking the AND of this with puppy is min(525, 550) = 525.

Since (c) results in the smallest size, we process it first.

14. (1 point) Why is it not enough for a search engine to only focus on maximizing recall? Choose the best answer.

    (a) It might only retrieve a single relevant document to get perfect recall.

    (b) It can't calculate an F1 score.

    (c) It might just retrieve all documents to get perfect recall.

    (d) It won't correctly weight the top returned results.

    C. Remember recall is the number of relevant documents that are actually retrieved. If we retrieve all documents, we are guaranteed to retrieve all relevant documents, which means we have a recall of 1.

15. (2 points) Suppose we know that the cosine similarity between the query: "stanford computer science" and some document D is s, using **ltc.lnn weighting**. Let the cosine similarity between the query "stanford computer science stanford computer science" and the same document D be p (also using ltc.lnn weighting). How do s and p compare?

    (a) $p = s$

    (b) $p > s$

    (c) $p < s$

    (d) Insufficient information in the problem to compare p and s

    B. The document vector stays exactly the same (using LTC weighting). Since the query vector uses LNN weighting, we are effectively just making a vector of raw term frequencies for each term. Since the term frequencies of the longer query is double that of the shorter query, the cosine similarity (which is just the dot product between the query vector and document vector) for the longer query will be larger.

Note: Use the following documents for the next three questions.

Doc 1: "*Stanford students bike around Stanford campus*"
Doc 2: "*Many students at Stanford are CS majors*"
Doc 3: "*CS 124 is a favorite for Stanford kids*"

16. (2 points) Build a positional inverted index for the above documents. Which of the following would appear in the positional inverted index? **Select all that apply.**

    (a) "Stanford" $\Rightarrow \{1 : [0, 4], 2 : [3]\}$
    (b) "students" $\Rightarrow \{1 : [1], 2 : [1], 3 : [6]\}$
    (c) "CS" $\Rightarrow \{2 : [5], 3 : [0]\}$
    (d) "bike" $\Rightarrow \{1 : [2]\}$

    C, D.
    (a) is missing that "Stanford also shows up at position 6 in Document 3.
    (b) incorrectly has "students" appearing in Document 3.

17. (1 point) Which of the following is the result of a **boolean retrieval** for the query "*Stanford students*"? (Boolean retrieval is the "AND" of the words in the query.)

    (a) Doc1, Doc 2
    (b) Doc1, Doc 2, Doc 3
    (c) Doc 1
    (d) Doc 2, Doc 3

    A. "Stanford" shows up in all three documents, but "students" only shows up in Documents 1 and 2. Therefore, boolean retrieval will only return Documents 1 and 2.

18. (1 point) What would be the results of the **phrase retrieval** for the query "*Stanford students*"?

    (a) Doc 1
    (b) Doc 1, Doc 2
    (c) Doc 2
    (d) Doc 1, Doc 2, Doc 3

    A. The entire phrase "Stanford students" only shows up in Document 1; in Document 2, although both words of the query appear, they are not in the same (and consecutive) order as the query.

Relation Extraction and Vector Semantics (10 points)

19. (1 point) Suppose we run the Dipre algorithm on the following text instances starting with the following seeds. For this question, the Dipre algorithm will extract any patterns that have at least a single occurrence.

Seeds:
(Picasso, artist)
(Rowling, author)

Text:
The artist Picasso is famous for his works in cubism
Rowling is the author of the Harry Potter series
Picasso, an artist, was born in Spain

Which of the following patterns will we have extracted after one iteration? Assume that the first entry in the seed tuple is replaced by $?x$ and the second entry is replaced by $?y$ with words and punctuation marks separated by spaces. Choose a single answer.

(a) $?x$, a $?y$
(b) The $?x$ $?y$ is famous for his works in cubism
(c) $?x$ is the $?y$ $.*$
(d) $?x$ is the $?y$ of the Harry Potter series

D.
(a) is incorrect because using the seed (Picasso, artist), we can use the third text to extract "?x, an ?y". Note that this pattern uses "an", not "a".
(b) has ?x and ?y in the wrong order; if switched, it would be correct (extracted from the first text).
(c) is incorrect because the Dipre algorithm does not have the capability to infer arbitrary regular expressions. (d) is correctly extracted using the seed (Rowling, author) and the second text.

20. (1 point) Which of the following is false about hand-building patterns for relation extraction?

(a) Having humans hand-build patterns requires a lot of time and effort
(b) Hand-built patterns are often low-precision
(c) Hand-building allows for domain-specific patterns
(d) Hand-built patterns can be used to gather seed tuples for bootstrapping algorithms

B. Hand-build patterns are often high-precision, but low-recall; they work very well at very specific situations.

21. (2 points) Which of the following relations can be extracted from the text below (assuming you have a perfect relation extractor) using the ACE set of relations. **Select all that apply.**

Alphabet Inc. is the parent company of Google. One of the founders of Google Larry Page serves as CEO of Alphabet.

(a) PER-GPE

(b) ORG-ORG

(c) PER-ORG

(d) PER-PER

B, C.
Remember that ORG is organization, GPE is geopolitical entity, and PER is person. We learn the relation ORG-ORG (Alphabet Inc., Google) from the first sentence and the relation PER-ORG (Larry Page, Google) from the second sentence.

Note: Use the following information for the next four questions.

Imagine that you are trying to extract LANGUAGE-OF-COUNTRY(X, Y) relationships, where X represents the country and Y represents the language. Now we are given the following (nonsensical) document, which we use to both obtain features and extract relations:

United States, English. People in China speak Chinese. Classes, students. France, French. The language of Spain is Spanish. People in class speak often. People in United States speak English. Natural, language.

From the above passage, we can see that the **gold set of relations** is: *(France, French), (United States, English), (China, Chinese), (Spanish, Spain)*

Assume we first run the **bootstrap algorithm** and perform **supervised learning** on our expanded feature set to extract the relations. We begin with the seed **(France, French)**.

22. (2 points) Using this bootstrap + supervised learning method, which of the gold relations do we extract? **Select all that apply.**

    (a) (France, French)

    (b) (United States, English)

    (c) (China, Chinese)

    (d) (Spanish, Spain)

    A, B, C.

    Explanation: Using the bootstrap + supervised learning method, we obtain the following new patterns:

    X, Y (from "France, French")
    X speak Y (from "... United States speak English")

    With these patterns, we extract the following gold relations:
    (France, French)
    (United States, English)
    (China, Chinese)

23. (1 point) Using this bootstrap + supervised learning method, how many "false positive" (ie, tuples that aren't part of the gold set) relations do we extract?

    (a) 0

    (b) 1

    (c) 2

    (d) 3

D.

Explanation: Using the relation patterns extracted in the previous problem, we also see that the following 3 non-gold relations get extracted:
(classes, students)
(class, often)
(natural, language)

24. (1 point) Which is **higher**, the precision or the recall?

    (a) Precision
    (b) Recall

    B. Recall is larger:

    - $R = \frac{\text{\# correctly extracted relations}}{\text{\# gold relations}} = \frac{3}{4}$
    - $P = \frac{\text{\# correctly extracted relations}}{\text{total \# extracted relations}} = \frac{3}{6} = \frac{1}{2}$

25. (2 points) What is the F1 score?

    (a) 0.60
    (b) 0.63
    (c) 0.67
    (d) 0.75

    A. Explanation: $F1 = 2PR/(P+R) = 2(\frac{3}{4})(\frac{1}{2})/(\frac{3}{4}+\frac{1}{2}) = 0.6$

Question Answering and Chatbots (6 points)

26. (1 point) True or False: The following dialogue is an example of implicit grounding:

    Chatbot: *What would you like to order?*
    Person: *I would like a hamburger with fries please.*
    Chatbot: *Okay, so you want a regular hamburger with French fries?*

    (a) True

    (b) False

    False. The chatbot responds by asking the user to confirm whether the chatbot is correct, which is explicit grounding.

27. (1 point) True or False: For all sets of questions and corresponding ranked answers, the following statement holds: mean reciprocal rank $\geq$ accuracy

    (a) True

    (b) False

    True. Accuracy is 1 when the correct answer is ranked #1 and 0 when the correct answer is ranked lower or unranked. Mean reciprocal rank is 1 when the correct answer is ranked #1, $\frac{1}{2}$ when the correct answer is ranked #2, and so on, and it is equal to 0 when the answer is unranked. Thus, mean reciprocal rank will always be greater than or equal to the accuracy.

Note: Use the following information for the next three questions.

You have been tasked with building a specialized Information Retrieval-based (IR-based) question answering system about countries. All questions it receives will be answerable correctly with the name of a country. For example:
Q: "*In what country did the biological father of Steve Jobs grow up?*"
A: "*Syria*"

28. (1 point) Which of the following resources would be the most appropriate source of documents for a **purely IR-based** approach to this problem?

    (a) GeoNames, a geospatial database (contains knowledge in table format, where each row consists of the name of a place, the country it's in, its "feature class"/what type of place it is (ex: city, park), and its latitude and longitude)

    (b) WordNet, an ontology that groups English words into sets of synonyms

    (c) Wikipedia, a free online encyclopedia

    (d) The David Rumsey Map Collection digital archive, an image database containing photos of historical maps and drawings

    C.
    (a) is incorrect. This is a knowledge-based resource. It does not have information in the form of passages within documents, so an IR-based approach would have difficulty using it well.
    (b) is incorrect. This is a knowledge-based resource. It does not have information in the form of passages within documents, so an IR-based approach would have difficulty using it well.
    (c) is correct. The IR system's document-retrieval step could retrieve articles, and the passage retrieval step could extract useful excerpts from them.
    (d) is incorrect. IR-based approaches involve documents and passages, not images.

29. (1 point) In what order do the following steps in your system need to be performed, in order to compute answers for a question?

    I. Answer Processing
    II. Document Retrieval
    III. Passage Retrieval
    IV. Query Formulation

    (a) I, II, III, IV

    (b) IV, II, III, I

    (c) II, III, I, IV

    (d) IV, III, II, I

    B. Answer processing requires passages to have been retrieved, these passages need to be retrieved from documents, and the query needs to be formulated for document retrieval to happen. So the ordering is Query Formulation → Document Retrieval → Passage Retrieval → Answer Processing.

30. (2 points) After building your IR-based question answering system, you give it to a contractor to evaluate on a set of queries. To your horror, they report your system had an accuracy of 0%! They also inform you the mean reciprocal rank was exactly 1/2 (0.5).

    You find these results hard to believe. . . but if they are true, there is one clear bottleneck, one subsystem most worth improving. To achieve higher accuracy on the queries the system was evaluated on, what do you need to make your system better at?

    (a) Retrieving better passages

    (b) Extracting answer candidates from the passages

    (c) Ranking the candidate answers

    Hint: How often were the correct answers among the answer candidates?

    C. Accuracy 0% and MRR 1/2 means that the correct answer was surfaced as a candidate for every question, and it was ranked #2 every single time. Retrieving passages and extracting answer candidates from the passages already did their job, getting the correct answer into the candidates list every single time. All that's needed to improve accuracy on these queries is rank the correct answers as 1 instead of 2.

Recommender Systems (6 points)

31. (1 point) What is "The Long Tail" problem faced by modern recommendation systems?

    (a) The vast majority of items aren't very popular/have very few reviews

    (b) Shelf space is a scarce commodity for traditional retailers

    (c) There is a "long tail" of users who only shop online

    (d) Recommendation systems fail to recommend movies starring cats, raccoons, mice, and other animals with long tails.

    The correct answer is (a) by definition. (b) is closely related, as it is a potential cause of the long tail problem, but is not the definition of the problem itself. (c) is incorrect, as the long tail problem is about items, not users. (d) is untrue in general, which is good news if you're a fan of Cats.

32. (1 point) Which of the following are challenges for content-based filtering? **Select all that apply.**

    (a) Picking features: it's hard to identify and extract what would be good features for the item vector.

    (b) Cold start: without many users in the system, it's hard to form recommendations.

    (c) First rater: it's hard to recommend an item that has never been rated before.

    (d) Overspecialization: it is hard to recommend items that are outside of the user's previous preferences.

    The correct answers are (a) and (d). (b) does not apply to content-based recommender systems, as even with only a single user in the system, as long as we have their user profile and the item profiles for all of the items, we can make recommendations. (c) is not correct, as the item profiles do not depend on user ratings, so even a completely new item that has never been rated before will have an item profile and can be recommended.

33. (2 points) Imagine that we added an extension to the Chatbot assignment to evaluate the chatbot's predictions by asking the user to watch the movies that were recommended and provide her true rating of the movie.

    The following table holds the predicted ratings for the various movies for User A.

Movie	Predicted Rating (on a scale of 1 - 10)
A Star is Born	8
Bohemian Rhapsody	9
Avengers: Infinity War	4

    After watching the movies, User A provides their ratings of the movies:

Movie	Actual Rating (on a scale of 1 - 10)
A Star is Born	9
Bohemian Rhapsody	7
Avengers: Infinity War	8

What is the root-mean-square error of the given recommendations?

(a) 2.645

(b) 5.686

(c) 7.000

(d) 4.58

The correct answer is (a). It can be calculated using the formula $\sqrt{\frac{(8-9)^2+(9-7)^2+(4-8)^2}{3}} = 2.645$. (b) is the answer you get if you take the squares before taking the difference. (c) is the answer if you forget to take the square root. (d) is incorrect.

34. (2 points) As Spotify's newest recommendation engineer, you have been tasked with overcoming the cold start effect for songs from newly-released albums by established artists. Which of the following approaches helps to overcome the cold-start effect?

I. Use user-user collaborative filtering to find users whose preferences are similar to the music-listening preferences of the artist, and recommend the new album to those users.

II. For each listener, count the percentage of listens that have been to this artist previously. If the percentage is above a threshold, recommend songs from the new album.

III. Extract features from the newly-released songs using machine learning and use content-based filtering to recommend the new song to listeners who like similar songs.

IV. Use item-item collaborative filtering on the artists, recommending the new songs from this album to users who previously enjoyed (listened to) similar artists (including this one!).

(a) II and III only.

(b) II, III, and IV.

(c) I, III, and IV.

(d) II and IV.

The correct answer is (b). Not I, since the music the artist listens to may not be representative of music they produce, and represents much less information than the listening habits of listeners to this artist, particularly if the artist is well-established. II would be a reasonable non-recommendation system approach that might be beneficial. III is just a good idea. IV is a valid approach, treating the artists as the items that are recommended in a recommendation system.

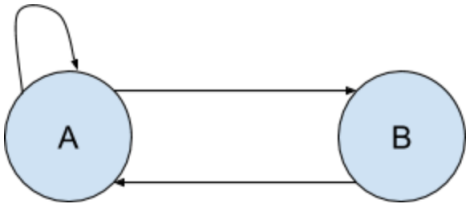
Networks (6 points)

35. (1 point) Anchor text is helpful in all of the following ways EXCEPT:

    (a) Anchor text gives user a cue of what the link is pointing at
    (b) Anchor text tells us the authority of the anchor page's website
    (c) Anchor text can be used for indexing a document
    (d) Anchor text can be used to find translations of items in different languages

    B. A is wrong since anchor text such as including "ibm" can help tell us what the link is pointing at ibm's website. C is wrong because anchor text is used for indexing a document, such as seeing anchor text "big blue" and "ibm" when indexing www.ibm.com. D is also wrong, for instance, if I looked at anchor text on us.gov that said "Germany" and that took me to the "Deutschland" government website (Germany is Deutschland in German). B is our answer by process of elimination.

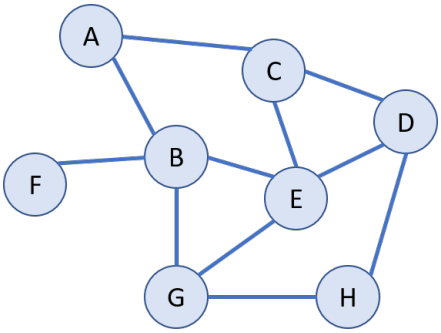
36. (1 point) Suppose we take a random walk with no teleporting on the following Markov chain. All of the following are valid random walks EXCEPT: (valid walks denotes as a sequence of states visited)



    (a) A, A, A, A, A
    (b) A, B, A, B, A
    (c) A, B, A, B, B
    (d) A, B, A, A, A

    C. Since we aren't teleporting, we can't walk from B to B. There is no edge from B to B. Therefore C is our answer.

37. (2 points) Suppose we have the following graph:



    What is the clustering coefficient of node E?

(a) 4/11

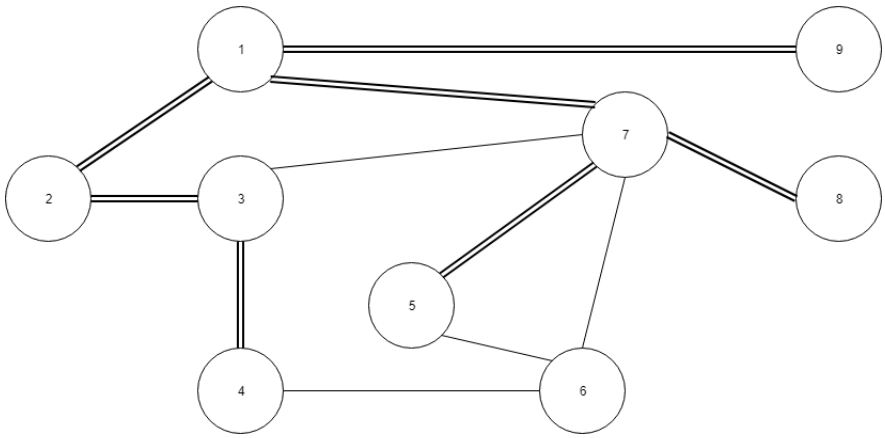
(b) 2/6

(c) 2/4

(d) 6/10

The answer is B. By definition of clustering coefficient" E is connected to B, C, D and G. BG, and CD is connected.

38. (2 points) Consider the following graph, with bold edges representing strong links and light edges representing weak links. How many additional edges would we expect to exist in the graph if the strong triadic closure property were to be satisfied?



(a) 5

(b) 6

(c) 7

(d) 8

The answer is D. One simply needs to draw out all possible edges which don't exist between two distinct nodes A and B where A and B are both strongly linked to a different node C.

Extra Credit - Two-Sided Recommendation Systems (4 points)

Please note that this problem is purely extra credit and is harder than the problems on the main final exam. We advise you to complete the main final before attempting this problem.

In this course, we've seen recommendation systems used in 1-sided markets: Amazon items do not have a preference for who purchases them and movies do not care who watches them. However, many real-world markets are 2-sided: in the monogamous dating market, the preferences of both sides of a prospective recommendation affect the quality of the recommendation; in the labor market, employers care about the quality of their employees and employees care about the quality of their employers; and in the residency-matching market, prospective doctors have ratings over hospitals and hospitals have ratings of candidates.

We will design a 2-sided recommendation system for the residency-matching market. Here's the setup: We have a set $\mathcal{D}$ of doctors and a set $\mathcal{H}$ of hospitals. We observe some, but not all, of the ratings of hospitals by doctors and the ratings of doctors by hospitals. Our goal is to determine system-wide "good matches" and recommend these matches to the participants.

We will define the quality of a match MN, where $M \in \mathcal{D}$ and $N \in \mathcal{H}$, to be the product of M's predicted or actual rating for N with N's predicted or actual rating for M. For example, if $M \in \mathcal{D}$ has rated $N \in \mathcal{H}$ as 6, and we predict (somehow) that N's rating for M would be 5.5, then the recommendation system would place a value of 33 on the MN pair.

As a concrete example, suppose we have four doctors A, B, C, D $\in \mathcal{D}$ and four hospitals W, X, Y, Z $\in \mathcal{H}$, and two partially observed ratings matrices, $\mathcal{M}$ (which represents candidates' ratings of hospitals) and $\mathcal{M}'$ (which represents hospitals' ratings of candidates).

$\mathcal{M}$	A	B	C	D
W	1	3	7	
X	4	2		6
Y	8	9	5	
Z		6	3	9

$\mathcal{M}'$	W	X	Y	Z
A	1	3	3	1
B	4		8	
C	2	3		2
D			4	

From these tables, we see that candidate C gave hospital Y a rating of 5 and hospital Y gave candidate B a rating of 8. Hospital Y did not provide a rating for candidate C. The pair BY is given a value of 72 by the system, since B rated Y as 9 and Y rated B as 8.

39. (2 points) Using item-item collaborative filtering, mean-centering columns (for both $\mathcal{M}$ and $\mathcal{M}'$), taking the weighted average (i.e. normalizing by the sum of the similarities), only using positive-similarity neighbors in the weighted average, determine the top two matches of highest value recommended by the system. Which of the following pairs are in the top two? **Select all that apply.**

    (a) BY

    (b) DY

    (c) CY

    (d) CW

    A and B (BY, DY). BY has value 72, DY has predicted value 36, CY has predicted value 15, and CW has value 14. The next highest off this list is AY with 24.

40. (1 point) Which of the following are true statements about candidates who provide ratings over all hospitals compared to candidates who provide ratings for relatively fewer hospitals? **Select all that apply.**

    (a) If this system does not normalize by the sum of the similarities when computing the predicted ratings, candidates with frequent ratings are over-recommended by the system, meaning that matches with that candidate are more likely to be among the top matches.

    (b) New candidates with no rankings will never be part of a recommended match until they rank at least one hospital.

    (c) The candidate with the fewest rankings will never be a part of a recommended match until they rank more hospitals.

    (d) If a hospital's first choice candidate's first choice is that same hospital, then this match will be the highest-value recommendation by the system.

    A and B. How often candidates have rated hospitals affects the system's predictions for what they think of the hospitals they haven't rated yet.

    A: Correct - If you don't normalize by the sum of the similarities, then the more ratings are in the sum in the numerator (which no longer has a denominator!), and the higher the output - possibly surpassing all of their other, "real" ratings. So the more hospitals you rate, the higher it predicts your ratings of other hospitals to be, and the more matches you get.

    B: Correct - Item-item filtering can't predict ratings for new candidates with no ratings, unless it gives them a 0, which can't be part of the top matches (assuming there are other, nonzero ratings in the system), so new candidates who have rated no hospitals won't get recommended.

    C: Incorrect - She could still be part of a recommended match, if hospitals rank her super highly relative to the other candidates and she ranks hers highly as well.

    D: Incorrect - Say Candidate A is Hospital X's first choice and vice-versa, but Candidate A is more conservative with her ratings than another Candidate B. Candidate A gives Hospital X a rating of 7 (her highest), and X gives A a rating of 8 (its highest). Now say Candidate B gives Hospital X a rating of 10, and X gives B a rating of 7. Then BX is a higher value recommendation: $7 * 10 = 70$, which is greater than $7 * 8 = 56$ for AX.

41. (1 point) An airline is planning a promotion to provide candidates with free flights to visit hospitals with which they would be a good match. The airline will award the flights to the doctor-hospital pairs with the highest predicted match value.

    Under this system, a candidate can "game the system" by giving every hospital a rating of 10, in order to maximize the match value of the doctor-hospital matches they belong to.

    One way to prevent the system from disproportionately awarding flights to doctors who rated every hospital a 10 is to mean-center the _____ before computing the match value. (Fill in the blank.)

    (a) rows of $\mathcal{M}$ (the doctors-rating-hospitals matrix)

    (b) columns of $\mathcal{M}$

    (c) rows of $\mathcal{M}'$ (the hospitals-rating-doctors matrix)

    (d) columns of $\mathcal{M}'$

    B. Mean-centering the columns of $\mathcal{M}$ would set the ratings by doctors who rate every hospital a 10 to 0, whereas doctors who had more varied ratings would have some made negative and others kept positive (and thus greater than 0). Mean-centering any of the other rows or columns would not address the problem.

— END OF EXAM —

— Scratch Space —