Indian Institute of Technology Kharagpur
Social Computing (CS60017)
End-Semester Examination, Autumn 2017-18

Full Marks: 60 **Time: 3 hours**

All parts of the same question must be answered together
Be precise in your answers, and state any assumptions made
If there are multiple ways to perform a computation, state which one you are using

Question 1 [3+6+6 = 15 marks]

(a) List and explain three possible sources of algorithmic bias.

(b) With respect to bias and discrimination, what is a protected group? What are the differences between individual and group discrimination? Why is proving discrimination hard?

(c) Using the contingency table below, define the following discrimination measures:
(i) Risk Difference, (ii) Odds Ratio and (iii) Extended Lift.

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

Question 2 [4+6+5 = 15 marks]

(a) If you were to develop an automated classifier to distinguish clickbaits and non-clickbaits, list at least four features that you could use for the classification.

(b) Apply the features you have listed against the above question, and identify from the following headlines, which ones are clickbait and why.

A. Advocates Who Want To Change Criminal Justice Are Rallying Around These Republican Governors
B. This Is What High Functioning Anxiety Looks Like
C. Answer 6 Questions About Food And We'll Tell You When You'll Get Married
D. This Community College Basketball Player Says He Was Kicked Off The Team For Not Taking Part In The National Anthem
E. Under Trump, The Pentagon Has Been Quietly Escalating Its Presence In Somalia
F. Only 10% of people can pass this brand pronunciation test

(c) Suppose a user has marked the following headlines as annoying and mentioned that she does not want to see headlines having similar pattern in future. Identify possible patterns to represent the headlines, and write the steps you used to arrive at the patterns.

A. Can You Guess These `Harry Potter' Characters?
B. Can You Play The `Shrek's Best Friend' Roles?
C. 18 Amazing Things Indians Do Not Anticipate
D. 11 Wrong Ways Moms Use The Tasty Cookbook
E. 27 Cool Projects Kids Will Want To Try

Hint: The Parts-Of-Speech tags of singular noun, plural noun, verb, adjective, determiner, preposition are NN, NNS, VB, JJ, DT, and IN respectively. You can use { and } to denote some tag being optionally included in a pattern. State any other assumption you make.

Question 3 [5+5 =10 marks]

Consider the following decision records of hiring decisions made in a company, based on the gender and qualification of the applicants.

Gender	Qualification	Hiring Decision
Male	Graduate	Yes
Female	Undergraduate	No
Female	Graduate	Yes
Male	Graduate	Yes
Male	Undergraduate	Yes
Female	Undergraduate	No
Female	Undergraduate	No
Female	Graduate	No
Male	Undergraduate	No
Male	Graduate	Yes

(a) Check and justify whether the following association rules are 2-discriminatory:

(i) "Female, Graduate -> No"

(ii) "Female, Undergraduate -> No".

(b) Using situation testing, check whether women are t-discriminated where t = 0.2. Assume the neighborhood size to be 2.

Question 4 [6 + 4 = 10 marks]

In a certain country, there are two political parties X and Y. In a search system that is popular in the country, three different algorithms A1, A2, A3 can be used. For a certain politics-related query q, the top five ranked results returned by the three algorithms are:

A1: r1, r5, r3, r4, r2

A2: r5, r4, r2, r1, r3

A3: r1, r5, r4, r2, r3

Here r1, r2, ..., r5 are five results that are relevant to query q. The left-most result is the top-ranked one in each ranked list.

Consider a binary bias score that is 0 for a result biased towards party X, and 1 for a result that is biased towards party Y. The bias score of the results are as follows: r1: 0, r2: 0, r3: 0, r4: 1, r5: 1.

0 (a) Compute a numeric bias score at rank 4, for the ranked lists output by the three algorithms. The score should account for bias at higher ranks, as opposed to bias at lower ranks.

✓ (b) It is claimed that the algorithms deployed by the search system are biased. Is the above information sufficient to verify this claim? If yes, say which of the three algorithms A1, A2, A3 are biased and justify your answer. If no, say what other information is necessary to verify the claim.

✓ Question 5 [5 marks]

The Twitter social media allows users to post an image with every textual tweet. A real-time image search system needs to be developed over Twitter, which, for a given topic (e.g., politics, sports, music), outputs a ranked list of popular images that are relevant to the topic. However, there is a problem - while some operations on images are efficient (e.g., deciding if two images are duplicates of each other is efficient and can be performed in real-time), there is no known efficient method to identify the topic of an image. Suggest how the real-time image search system can be constructed. You can assume that all tweets are available in real-time.

✓ Question 6 [5 marks]

Consider the sub-graphs marked A and C in the network shown below. Which of the sub-graphs is a better community, according to the measures (i) conductance, (ii) internal density? Show the calculations of the measures for the two sub-graphs.