

Sample Problem Set

COSC 4397: Natural Language Processing

Name: _____ UHID _____

Instructions:

1. This is a closed book examination
2. The paper has 22 questions, the full mark is 100, time allotted is 3 hours.
3. Answer briefly and to the point in the space provided respecting the points.
4. Show computation/reasoning for justification/derivation problems.
5. Ideally, you should not need extra space, but in case you need, please feel free to use 4 extra sheets at the end.

I. NLP Basics and Text Retrieval [20 points]

1: Consider the following mini corpus containing 4 documents [5+5+1+1 = 12 points]:

Doc-1: breakthrough drug for schizophrenia

Doc-2: new schizophrenia drug

Doc-3: new approach for treatment of schizophrenia

Doc-4: new hopes for schizophrenia patients

(a) Compute the binary term-document incidence matrix for this collection.

(b) Compute the inverted index for this mini corpus.

(c) What is the result of the query: schizophrenia AND drug ?

(d) What is the result of the query: for AND NOT (drug OR approach) ?

2. This question is based on statistical distribution of words in English. [1 + 1 + 2 + 2 = 6 points]

- (a) What is Zipf's law?
- (b) Mandelbrots' law? Provide formulae for them.
- (c) If f and r denote the frequency and rank of the terms/words in a corpus, and we plot $\log(f)$ on y-axis and $\log(r)$ on x-axis, then The relationship of y and x is roughly (choose one): (1) quadratic, (2) inverse, (3) linear, (4) None
- (d) The slope is roughly (choose all that apply): (1) positive (2) negative (3) undefined (4) constant

3. Briefly define the following: [1 + 1 = 2 points]

- (a) Stopwords
- (b) Content/Function words

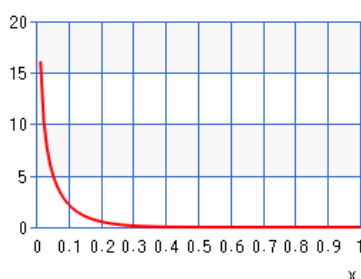
II. Mathematical Foundations for NLP [20 points]

4. Suppose 5% of men and 0.25% of women in some African tribe are color blind. A random person is chosen from the tribe and was examined to be color blind. What is the probability that the person is male? Assume males and females are equal in numbers. An exact value is not required. A numerical expression is sufficient. [6 points]
5. For some events A and B in a given sample space, we have, $P(A) = 1/3$ and $P(B^c) = 1/4$, where B^c denotes the complement of B , i.e, non-occurrence of B . Can the events A and B be disjoint? Justify your answer. [3 points]
6. Seven balls are distributed randomly into 7 baskets, i.e., each basket can get anywhere between 0 to 7 balls and the sum of all balls in each basket should be 7. We define the random variable: $X_i = \text{The number of baskets containing exactly } i \text{ balls}$
What are the possible values for X_3 ? [2 points]

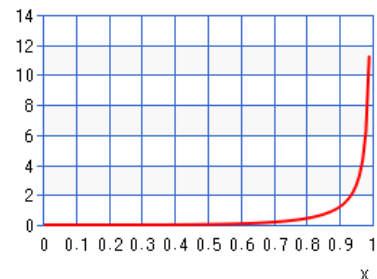
7. Recall that if $X \sim \text{Beta}(\alpha, \beta)$, then the mean value of the random variable is given by $E[X] = \frac{\alpha}{\alpha + \beta}$. Using this result which of the following density plots (PDF) on the y -axis for different values of x , best describes the distribution $\text{Beta}(\alpha = 5, \beta = 0.2)$? Provide a justification for your choice. [3 points].



Plot A



Plot B



Plot C

8. Express the probabilities of the following events in terms of $P(A)$, $P(B)$, and $P(A \cap B)$:
[2+2+2 = 6 points]

- (a) either A or B or both
- (b) either A or B but not both
- (c) at least one of A or B

III. Collocations, Hypothesis testing, N-gram Language Models [20 points]

9. Write brief answers to the following [1 + 1 + 1 + 1 + 1 + 1 = 6 points]:

- (a) What is a collocation (of consecutive words)?
- (b) How are they discovered?
- (c) Are all collocation informative?
- (d) What tends to make a collocation interesting/informative?
- (e) What is the main drawback of t-test?
- (f) How does Chi-squared (χ^2) get around it?

10. Answer the following based on the main idea of t-test [1+2 = 3 points].

- (a) Does higher t value in the t-test indicate lower confidence in rejecting the null hypothesis?
- (b) Arrange the p-values in the increasing order for three experiments A, B, C whose t values (obtained using a t-test) are as follows:
t-test value, for experiment A, $t_A = 2.156$
t-test value, for experiment B, $t_B = 1.656$
t-test value, for experiment C, $t_C = 3.556$
i.e., Arrange the p-values of the experiments, p_A , p_B , and p_C in the increasing order assuming each experiment is of a similar kind (more specifically, they and the associated t-tests have the same degrees of freedom).

11. Compute these unigram and bigram probabilities based on the POS-tagged text snippet below.

I/PRO am/VERB a/ART nobody/NOUN.

Nobody/NOUN is/VERB perfect/ADJ.

Therefore/ADV I/PRO am/VERB perfect/ADJ.

Show the numerator and denominator (e.g., 5/10) rather than just the resulting probability (e.g., 0.5) to ensure you get the counts correct! Provide your final answer in the blanks. [1 + 1 + 1 + 1 + 1 = 5 points]

(a) $P(\text{perfect}) =$ _____

(b) $P(\text{VERB}) =$ _____

(c) $P(\text{am} \mid \text{I}) =$ _____

(d) $P(\text{ADJ} \mid \text{VERB}) =$ _____

(e) $P(\text{VERB} \mid \text{ART}) =$ _____

12. (a) What is smoothing? (b) Why do we need it? (c) Mention a practical application [1 + 1 + 1 = 3 points]

13. Higher order n-gram suffer with: (choose all that apply) [3 points]

(a) Smoothing

(b) Data sparsity

(c) Usually have lower probability than lower order n-grams

(d) OOV words

(e) All the above

IV. Markov Models, POS Tagging, Grammar and Parsing [20 points]

14. Show the equation that would be used to compute the probability of $P(PRP\ NN\ RB\ VBZ\ NN|My\ dog\ also\ eats\ spinach)$ using a trigram POS tagger [4 points].

15. In context of POS tagging, [1 + 1 + 1 = 3 points]

(a) why is direct estimation of the most likely tags given an observed word sequence, i.e.,

$$\widehat{T_1 \dots T_n} = \operatorname{argmax}_{T_1 \dots T_n} P(T_1 \dots T_n | W_1 \dots W_n)$$
 not practical?

(b) What is the time complexity of direct estimation assuming a total of k tags?

(c) Which algorithm and technique is used to get around the problem?

16. Explain briefly the 3 main estimation problems in a Hidden Markov Model (HMM)? What technique is used for each? [1 + 1 + 1 = 3 points]

17. What is phrase chunking? How is it related to syntactic parsing and why is it called *shallow parsing*? [2 + 1 = 3 points]

18. Referring to parsing, answer the following:

- (a) Briefly state the difference between top-down and bottom-up parsing.
- (b) What is ambiguity in parsing?
- (c) For a Probabilistic Context Free Grammar (PCFG), how would you decide which parse tree is the most likely parse given multiple parse trees for a sentence? [3+ 2 + 2 = 7 points]

V. Text Categorization [20 points]

19. Given the classification results in the following confusion matrix, compute the classification *accuracy*, and the *precision*, *recall* and *F* score of the positive data. It is sufficient to provide the expression. [1 + 1 + 1 + 1 = 4 points]

Classified as		Actual/Correct class
Positive	Negative	
50	10	Positive
5	200	Negative

20. (a) Why do we say the decision function for an SVM classifier is linear? (b) Can we say the same for Naïve Bayes (NB) or Decision trees (DT)? Justify (c) Can we directly use continuous attributes/features (e.g., height in inches, $h \in [50, 84]$) in NB, SVM, DT? (d) How to deal with continuous attributes/features when we cannot use them directly? [1 + 2 + 2 + 1 = 6 points]

21. We know that Naïve Bayes' final decision is based using the probability function $P(c|X) = \underset{c}{\operatorname{argmax}} P(X|c)P(c)$ where c denotes the class variable and X the instance which needs to be classified. Answer the following: [2 + 3 + 2 = 7 points]
- (a) Is the decision function obtained using MAP (Maximum a-priori) estimation or MLE (Maximum likelihood estimation)? Justify your answer and identify the prior, likelihood and posterior.
 - (b) How would the Naïve Bayes classifier's final decision function differ if we were to use maximum likelihood estimation (MLE)? i.e., Assuming no class prior or uniform class prior (each class is equally likely).
 - (c) What makes Naïve Bayes "naïve"? Explain briefly

22. Briefly answer the following [1 + 1 + 1 = 3 points]
- (a) What is cross validation (CV)? Give one application of it.
 - (b) When does k -fold CV equal leave one-out CV for a dataset of n instances/examples?
 - (c) Which algorithm out of SVM, NB, and DT performs automatic feature selection?

[This is additional space provided in case you need more space.]
[Please provide question/problem number if you are answering any part here.]

[This is additional space provided in case you need more space.]
[Please provide question/problem number if you are answering any part here.]

[This is additional space provided in case you need more space.]
[Please provide question/problem number if you are answering any part here.]

[This is additional space provided in case you need more space.]
[Please provide question/problem number if you are answering any part here.]