PDK

NLP Quiz 1 Total Marks: 15 Time: 1 hour

NOTE: ANSWERS TO QUESTIONS IN PART A SHOULD BE CLEARLY SEPARATED FROM ANSWERS TO QUESTIONS IN PART B. START PART B ON A FRESH PAGE.

Be Precise, use bullet points rather than verbose text wherever appropriate.

PART A

1. In Laplace smoothing we add a fixed quantity to each count. In absolute discounting, we subtract a fixed quantity from each count. Given this similarity, why is it claimed that the latter is better than the former? [1]
2. In N-gram models we compute the conditional probability $p(w_i|w_{i-1})$. If we wanted to compute the joint probability $p(w_{i-1}w_i)$, how would you obtain the ML estimate? Assume the distribution is stationary. [1]
3. How would you treat a back-off model as a generalized interpolation model? [1]
4. In Good-Turing discounting, explain the need for building a predictor for the frequency of frequencies. [1]

PART B

5. Briefly outline an approach to *efficiently* identify candidate words from a dictionary that can be potential replacements for a misspelt word. (The motivation is to avoid costly operations such as edit distance computations over words that have very little overlap with the misspelt word.) [2]
6. Find the edit distance between "msdhoni" and "madonna" using Dynamic Programming, where the costs of insertion, deletion and substitution are 1, 1 and 2 respectively. Show clearly your table of sub-problems. [3]
7. What mathematical operation does a set of parallel Finite State Transducers realize? Why is this useful in the context of word morphology? [1]
8. How would you construct a distance metric out of K-L Divergence? [0.5]
9. Is it possible for K-L Divergence to be symmetric? If yes, how? If no, why? [0.5]
10. Give an example of an Information Theoretic measure of semantic relatedness between a given pair of words and show CLEARLY the steps in estimating it from a given corpus. Give attention to details. Mention limitations of the measure you used, if any. [2.5] Use Wordnet & assume the words have entries Word
11. Point out syntactic ambiguity (if any) in the sentence "The old man the boat", and show the parse tree(s). [1]
12. In estimating distributional similarity using context words, what is the impact of the size of the context window on precision and recall, when the relatedness measure is used for retrieval? [0.5]

NLP Quiz 1 2014 Total Marks: 20 (to be scaled to 15) Time: 1 hour
Be Precise, use bullet points rather than verbose text wherever appropriate.

---

1. What is the difference between semantics and pragmatics? Illustrate with a *crisp* example. [2]
2. What is the single most important reason for Machine Learning to have gained prominence in the NLP context in recent years? [1]
3. We discussed a context-sensitive spell-check algorithm in class.
   a. How would you detect a spelling error where *advice* is wrongly typed in as *advise*? Give two distinct approaches each of which eliminate the need for checking correctly typed words, as far as possible.
   b. Identify one parameter that is extremely important in determining the effectiveness of the algorithm and needs to be tuned appropriately.
   c. What is the role of a chi-squared test in the context of this algorithm?
   d. The chi-squared test cleverly exploits a fundamental theorem in statistics. What is the theorem, and how is it exploited? [2+1+1+2]
4. How can you make an edit distance based spell-check algorithm (that disregards context) sensitive to the following two error sources? (a)errors made in typing, with no recorded statistical data of typos and corrections (b) OCR errors. [2]
5. I have constructed a vector space which has valid English words as vectors and letters as dimensions. I want to use a cosine measure to suggest spelling corrections, by positioning a typo as a vector in the space and identifying similar words.
   a. Identify as many important and distinct limitations of this approach as you can. Suggest ways to repair these problems while retaining the basic idea of cosine similarity in a vector space.
   b. Vectors space approaches are often not used in highly effective spellcheck systems. However, these systems can still benefit from your technique in (a) above after repairing (if you have done the right repairs), and this idea is actually used in practice. How? [2+2]
6. Both distributional measures and Information Content based of similarity are corpus based. Between two objects A and B, the distributional measure is dis(A,B) and the Information Content based measure is inf(A,B). Show why this choice of notation may not be ideal. Are the goals of these measures identical? [1+1]
7. Prof. Enelpee is grey-haired, and has taught the NLP course 25 times. He has statistics of students who earned S, A and B grades. [1] He thinks he knows who will get what grades based on the values of three features F1, F2 and F3 which he detects as early as the first week. Construct a noisy channel to model this, identify source and output correctly, give an expression to estimate the quantity of information (in bits) lost in the channel and show how you will estimate it in practice. [3]

The End.

---

[1] None of his students ever got a grade lower than that, except for one by the name Profound who is now, in absence of a better thing to do, a professor greying his hairs somewhere else.

NLP (CS 6370) Quiz 2 Time: 1 hour Full Marks:15 Be Concise

1. How are document-document similarities and term-term similarities evaluated after carrying out SVD on a term-doc matrix? (1.5)
2. Use the SVD of a rectangular term-doc matrix M to show that the eigenvalues of MM^T and M^TM are identical. (1.5)
3. LSI is not ideal for dealing with polysemy. Why? (1)
4. Briefly describe how WSD can be done using the idea of word sense dominance. (The algorithm integrates knowledge from WordNet based semantic relatedness and corpus-based distributional similarity). (2)
5. You are given a set of CFG rules with probabilities attached to these rules. Under what conditions would this define a standard PCFG? (1)
6. What is bottom-up filtering? Is it of any use in improving efficiency of a top-down parser? If yes, show how, with an example. If no, justify. (1.5)
7. If lexical chains were treated as features, what would be the MOST IMPORTANT difference between these features and the ones mined using Latent Semantic Analysis? What are the parameters that would dictate the appropriate number of features in each of these schemes? (1.5)
8. The notion of precision and recall in WSD is different from the corresponding notions in Information Retrieval. How? (1)
9. When parsing using PCFGs, there are occasions when an error is made in that inappropriate parse tree is prefered over the desired one. Identify how this can happen and suggest a fix. Use a simple example to illustrate your answer. (2)
10. What is the motivation for using EM for learning parameters in PCFG? Are there situations where EM is not effective in learning a "good" set of parameters for PCFG? Can the parameters after termination be worse than the parameters it started off with? (More fundamentally) When would a set of parameters be called "good"? (2)

$X^TX = D\Sigma^2D^T$