Name:
Matriculation number:
E-Mail:

Lecture "Statistical Natural Language Processing"

Prof. Dr. D. Klakow

Exam

Friday, July 17^{th}, 2020
8.00h - 10:00h
Building A17; Room 5.1 and Günter Hotz Lecture Hall

Please read these instructions carefully before you start.

There are two types of questions: easy and medium. In each category, you have to answer a certain number of questions. A fully correct answer to an Easy question gives you 4 points and a Medium question 6 points.

Easy: you need to answer 9 out of 10 questions. Please mark the 9 to be graded:

1	2	3	4	5	6	7	8	9	10

Medium: you need to answer 2 out of 4 questions. Please mark the 2 to be graded:

11	12	13	14

- Keep 1.5 meters distance at all times.
- Wear a mask until you arrived at your designated seat.
- Put all your belongings (back packs etc.) in front of the blackboard.
- At your seat you are only allowed pencils, a ruler and your student ID.
- Go to your designated seat.
- Put your student ID at a neighbouring seet at around 1m distance.
- Use a *separate* sheet of paper to answer each question.
- Note the question number on the *top right* corner of your sheet.
- Write your name & matriculation number on the *top right* of each sheet.
- If you answer more questions than required, then only the first n required answers will be considered not the best ones.
- After the exam *sort* the sheets by question number.
- **Wait at your seat until you are called to hand in and leave!**

Sample sheet

You have 120 minutes to complete the entire exam.

Good Luck!

1 Difficulty level: Easy

1. *Zipf's Law:*

   Mandelbrot derived the following relationship between the rank of the word and its frequency:

   $$f = P(r + p)^{-B}$$

   In this equation f stands for frequency, r is rank and P, B and p are additional parameters that measure the richness of the text's use of words.

   (a) Find the values of B and p which simplify Mandelbrot's law to Zipf's law.

   (b) Imagine that you have a generator that randomly samples from a set of 26 characters from the alphabet and the space. Consider a sequence of letters followed by a single space as a word. How many words of length $n + 2$ can be generated compared to the words of length n? Start by writing down the probability of generating a word of length n.

   Which words will be generated more frequently and how this relates to Zipf's law?

2. *Entropy:*

   The cross-entropy between a random variable X with true probability mass function $p(x)$ and another probability mass function q is given by the following formula:

   $$H(X, q) = H(X) + D(p||q) = -\sum_x p(x) \log q(x)$$

   where H is entropy and D is Kullback-Leibler divergence.

   (a) Prove that the equation above holds.

   (b) Show how joint entropy between two random variables X and Y can be expressed in terms of conditional entropy. Briefly explain the meaning of $H(Y|X)$.

3. *Mutual Information:*

   (a) Write down the definition of mutual information between two random variables and explain the difference between mutual information and PMI (Pointwise Mutual Information).

   (b) Mention at least two different tasks relevant for statistical NLP where PMI can be used. What are potential problems when using PMI as a measure of association?

4. *Perplexity:*

   (a) In the lecture we have encountered two different formulas for perplexity. Give both formulas and explain each one in not more than two sentences.

   (b) Prove that these two perplexity formulas are identical. This should be done for the general case, not for the case of zerograms!

5. *Language Modeling*

   We are given the following corpus:

   ```
   I am Sam
   Sam I am
   I am Sam
   I do not like Sam
   ```

   (a) State the vocabulary size and word counts for all words of the vocabluary. What is the bigram probability for the word sequence "I like Sam" using Lidstone smoothing with $\epsilon = 0.5$?

   (b) Use linear interpolation with $\lambda_1, \lambda_2 = \{0.5, 0.5\}$ to calculate the probability for the word sequence "I like Sam" using a bigram language model without smoothing.

   (c) What is the advantage of using linear interpolation as compared to Lidstone? Describe the difference between back-off and interpolation.

6. *Backing-off Language Modeling:*

   (a) Write down the formula for a unigram backing-off language model using absolute discounting. Give a one-line description for each element of the formula.

   (b) Briefly explain how the value of discounting parameter d can affect language model performance.

   (c) Explain the intuition behind Kneser-Ney smoothing and how it extends the idea of absolute discounting.

7. *Classification Algorithms:*

   (a) Margin-based classifiers, such as support vector machines (SVM), are a subset of a family of classification algorithms known as discriminative classifiers. Explain how discriminative classification methods work and how they differ from the generative methods (e.g., naive Bayes).

   (b) Explain how SVM classifiers construct a decision boundary for a binary classification task. Illustrate your answer with a sketch and mathematical formulas when necessary. In total explain in not more than half a page.

8. *Feature Selection:*

   (a) Explain why we need feature selection in statistical NLP and mention at least 3 different types of feature selection methods.

   (b) What is the term strength and when is it needed? Explain the difference between the term strength and tf-idf.

9. *Word Sense Disambiguation:*

   The flip-flop algorithm is a way to do word sense disambiguation in case you have a parallel corpus where senses are implicitly defined by possible translation.

   (a) Write down the training algorithm in pseudo-code.

   (b) In the language modeling chapter we also talked about word classes. Explain in up to three sentences how the flip-flop algorithm could be used for word clustering.

10. *Information Retrieval:*

    (a) Write down the formula to compute tf-idf weights and explain the components. How are tf-idf weights used to retrieve relevant documents in information retrieval?

    (b) How can language models be used for IR? Show the basic formulation and explain how a relevant document is retrieved.

2 Difficulty level: Medium

11. *Compression*

    (a) A networking company uses a compression technique to encode the message before transmitting it over the network. Suppose that a new message contains the following characters with the corresponding frequencies:

character	**frequency**
A	16
B	5
C	13
D	9
E	45
F	12

    Assume that each character in the input message takes 1 byte (8 bits).

    If the compression technique used by the company is Huffman algorithm, how many bits will be saved by encoding the message? Draw a corresponding Huffman tree and justify your answer.

    (b) Write down a formula for Kraft's inequality. Can we use it to decide whether some code is a prefix code? Prove or provide a counter example.

Turn page for next question!

12. *Naive Bayes classification*

(a) Describe and give the formula for Naives Bayes. Why is it called Naive? Why do we use it as opposed to the non-naive formulation? Finally, give a scenario in which the Naive Bayes classifier will perform poorly compared to a non-naive classifier. No more than one to two sentences per question.

(b) Suppose we have the following document-term matrix:

Word	Document 1	Document 2	Document 3
tax	1	0	2
theater	0	1	1
movie	0	2	1
money	1	0	0
plan	1	0	0

Table 1: Each cell represents the count of a word in that specific document.

Additionally, we also have conditional probabilities for the appearance of a word in a document on a specific topic, a.k.a. $P(word|topic)$ values:

Word	Topic: *budgets*	Topic: *arts*
tax	$P(w = \text{tax}|topic = \text{budgets}) = 0.5$	0.01
theater	0.01	0.2
movie	0.05	0.5
money	0.2	0.01
plan	0.2	0.001

Table 2: Conditional probabilities. Each cell represents a different $P(word|topic)$ value.

Predict the topic of each of the three documents in table 1, assuming that each topic appears across documents with the same probability. Use Bayes' theorem to make these predictions, assuming that each document is a bag of words and the words are conditionally independent given the topic.

Turn page for next question!

13. *Information Retrieval:*

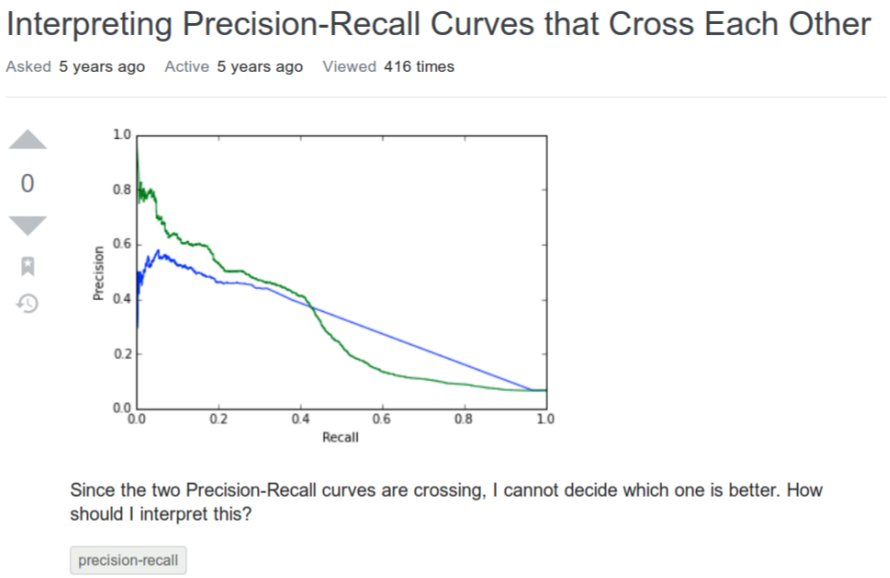
Suppose we have a set of 10,000 documents. In this case, 64 of the documents are relevant for our query. We evaluate two IR methods with this query, obtaining the following results:
Method I retrieves 183 documents of which 61 are relevant.
Method II retrieves 84 documents of which 32 are relevant.

(a) Construct a confusion matrix and compute precision, recall and accuracy for both methods. It is sufficient to write the values as fractions.

(b) When would you use which metric? Justify your answer.

(c) What is a precision-recall graph? Make a typical drawing and explain in up to five sentences.

(d) On stack overflow there is a post:



What would you reply? Please give an answer in not more than 5 sentences.

14. *Expectation Maximization Algorithm:*

(a) What is the purpose of the E-step and the M-step of the EM-algorithm? Provide a verbal description of how the algorithm works.

(b) In which situations is the EM-algorithm useful? Briefly describe the advantage of using EM and name at least two different tasks when it can be applied.

(c) Imagine a statistics class where the probability that a student gets a grade A is 0.5 and for grade B the probability is μ. Assume that you don't know how many students got exactly A or exactly B. However, you know that h students in total got either A or B. This can be represented as follows: $h = a + b$ where a and b stand for the amount of students who got A or B respectively. Use the E-step of the EM-algorithm to determine the expected values of a and b given μ.

(d) Does the EM-algorithm always converge to a local optimum (which may or may not also be a global optimum)? Explain in one or two sentences.