Name:
Matriculation number:
E-Mail:

Lecture "Statistical Natural Language Processing"

Prof. Dr. D. Klakow

Exam

Friday, July 21st, 2017
16.00h - 18:00h
Building C4 3; Room 21

Please read these instructions carefully before you start.

There are three types of questions: easy, moderate and difficult. In each category you have to answer a certain number of questions. An easy question gives 4 points, a medium question 6 and a difficult one 8.

Easy: you need to answer 7 out of 9 questions. Please mark the seven to be graded:

1	2	3	4	5	6	7	8	9

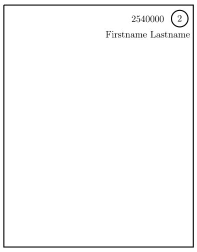
Medium: you need to answer 2 out of 4 questions. Please mark the two to be graded:

10	11	12	13

Difficult: you need to answer 1 out of 2 questions. Please mark the one to be graded:

14	15

- Put all your belongings (back packs etc.) in front of the blackboard.
- At your seat you are only allowed pencils and a ruler.
- Paper will be provided.
- Use a *separate* sheet of paper to answer each question.
- Note the question number on the *top right* corner of your sheet.
- Also write your name and matriculation number on the *top right* of each sheet.
- After the exam *sort* the sheets by question number, put this question sheet on top and staple them. Staplers will be provided.



Sample sheet

You have 2 hours to complete the entire exam.

Good luck!

1 Difficulty level: Easy

1. *Zipf's Law:*

   a) State Zipf's law and draw the corresponding plot with proper description. What is the main information that one can extract from this plot?

   b) Suppose a language has two characters only a space and "a". A text has been generated observing the following rules:

      – space is generated with a probability $p = 0.2$ while character"a" has probability $(1 - p)$

      – space is never generated twice in succession, i.e., space is never directly followed by another space.

   Under the above rules, what will be the probabilities of the following combinations:
   i. A single character word followed by a space.
   ii. A 1000 character word followed followed by a space.

2. *Mutual Information:* Prove that $I(X;Y) = H(X) - H(X|Y)$. Also, show that $D(p||q) = H(p,q) - H(p)$. Hint: first write down the definitions of $I(X;Y)$ $H(X)$ and $H(X|Y)$. Then try to find a way to pove the claimed equality.

3. *Perplexity:* Assume that you are given a sequence of N words $w_1, \ldots, w_N$ that are taken from an English text.

   i) There are two different formulas for perplexity. Give both and show that they are equivalent.

   ii) Write down the formula for the likelihood of the entire text.
   Hint: You may use either the definition in the lecture notes (which is essentially the log-likelihood) or the usual likelihood you may know.

   iii) Show the relationship between the perplexity and the likelihood, as you defined them in parts 3i and 3ii respectively.
   Hint: Write down one of the two formulas, e.g. the perplexity. Now with appropriate transformations derive an expression containing the other formula.

4. *Model Selection:* How can you determine the discounting parameter in absolute discounting (two methods, provide formula or sketch for each method as appropriate). Given a training set and two separate test sets, describe hold-out validation method of model selection.

5. *Kneser-Ney Smoothing:*

   a) What is the additional idea in Kneser-Ney smoothing beyond absolute discounting?

   b) Write down the starting formula for marginal-constraint-backing off.

   c) Which final formula for the counts of the backing-off distribution $\beta(w|\hat{h})$ does this result in?

6. *Text Categorization with KNN:*

   a) What is the basic hypothesis in using the Vector Space Model for classification? State and describe the hypothesis?

b) Explain briefly the rationale of K-Nearest Neighbor (KNN) text classification.

c) The table below shows a test set consisting of 10 documents, where each document is annotated with its corresponding class **(A, B, C)**. We wish to use this test set to find to which class, document X belongs. The distance between document X and the documents in the test set is given in the table. Use **majority voting scheme KNN** with k = 5 to calculate which class document X belongs to. Give a justification of your answer and necessary intermediate steps.

Document	1	2	3	4	5	6	7	8	9	10
Class	A	A	B	A	C	C	B	C	B	B
Distance from X	5.31	2.89	0.63	1.12	3.52	4.31	2.33	1.35	0.89	2.15

Table 1: KNN

7. *Sequence Labelling Tasks:*

a) Using a diagram and bayes theorem, describe Hidden Markov Model (HMM). Which independence assumptions are made in HMM?

b) What is the major advantage of Conditional Random Field over HMM?

c) Briefly explain the task of NER.

8. *Tf-idf:*

a) State the formula for tf-idf and explain both components of this formula. Why can't we use tf instead of tf-idf weighting? How does idf help in modifying tf?

b) Write down the Inverted Index representation for querying documents.

Terms	economy	Scotland	growth	banks	business
Document 1	10	8	0	2	1
Document 2	0	0	9	9	8
Document 3	2	2	4	4	6
Query	1	1	1	1	1

Table 2: Word frequencies in documents and query

c) Compute the Cosine-Similarity measure for each of the three documents and the query. Based on the result, which document is the best match for this query and why? What are the strengths and weaknesses of the cosine measure?

9. *Precision and Recall:*

a) Describe in your own words what Precision and Recall is. Also provide the corresponding formulas.

b) Two retrieval systems, **X** and **Y**, are being compared. Both are given the same query, applied to a collection of 1500 documents. System **X** returns 400 documents, of which 40 are relevant to the query. System **Y** returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents

relevant to the query.
Tabulate the results for each system, and compute the precision and recall for both **X** and **Y**. Gibe intermediate steps in the calculation.

c) Both precision and recall need to be taken into account when evaluating retrieval systems. Why is it not sufficient to pick one and use only that?

2 Difficulty level: Medium

10. *Maximum Likelihood Estimation:*

Task 10.1

The random variable Y has a probability density function,

$$\begin{aligned} p(Y) &= (1-\theta) + 2\theta Y, \quad 0 < Y < 1 \\ &= 0 \qquad\qquad\qquad \textit{otherwise} \end{aligned} \tag{1}$$

for $-1 < \theta < 1$. There are n observations y_i, $i = 1, 2, ..., n$ drawn independently from this distribution. Derive the expected value for y, showing all the steps.

Task 10.2

The Gaussian distribution is given by,

$$N(x_i|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big\}, \forall x_i \tag{2}$$

Suppose you are given a dataset $\{x_1, x_2, ..., x_n\}$ consisting of n observations where every observation x_i is drawn independently from the above Gaussian distribution.

a. Write the expression for the log-likelihood function of the observations.
b. Derive the maximum likelihood estimate for the parameter σ.

11. *Absolute discounting:* Assume that you are given a sequence of N words $w_1, \ldots, w_N$ with $V = \{\text{“}apple''\text{, “}tree''\text{, “}tart''\text{, “}the''\text{, “}grow''\}$ and $N = 100$

- Using Absolute discounting model with $d = 0.5$:

$$P(w_i|w_{i-1}) = \begin{cases} \frac{N(w_i,w_{i-1})-d}{N(w_{i-1})} + \alpha(w_{i-1})P(w_i) & \text{if } N(w_i, w_{i-1}) > 0 \\ \alpha(w_{i-1})P(w_i) & \text{else} \end{cases}$$

$$P(w_i) = \begin{cases} \frac{N(w_i)-d}{N} + \alpha\frac{1}{V} & \text{if } N(w_i) > 0 \\ \alpha\frac{1}{V} & \text{else} \end{cases}$$

- NB: The only bigrams with the history “apple” are:

$N(w_i, w_{i-1})$	**bigram**
1	“apple tart”
2	“apple tree”

- $N(\text{“}apple\text{''}) = 3$
- $N(\text{“}tree\text{''}) = 9$

Estimate the bigram probabilities $P(\text{“}cider\text{''}|\text{“}apple\text{''})$ and $P(\text{“}tree\text{''}|\text{“}apple\text{''})$.

12. *Author Identification with Bayes Theorem*

    a) State and explain the Bayes Theorem.

    b) Write down the principle for Bayes classification. What modification is made to the principle to make it useable and why?

    c) In a linear Naive Bayes classifier, under the Binary classification mode, what is Naive and why is it called linear? Explain how would you use a Naive Bayes classifier in a multiclass classification setting.

    d) Given a piece of text and three possible authors suppose we have prior probabilities,

    $$\begin{aligned} p(author1) &= 0.3 \\ p(author2) &= 0.3 \\ p(author3) &= 0.4 \end{aligned} \tag{3}$$

    and the probability that the text T is written by $author1$ is 0.2, probability that the text is written by $author2$ is 0.25 while the probability of the text being written by $author3$ is 0.4. Then what is the probability $P(author2|T)$ that given the text, $author2$ has written it?

13. *Text classification*

    a) Describe and give the formula for Naive Bayes classification.

    b) We have a set of documents containing certain specific words expressed in the following matrix with document-term counts:

Word	**Document A**	**Document B**	**Document C**
tax	1	0	2
theater	0	1	1
movie	0	2	1
money	1	0	0
plan	1	0	0

Also we know the conditional probability of appearance of a word in a document about certain topic:

$$\begin{array}{ll} P(word = tax|topic = budgets) = 0.5 & P(word = tax|topic = arts) = 0.01 \\ P(word = theater|topic = budgets) = 0.01 & P(word = theater|topic = arts) = 0.2 \\ P(word = movie|topic = budgets) = 0.05 & P(word = movie|topic = arts) = 0.5 \\ P(word = money|topic = budgets) = 0.2 & P(word = money|topic = arts) = 0.01 \\ P(word = plan|topic = budgets) = 0.2 & P(word = plan|topic = arts) = 0.001 \end{array}$$

Predict the topic of each of these three documents, assuming that each topic appears across documents with the same probability. To make this prediction, use Bayes' theorem, assuming each document is a bag of words and the words are conditionally independent given the topic.

3 Difficulty level: Hard

14. *Graphical Model:*

    Let X, Y, Z be three disjoint subsets of random variables.

    a) Suppose $p(x, y, z) = p(x|y)p(y|z)p(z)$. Draw the corresponding Bayesian network. Does it imply $p(y|x, z) = p(y|z)$? If yes, prove it, otherwise provide a counter example.

    b) Suppose $p(x, y, z) = p(x)p(y)p(z|x, y)$. Draw the corresponding Bayesian network. Does it imply $p(x|z) = p(x|z, y)$? If yes, prove it, otherwise provide a counter example.

    c) Suppose $p(x, y, z) = h(x, y)g(y, z)$. Draw the corresponding Markov random field. Does it imply $p(x|z) = p(x|z, y)$? If yes, prove it, otherwise provide a counter example.

    d) Draw the conditional random field for sequence labeling. Explain how to model transition and emission probability.

15. *Information Theory:*

    Suppose event X is generated from $p(x|z)$. z is drawn from $p(z)$. If we are going to encode X with minimum average bits, we need to know $p(X)$. A possible way to compute it is $\sum_z p(x|z)p(z)$, but $p(x|z)$ is near zero for most z. It's computationally too expensive to sample a meaningful z for every x. Therefore, instead of sampling from $p(z)$, we try to find a way to evaluate $p(x)$ with only sampling from $p(z|x)$

    a) When we use another distribution $q(z|x)$ to approximate the real $p(z|x)$, how many more bits on average do we need to encode event X? No proof is needed for this question.

    b) Prove when we use another distribution $q(z|x)$ to approximate the real $p(z|x)$, the minimum average bits we need to encode X is $\mathbb{E}_{p(x)}[-\mathbb{E}_{q(z|x)}(\log(p(x|z))) + \mathrm{KL}(q(z|x)||p(x))]$

    c) Show the needed bits in b) can be rewritten as $\mathbb{E}_{p(x)}[-\mathbb{E}_{q(z|x)}(\log(p(x|z))) - H(q(z|x))] + H(q(z), p(z))$.

    d) Now prove the assumption you draw in 1). Namely, show that when we use the real probability to encode X, the average needed bits is the result you obtained in (b) minus (a).