DATA.STAT.840 Statistical Methods for Text Data Analysis,
Exam December 19, 2025 12-16

This exam is open on December 19, 2025 between 12 and 16.

This examination has five problems each worth 10 points. This exam can be answered in the online form or on paper. Answering in the online form in the preferred option and answering on paper is a backup. If you answer in both ways, your answers from the form and from the paper will be combined during exam evaluation. You can answer in English or in Finnish. Please write clearly.

Each participant must complete the exam on their own without collaboration. In this bring-your-own-laptop exam, you can have a calculator, use of course material is allowed, and use of programming environments such as Python is allowed. Use of text or code that you have previously created for the course, such as your own exercise answers or your own course notes, is allowed. However, copying of text or code from others (e.g. from other students or from online material), or use of any online service that writes text for you (such as artificial intelligence tools), is not allowed.

If you answer in the online form: use the link below to access the form.

https://forms.office.com/e/abE3wDBc2Z

- You must keep track of time! The exam cannot be returned after the deadline and the online form IS NOT AUTO-RETURNED when the time runs out. In this exam, the online form closes strictly at 16:00, you must upload all files before then!
- In the online form, your answers ARE NOT AUTOSAVED during the exam. It is recommended to type your answers in an external file (e.g. in Word), and finally copy them to the form fields.
- The online form does not support all file formats, in particular it does not support uploading code files like Python directly. You must convert them to e.g. Word or PDF before uploading. Please reserve enough time to convert and upload the files!
- The form supports submitting multiple answers, but you may need to upload all parts of your answers again if you make a second answer.

If you answer on paper: You must write the following information on each paper sheet: **First name, Last name, Student number, Date, Sheet number** out of maximum (e.g. 1 of 3, 2 of 3, etc.). You must also write, on the first paper sheet, that **you have not used an AI tool**, such as ChatGPT, in any way to generate any of your answers in this exam. AI tools are not allowed to be used in the exam in any way, neither to directly solve questions nor to ask for hints or verify answers.

A Zoom chat channel is open during the exam in order to ask questions from the lecturer. The link is: **https://tinyurl.com/exam-dec19-chat**

To pass the course you must also pass the exercise packs. Results of this examination are valid for one year after the examination date. The results will be announced online.

Problem 1: Multiple-choice questions (10 points)

For each of the multiple-choice questions below, only one of the proposed answers is correct. For each question, you must write down an answer (A, B, C, D, ...), and also write down your confidence "high" or "low" for each answer; if the confidence is missing it will be treated as "low". The grading for each question is as follows:

- +2 points if the answer is correct and has "high" confidence
- +1 points if the answer is correct and has "low" confidence
- 0 points if the answer is missing
- -1 points if the answer is incorrect and has "low" confidence
- -2 points if the answer is incorrect and has "high" confidence

The total score for this problem is between 0 and 10 points.

Question 1.1:

A) The hypothesis test for collocations examines whether their frequency of occurring next to each other is higher than average.
B) Expectation-Maximization always increases the likelihood in every iteration.
C) These probabilities are possible in a bigram model: p(w1='weekly')=0.01, p(w2='meeting')=0.005, p(w2='meeting'|w1='weekly')=0.4.
D) In n-gram probability estimation, the MAP estimate is the same as the maximum likelihood estimate when a uniform prior is used.
E) A chatbot fails the Turing test if a series of trials is run and in any of the trials the evaluator can correctly tell it is a machine.
F) All the answers A-E are correct.
G) All the answers A-E are incorrect.

Question 1.2:

A) Recall means the proportion of all retrieved results that were correct.
B) The LSTM neuron computes its output as the multiplication of three gate functions.
C) In collocation analysis, the strongest collocation is the pair of words for which the greatest number of documents contain both words.
D) PLSA builds a generative model of new documents and their topics.
E) In a HMM, the probability of a sentence can be computed using only Forward probabilities without Backward probabilities.
F) All the answers A-E are correct.
G) All the answers A-E are incorrect.

Question 1.3:

A) In a PCFG, it is not necessary to use both Inside and Outside probabilities to compute the probability of a sentence.
B) Every PCFG has a strongly equivalent version in Chomsky normal form.
C) Unlike Latent Semantic Indexing, Latent Semantic Analysis is probabilistic.
D) It is possible to cleanly delineate correct from incorrect language using a deep neural network as long as it has sufficiently many LSTM layers.
E) In a mixture of Gaussians model for TF-IDF documents, each Gaussian models the distribution of one of the top words.
F) All the answers A-E are correct.
G) All the answers A-E are incorrect.

(The questions of Problem 1 continue on the next page)

(Problem 1 continues from the previous page)

Question 1.4:
A) Vocabulary pruning is used to speed up computation but is not otherwise needed for language modeling.
B) Regular expressions are overly common phrases that should be removed from text as part of data cleanup.
C) The z-score of a word has variance 1 in the collection.
D) The unigram model samples the length of the document from a multinomial distribution.
E) The cosine similarity between two documents is 1 if and only if they have identical text.
F) All the answers A-E are correct.
G) All the answers A-E are incorrect.

Question 1.5:
A) A concordance view plots a matrix of average distances between co-occurring words.
B) The top collocated word pair is the same as the word pair with the largest bigram probability.
C) In paragraph embedding, the paragraph vector is the average word vector of words in the paragraph.
D) The F1 score is the arithmetic average of precision and recall.
E) A lexical dispersion plot plots the variance of word counts over the collection for each word.
F) All the answers A-E are correct.
G) All the answers A-E are incorrect.

Problem 2: Discussion and analysis (10 points)

Question 2.1:
Discuss the similarities and differences between a) an n-gram model, b) a topic model, and c) a hidden Markov model. (3 points)

Question 2.2:
Discuss the similarities and differences between a TF-IDF paragraph representation and a paragraph embedding representation. (2 points)

Question 2.3:
Discuss the similarities and differences between a recurrent neural network, a LSTM network, and a transformer network. (3 points)

Question 2.4:
Suppose an n-gram model with n=10 has been trained, but when it is used to generate new documents, it only produces copies of existing pieces of text from the training data. Explain why this can happen. (2 points)

Problem 3: HMM computation (10 points)

Consider a hidden Markov model with the following vocabulary of six words:
{one, can, work, round, still, light}
and three states (z=1, z=2, z=3), where the states have the initial state distribution, emission distributions, and transition probabilities given below.

Compute the most likely state sequence corresponding to the observed word sequence:
"light can still work". Report your computation steps and your answer. (10 points)

$$\pi_1 = 0.5 \ , \ \pi_2 = 0 \ , \ \pi_3 = 0.5$$

	one	can	work	round	still	light
$\{\beta_1(w)\} =$	{0.2,	0,	0,	0.3,	0.25,	0.25}
$\{\beta_2(w)\} =$	{0,	0.3,	0.4,	0,	0,	0.3}
$\{\beta_3(w)\} =$	{0.1,	0.2,	0.3,	0.1,	0.1,	0.2}

$$\{\theta_{z_t|z_{t-1}}\} = \begin{array}{cc|ccc} & & z_t & & \\ & & 1 & 2 & 3 \\ z_{t-1} & 1 & \{ 0.3, & 0.2, & 0.5, \\ & 2 & 0.6, & 0.3, & 0.1, \\ & 3 & 0, & 1, & 0 \ \} \end{array}$$

Problem 4: Grimm's Fairy Tales (10 points)

Download the file

https://homepages.tuni.fi/jaakko.peltonen/grimms-fairy-tales.zip

which contains the following contents:

1) The file "grimms-fairy-tales-lowercase.txt" contains the text of "Grimms' Fairy Tales" with some initial and concluding material left out, lowercased and most special characters removed: the only remaining punctuation is periods between sentences. No other processing has been applied.

2) The directory "fairy-tales-pruned-paragraphs" contains 251 paragraphs of "Grimms' Fairy Tales" as separate documents (files "paragraph000.txt" - "paragraph250.txt"). Each paragraph has already been lowercased, lemmatized and vocabulary-pruned: you do not need to redo those processing steps.

(the problem 4 description continues on the next page)

(problem 4 description continued from the previous page)

Your task is to create and run programs for the following tasks. You can create the programs in any programming language, and using any existing libraries. You are also allowed to reuse any code provided on the course, or any code you have created earlier during the course. For each task, report your code and the requested information produced by running the code.

a) Use a text generation method from the course to learn a model of the Grimm's Fairy Tales texts, and generate two new paragraphs of 50 words based on the model. Give the resulting two paragraphs of text. You can use any method that is more advanced than a unigram model. (5 points)

b) Create a TF-IDF representation for the 251 documents. Then retrieve the closest other document according to cosine similarity, for each of the following three documents: "paragraph000.txt", "paragraph030.txt", and "paragraph100.txt". For each of the three documents, give the filename of the retrieved closest document, and the value of the cosine similarity. (5 points)

Problem 5: Essay (10 points)

Consider analyzing conversations in Reddit related to energy politics in the European Union. Reddit comments have been gathered by searching for posts (conversation starting submissions) satisfying at least 2 of 3 criteria: 1) mentions 'energy', 'renewable', 'oil', 'coal', 'gas', 'nuclear', or 'hydro'; 2) mentions 'policy', 'vote', 'treaty', 'member state', 'market', or 'commission'; 3) posted in one of 10 EU politics subreddits. For such posts, also all response comments have been gathered.

Several types of information have been recorded.
- There are 1 Million comments in total in the data set.
- Comments arise from 100000 user accounts, 80000 of which are in the EU.
- The country of residence has been recorded for 27000 EU accounts (1000 from each of the 27 member states).
- Comments have been recorded from January 2022 to October 2025; the time of each comment has been recorded (month, day, hour, minute).
- For each post/comment, it has been recorded whether it had one of 2000 flair tags.
- For each comment, it has been recorded what parent post/comment it responds to.
- For each day, the rank of each post in Reddit's "Top" and "Hot" ranking has been recorded, in the subreddit of the post and in Reddit's overall ranking.
- For each post/comment, the number of upvotes and downvotes has been recorded
- For each user account, age of the account and its "post karma" and "comment karma" (total received upvotes of posts/comments) have been recorded.
- For each user account it has been recorded if it is an account of an EU parliament member (MEP), EU commissioner, EU representative, or national politician of a member state. For parliament members, parliamentary group has been recorded; for national politicians, party affiliation has been recorded.
- For user accounts of news organizations or reporters, the organization has been recorded.

Each measurement has been recorded in an appropriate scale: time is in year/month/day/hour/minute, upvote/downvote counts are integers, affiliations are binary, etc.

(The problem 5 description is continued on the next page.)

(problem 5 description continued from the previous page)

You want to perform text analysis on the data for tasks a), b), and c) below. For each task describe the methodology you will follow and reasons for your choices. Include pre-processing steps, post-processing steps, quality criteria, text analysis methods, etc. that you would use. The tasks are:

a) You want to use statistical methods to analyze the comment collection as a whole: for example, to detect if there are previously unknown themes discussion or subsets of the data, or previously unknown types of changes over time or between the different locations. (4 points)

b) You also want to study what kinds of meanings users associate to words like environment, conservation, security, fairness, trust, sustainability, pollution, emission, trading, climate, and so on, for example whether they consider word relationships between climate, environment, and conservation to be similar to relationships of infrastructure, market, and security. (3 points)

c) Next, you want to study whether some posts and replies might be generated by bots rather than human participants. To do so, you would like to check whether human readers can tell real replies and posts from ones generated by automated posters (bots). How could you use methods of this course to accomplish this? (3 points)

This Reddit domain is complicated, one can study several questions using text analysis methods. The descriptions a), b), and c) do not specify everything, you must specify missing details so you can choose a methodology to use. No need to analyze everything, it is enough to pick some interesting questions and ways to analyze them. Special knowledge of the EU or its energy politics is not needed, it is enough to use the information above.

Use at most 2 A4 pages in size 12 font. Use full sentences. Structure your answer in paragraphs. Write in a way understandable to a student who asked you about the topic. You can assume they have the prerequisites to take the course but have not taken it.