DATA.STAT.840 Statistical Methods for Text Data Analysis, Exam December 20, 2024 12-16

Problem 1: Multiple-choice questions (10 points)

For each question, only one proposed answer is correct. For each question, write down an answer (A, B, C, D, ...), and your confidence ”high” or ”low”; if the confidence is missing it will be treated as ”low”. The grading for each question is:
+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 from 0 to 10 points.

Question 1.1:

A) PLSA builds a generative model of new documents and their topics.
B) Precision means the proportion of all correct results that a retrieval engine managed to retrieve.
C) A probabilistic context free grammar is a proper generative model if the rule probabilities that apply to each left-hand side sum to 1.
D) In collocation analysis, the strongest collocation is the pair of words for which the greatest number of documents contain both words.
E) The z-score of a word has variance 1 in the collection.
F) All answers A-E are correct
G) All answers A-E are incorrect

Question 1.2:

A) In Latent Dirichlet Allocation the model chooses a topic for the document and then generates the words of the document based on the topic.
B) Vocabulary pruning is used to speed up computation but is not otherwise needed for language modeling.
C) In a HMM, the probability of a sentence can be computed using only Forward probabilities without Backward probabilities.
D) The first transformer model was based on recurrent connections between stacked layers.
E) Regular expressions are overly common phrases that should be removed from text as part of data cleanup.
F) All answers A-E are correct
G) All answers A-E are incorrect

Question 1.3:

A) 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.
B) It is possible to cleanly delineate correct from incorrect language using a deep neural network as long as it has sufficiently many LSTM layers.
C) The LSTM neuron computes its output as the multiplication of three gate functions.
D) In paragraph embedding, the paragraph vector is the average word vector of words in the paragraph.

E) Latent Semantic Analysis can be done by eigendecomposition of the document-term matrix.
F) All answers A-E are correct
G) All answers A-E are incorrect

Question 1.4:

A) Expectation-Maximization always increases the likelihood in every iteration.
B) A lexical dispersion plot plots the variance of word counts over the collection for each word.
C) Every PCFG has a strongly equivalent version in Chomsky normal form.
D) Two documents with different text can have cosine similarity 1 in a vector space model.
E) An n-gram model predicts the next word based on the n previous words.
F) All answers A-E are correct
G) All answers A-E are incorrect

Question 1.5:

A) Stop words are the terminal symbols in a grammar.
B) In a mixture of Gaussians model for TF-IDF documents, each Gaussian models the distribution of one of the top words.
C) The F1 score is the arithmetic average of precision and recall.
D) In a PCFG, both Inside and Outside probabilities are needed to compute the probability of a sentence.
E) The unigram model samples the length of the document from a multinomial distribution.
F) All answers A-E are correct
G) All answers A-E are incorrect

Problem 2: Discussion and analysis (10 points)

Question 2.1:

Discuss the similarities and differences between a) latent semantic analysis, b) probabilistic latent semantic analysis, and c) latent Dirichlet allocation. (3 points)

Question 2.2:

Discuss the similarities and differences between a hidden Markov model and a probabilistic context-free grammar. (2 points)

Question 2.3:

Discuss the similarities and differences between a) an n-gram model, b) a continuous bag of words model, and c) a skip-gram model. (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 five words:
{may, plan, work, then, that}
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 probability of the observed word sequence:
"that plan may work". Report your computation steps and your answer. (10 points)

$$\pi_1 = 0.5 \;,\; \pi_2 = 0.25 \;,\; \pi_3 = 0.25$$

$$\begin{array}{llllll}
 & \text{may} & \text{plan} & \text{work} & \text{then} & \text{that} \\
\{\beta_1(w)\} = & \{\,0.5, & 0.2, & 0, & 0.1, & 0.2\,\} \\
\{\beta_2(w)\} = & \{\,0.3, & 0.4, & 0.3, & 0, & 0\,\} \\
\{\beta_3(w)\} = & \{\,0, & 0.4, & 0.4, & 0, & 0.2\,\}
\end{array}$$

$$\{\theta_{z_t|z_{t-1}}\} = \begin{array}{lllllll}
 & & & & z_t & & \\
 & & & 1 & 2 & 3 & \\
z_{t-1} & 1 & \{ & 0.2, & 0.4, & 0.4, & \\
 & 2 & & 0.4, & 0.2, & 0.4, & \\
 & 3 & & 0.5, & 0.5, & 0 & \}
\end{array}$$

Problem 4: Hans Andersen's Fairy Tales (10 points)

Download the file https://homepages.tuni.fi/jaakko.peltonen/hans-andersens-fairy-tales.zip which contains the following contents:
1) The file "fairytales-lowercased.txt" contains the text of "Hans Andersen's Fairy Tales" with some initial and concluding material left out, lowercased and most special characters removed but no other processing applied.
2) The directory "fairytales-all-paragraphs" simply contains each paragraph of the above text as separate documents (files "allparagraphs0000.txt" - "allparagraphs1205.txt", in case that format is convenient for you.
3) The directory "fairytales-250-paragraphs" contains a subset of 251 paragraphs of "Hans Andersen's Fairy Tales" as separate documents (files "paragraph000.txt" - "paragraph250.txt"). These paragraphs have been further processed: they have been lowercased, lemmatized and vocabulary-pruned: you do not need to redo those processing steps.

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 fairy tale texts, based on the data either in "fairytales-lowercased.txt" or in the directory "fairytales-all-paragraphs". Then use the model to 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 250 documents in the directory "fairytales-250-paragraphs", using length-normalized frequency and smoothed logarithmic inverse document frequency. Then retrieve the closest other document in that directory, according to cosine similarity, for each of the following three documents: "paragraph000.txt", "paragraph100.txt", and "paragraph200.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 Reddit conversations on ethics, benefits & dangers of online data harvesting and use of such data for artificial intelligence (AI) based analysis and for training large language models or other AI systems. Reddit comments have been gathered by searching for posts (conversation starting submissions) satisfying at least 3 of 4 criteria: 1) mention 'privacy', 'tracking', 'profiling', 2) mention 'machine learning', 'AI', 'large language model' or 'data analysis', 3) mention one of 500 related companies, or 4) posted in one of 20 subreddits on AI/ethics/privacy/politics. For such posts, also all response comments have been gathered.

Several types of information have been recorded.
• There are 1 Million comments in the data set.
• Comments arise from 100000 user accounts: 70000 in the US, 30000 in the EU.
• State of residence has been recorded for 50000 US accounts, and country of residence for 20000 EU accounts.
• Comments have been recorded between Jan 2023 - Dec 2024. 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, rank of each post in Reddit's "Top" and "Hot" ranking has been recorded, in the post's subreddit and overall.
• For each post/comment, the number of upvotes and downvotes has been recorded
• For each user account, account age and "post karma" and "comment karma" (total received upvotes of posts/comments) have been recorded.
• For each user account it has been recorded if it belongs to 1) one of 500 companies, 2) one of 100 news organizations, 3) a member of a US state/federal government or congress/senate, 4) a member of an EU state parliament/government or EU parliament/commission.
• Account affiliation has been recorded: for companies/employees the company, for news organizations/reporters the organization, for members of parliament/congress/senate/commission the party.
Each measurement is in an appropriate scale: time as year/month/day/hour/minute, upvote/downvote counts as integers, affiliations as binary, etc.

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 methods of this course 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 want to study the meanings users associate to words like algorithm, privacy, ethics, bias, discrimination, anonymity, trust, transparency, fairness, and so on, for example whether they consider word relationships between ethics, fairness, and transparency to be similar to relationships of ethics, trust, and privacy. (3 points)

c) Next, you want to study whether some posts and replies might be generated by automated posters (bots) rather than human participants. To do so, you would first like to learn to predict, in an automated way, whether a post is created by a bot. Next, you would like to test whether human readers can tell real replies and posts from ones generated by bots. How could you use methods of this course to accomplish these tasks? (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 AI ethics or the USA/EU 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.