ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (CS–431)
Bosselut, A., Chappelier, J.-C. & Rajman, M.

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (CS–431)

Fall 2023 Examination

Friday, January 26th, 2024.

NAME: Hanon Ymous

SCIPER: 000000

Seat: 0

Instructions:

You have three hours (9:15–12:15) for this exam, which consists of five independent questions with different weights. For each question, the points are indicated and the total number of points is 110 ($10 + 5 + 10 + 45 + 40$).

All documents are allowed. However, no electronic device (computer, cell phone, calculator, etc.) is authorized.

Write your answers directly on the exam sheets. If you run out of space, ask for additional official exam sheets. Do not use your own sheets; they will not be considered.

QUESTION I : What do you mean? [10 pt]

Consider the following sequence of words:

Thrse bards are nesting on spring

① **[2 pt]** Propose at least three possible solutions where the above sequence has been corrected into a meaningful English sentence.

② **[2 pt]** Tell which of the corrected versions you have proposed seems the most plausible to you and **justify why**.

③ **[6 pt]** For the corrected version you have selected as the most plausible, indicated the number of mistakes that needed to be corrected, and, for each of these mistakes, indicate:

- at which processing level it can be detected;
- and at which processing level it can be corrected.

Justify your answers.

QUESTION II : Canaries [5 pt]

Consider the following sentence:

My friends spent their holidays in Tenerife and they loved it.

① **[2 pt]** Indicate which words in this sentence instantiate an "*anaphoric reference*"; i.e., a reference to some other word(s), either within or outside the sentence.

For each of the words instantiating an anaphoric reference indicate:

- its grammatical category;
- the list of possible references it is referring to.

② **[3 pt]** For words possibly referring to several references *within* the sentence, indicate which of the possible references seems the most plausible to you.

Justify your answers.

QUESTION III : Evaluation campaign [10 pt]

Human annotators have been recruited for an evaluation campaign of automated NLP tools. In this framework, they have received the following annotation guidelines:

> You will receive a set of sentences processed by an automated NLP tool that had to fulfil the following task:
>
> > For each of the non-grammatical words (nouns, verbs, adjectives, adverbs) present in each of the sentences, determine whether the meaning of the word in the sentence corresponds or not to its most frequent meaning.
>
> The outputs of the system will have the following format:
>
> - one word per line;
> - for grammatical words, no additional information;
> - for non-grammatical words, a "`yes`" or "`no`" tag indicating whether the meaning is the most frequent one or not.
>
> Your task as annotator will be to indicate whether the outputs produced by the NLP system you are evaluating are correct or not, and you will perform this task by adding at the end of each line a "+" when you think that the output is correct, and a "−" otherwise.

① **[1 pt]** Indicate whether the annotation task performed by the automated NLP tool is of lexical, syntactic, semantic or pragmatic nature. **Justify** your answer.

② **[4 pt]** Produce the required annotations for the output produced for the sentence:

the spring has jumped out of the box

given here:

```
the:
spring: yes
has:
jumped: yes
out: yes
of:
the:
box: no
```

Justify each of your annotation decisions.

③ **[5 pt]** Do you think the provided annotation guidelines are well defined?

Indicate the difficulties the annotators may be faced with when trying to apply them.
What is the impact of these difficulties on the quality/expoitability of the produced annotations?
How could this impact be measured?

Provide a ***detailed* justification** for your answers and use the concrete example given in the previous question whenever possible.

QUESTION IV : From characters to documents [45 pt]

① **[3 pt]** Using a 4-gram model of characters, what is the expression of the ratio $P(\text{around})/P(\text{rounds})$?

Provide your answer as a formula using only model parameters and with the fewer possible terms.

② **[4 pt]** Still being a 4-gram model of characters, how can the model be improved to take into account that "*around*" and "*rounds*" are actually words (as opposed to substrings in the middle of a word).

Motivate your answer and explain how your proposition would modify your answer to sub-question ①.

③ **[4 pt]** Assume that the considered alphabet consists of 128 different characters.

What is the maximum-likelihood estimate of the parameter corresponding to a 4-gram (of characters) that appears only 5 times in a corpus of 3'493'743 words, resulting in a total of 12'619'400 characters?

What is its estimated value using a Dirichlet prior with a uniform parameter set to $3 \cdot 10^{-3}$?

Justify your answers.

④ **[5 pt]** Considering the probability of a **word** sequence $w_1 ... w_n$, what is the fundamental difference between a 2-gram language model and an order-1 HMM Part-of-Speech tagger?

Support your claim by providing the formula of $P(w_1, ..., w_n)$ in both cases.

⑤ **[12 pt]** Consider the following sentence:

the quick fox jumps over the lazy dog

and an order-1 HMM for Part-of-Speech tagging with the following parameters (not exhaustive, but no missing information to solve the question):

the: `Det`
quick: `Adj`: $2 \cdot 10^{-4}$, `Adv`: $9 \cdot 10^{-4}$, `N`: $4 \cdot 10^{-4}$
fox: `N`: $2 \cdot 10^{-4}$, `V`: $8 \cdot 10^{-4}$
jumps: `N`: 10^{-4}, `V`: $3 \cdot 10^{-4}$
over: `Prep`
lazy: `Adj`
dog: `N`: $6 \cdot 10^{-4}$, `V`: $7 \cdot 10^{-4}$

	Adj	Adv	Det	N	V	Prep
Adj	0.15	0.1	0.3	0.2	0.05	0.25
Adv	0.05	0.2	**0**	0.1	0.15	**0**
Det	0.02	0.1	**0**	0.04	0.05	0.3
N	0.4	0.1	0.7	0.3	0.45	r
V	0.3	0.4	**0**	0.25	0.1	s
Prep	0.02	0.1	**0**	p	q	**0**

(a) [8 pt] Provide the tightest possible condition(s) between p, q, r and s so that the tag of "*jumps*" in the most probable sequence of tags for the above sentence is `V`.

(b) [4 pt] If these conditions are fullfiled, what is the most probable sequence of tags for the above sentence?

Fully justify your answers. (There is also room for answer at the back.)

⑥ **[6 pt]** Assume now you want to do some classification (e.g. spam filtering) of documents containing spelling errors. And we know we have different spelling errors depending on the class (for instance, spelling errors in spam emails are not of the same kind as in non-spam emails).

Propose a simple probabilitic model combining Naive Bayes classification and probabilistic spelling error correction to perform such a task. Fully explain all your notations.

⑦ **[5 pt]** Consider the following two documents:

d_1: the quick brown fox jumps over the lazy dog

d_2: the amber hound of the lazy fox hunter jumped and chased the wise fox

The indexing set reduces to: cat, dog, fox, jump, quick, run, wise

Over 1'000'000 documents, the number of documents that contain a given word is:

brown	cat	chase	dog	fox	hound	hunter	jump	lazy	quick	run	amber	wise
2'000	20'000	500	10'000	1'000	800	1'500	10'000	15'000	1'000	30'000	500	100

What is the cosine similarity between d_1 and d_2 using *simple* preprocessing and tf-idf weighting? Provide your answer as a *formula* with *numerical values* and **justify** your answer.

⑧ **[6 pt]** Finally, we consider the evaluation of a document retrieval system.

The IDs of the documents that are considered to be relevant for each of the 3 queries are:

```
query 1: 1 2 3 4 5
query 2: 2 6 7
query 3: 3 4 6 8
```

And the ranked output of the system to be evaluated are (best document first):

```
query 1: 2 1 5 6 4 3
query 2: 7 5 6 2
query 3: 8 6 4 2 3
```

Compute

a) [1 pt] P@3 for each query;

b) [2 pt] R-precision;

c) [3 pt] and MAP.

Justify your answers. (There is also room for answer at the back.)

QUESTION V : Automated Question Answering System [40 pt]

You are a student assistant (SA) for a class at EPFL and think you could create an automated system to answer student questions on Moodle and Ed discussion boards. You decide to use a deep learning-based chatbot to do this! With your system in hand, you'll be able to let it answer student questions and you can go skiing more often!

Luckily, you have access to 10'000 previous interactions between students and SAs from previous iterations of the course. Students never responded back to these messages (how rude!), so all of this data is in the form of a question x and answer y.

You decide to build a model using a transformer-based language model.

① **[1 pt]** On how many of these question-answer pairs would you train your model?
Justify your answer.

② **[1 pt]** You decide to use a vocabulary size of 12'000 tokens. However, your corpus has approximately 15'000 unique words in it. Assuming you want each token to be a full word from your corpus, how should your model process the remaining tokens that will not be in the vocabulary?

You're not sure you have enough data to train a good system, so you decide to pretrain a set of word embeddings on a larger corpus of textbooks.

③ **[2 pt]** Explain the technical difference between the continuous bag of words (CBOW) and skip-gram algorithms for training word embeddings.

continues on back ☞

You decide to use the continuous bag of words algorithm to train your word embeddings. To test whether your training algorithm works correctly, you test it with a small vocabulary of five words and provide it the sequence of words "*what day is the exam*" with the following embeddings:

$$\begin{aligned} \text{what} &= \left[\ln 2, \quad \ln 0.5 \right] \\ \text{day} &= \left[\ln 0.5, \ \ln 2 \right] \\ \text{is} &= \left[\ln 0.5, \ \ln 0.5 \right] \\ \text{the} &= \left[\ln 1.5, \ \ln 0.5 \right] \\ \text{exam} &= \left[\ln 2, \quad \ln 2 \right] \end{aligned}$$

(where $\ln$ is the natural logarithm function of base e);
and output vocabulary projection U:

$$U = \begin{pmatrix} 0 & 1 & 2 & 1 & 0 \\ 1 & 2 & 3 & 2 & 1 \end{pmatrix}$$

You can assume each column of U corresponds to the following vocabulary items:
what, day, is, the, exam.

④ **[6 pt]** Using a window size of 2, what is the probability of the word "*is*" according to the continuous bag of words network?
Justify your answer.

⑤ **[2 pt]** Using a window size of 1, what is the probability of the word "*the*" according to the continuous bag of words network?
Justify your answer.

Now that your embeddings are pretrained, you train your transformer language model. For the following questions, assume a single-headed attention function and use the following input embeddings as key vectors:

$$\begin{aligned} \text{what} &= [\ 2, \quad 0.5\ ] \\ \text{day} &= [\ 0.5, 2\ ] \\ \text{is} &= [\ 0.5, 0.5\ ] \\ \text{the} &= [\ 2, \quad -2\ ] \\ \text{exam} &= [\ 1, \quad 1\ ] \end{aligned}$$

⑥ **[6 pt]** Using scaled dot product attention, what is the attention distribution over key vectors for the word "*exam*" as the query in the first attention layer? You can ignore position embeddings. Assume that W^K, W^V are identity matrices and

$$W^Q = \begin{pmatrix} \sqrt{2}\ \ln(4) & 0 \\ 0 & \sqrt{2}\ \ln(4) \end{pmatrix}$$

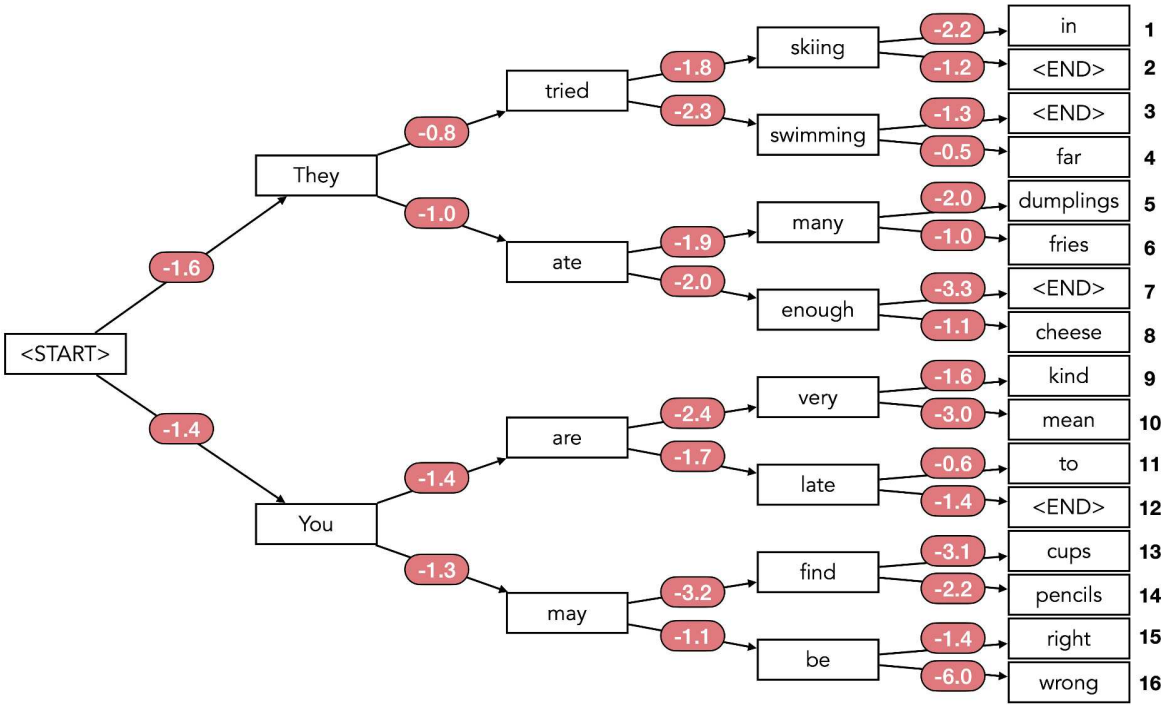
Justify your answer and provide all the steps of your computation.

⑦ **[2 pt]** What is the attention distribution if the position embedding in the first position is $[-1,\ 0.5]$ and the others are $[0,\ 0]$?
Justify your answer.

continues on back ☞

⑧ **[2 pt]** Assuming you provide a student question x, and an answer y of length $T+1$ with tokens $[y_0,\ y_1,\ \cdots,\ y_t,\ \cdots,\ y_T]$, write out the objective you would optimize to maximize the likelihood that the model produces answers that were similar to the ones in your dataset. Specify whether you would *maximize* or *minimize* this objective.

⑨ **[15 pt]** Now that you've trained your model on your dataset, you can produce text from it. You pre-compute the step-by-step probability distributions over all tokens for four steps. Below, we show the top-2 highest probability tokens in these distributions at each step (along with their **log probability**):



For each of the following sub-questions **a)** to **e)**, give your answer using the indices (1–16) to the right of the above figure.

a) **[3 pt]** What is the optimal sequence? **Justify** your answer.

b) **[2 pt]** Which sequence are you ***least*** likely to produce using top-k sampling with k = 2? **Justify** your answer.

c) **[3 pt]** What sequence would be produced using beam search with a beam size of 2? **Justify** your answer by annotating the graph.

d) **[2 pt]** In the limit, what sequence would you be most likely to produce if you were to re-compute your probability distribution over tokens at each step, but set a temperature coefficient that approaches 0? **Justify** your answer.

e) **[5 pt]** List the sequences that can be generated if you use top-p sampling with $p = 0.27$? **Justify** your answer

Hints: Assume $\ln(0.27) \simeq -1.3$. In top-p sampling, you sample from all tokens until the cumulative distribution *exceeds* the threshold.

continues on back ☞

⑩ **[3 pt]** Unfortunately, you didn't clean the dataset of previous interactions you used to train the model so the model was trained on the raw interactions from the Moodle discussion board.

Are there examples in your data that might lead your model to produce harm from the perspective of:

a) Leaking private information

b) Disinformation

c) Toxicity

Justify your answers.