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 — Solution of the exam

Friday, January 26th, 2024.

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.

The proposed sentences must be:

- credible corrections of the original sequence, i.e. at a reasonable lexicographic distance;
- correct English sentences, both at lexical and syntactic level;
- meaningful English sentences, i.e. having a possible literal meaning (syntactic level).

The words at reasonable lexicographic distance are:

Thrse: These, Those, Three

bards: bars, bards, barns, bands, birds, boards, beards

are: are

nesting : nesting, resting, testing

on: on, in

spring: spring, springs

Most plausible answers:

(a) These birds are nesting in spring

(b) Three birds are nesting in spring

(c) These bards are resting in spring

Other possibilities:

- Three boards are resting on springs
- These birds are nesting on springs
- Those birds are resting in springs
- Three birds are resting on a spring
- These bands are resting in spring

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

The most plausible correction is 1.

(a) is preferred over (b) because it exhibits less dependence wrt. any potential additional context (no necessity for justifying why 3 and not any other number).

(a) is preferred over (c) not because it is semantically meaningful (so is (c)), but because it is pragmatically more convincing (birds are indeed nesting in spring, while there is no specific information about bards resting in spring).

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.

The errors are therefore:

(a) "*Thrse*" replaced by "*These*"

(b) "*bards*" replaced by "*birds*"

(c) "*on*" replaced by "*in*"

- Error (a) can be detected at the **lexical** level because "*Thrse*" is not a valid English word.
  Possible corrections at minimal lexical distance (here 1) are:
  "*Terse*", "*These*", "*Those*", "*Three*".
  All these corrections are acceptable at the syntactic level.
  "*Terse birds*" does not correspond to a semantically meaningful combination, but "*These/Those/Three birds*" are all semantically acceptable.
  The correction can therefore only happen at the **pragmatic** level.
- Error (b) can be detected at the **semantic** level because bards are not typically nesting.
  Possible corrections at minimal lexical distance (here 1) that are also corresponding to a syntactically correct combination are:
  "*bars*", "*bands*", "*barns*", "*birds*", "*boards*", "*cards*", and "*yards*".
  Only "*birds*" is acceptable at the **semantic** level, where the correction can thus be made.
- Error (c) can be detected at **syntactic** level, because "*on* spring" is not a syntactically valid combination.
  The only possible corrections at minimal lexical distance (here 1) that is syntactically correct is "in".
  The correction can thus be made at **syntactic** level as well.

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.

Anaphoric references:

(a) My: Possessive determiner
External reference to the author of the sentence

(b) their: Possessive determiner
Internal reference to "*friends*";

(c) they: Pronoun
Internal reference to either "*friends*", or "*holidays*";

(d) it: Pronoun
Internal reference to "*spent their holidays in Tenerife*" or to "*Tenerife*" only.

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.

The pronouns "*they*" and "*it*" are potentially ambiguous.

For "*they*", the ambiguity can be resolved at the semantic level, as only animated entities can love, which imposes that the correct reference is "*friends*".

For "*it*", from a pragmatic perspective, the most plausible reference is likely to be the location "*Tenerife*", because when people talk about their holidays, they often express their enjoyment or love for the place they visited. However, without more context, it's not possible to definitively determine the intended reference, and it could still be open to interpretation.

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.

The annotation task is of semantic and pragmatic nature.

It is of course of semantic nature because it is about the meaning of the words.

However, in case of polysemous words, finding the meaning instantiated within a given sentence may require to refer to some general (societal or cultural) context.

For example, in the sentence
"*The professor asked the students to bring a notebook*"
the word "*notebook*" may refer to a "*book of blank pages used for recording notes*" or to a "*laptop computer*".

Identifying the right meaning will then depend on the societal context, and thus be of pragmatic nature.

Therefore, the annotation task is also of pragmatic nature.

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.

This question must be done in two steps:

(a) deciding what are the meanings instantiated in the sentence;
Clearly, "*spring*" means the coiled device, "*jumped*" the action of moving, "*box*" the container.
"*out*" is a grammatical word and should not be tagged (thus the annotation must be "−").

(b) Once the instantiated meanings are identified, a decision must be taken on whether they correspond to the most frequent ones or not; this is fully *subjective* and the annotation can thus be either "+" or "−", provided that "most frequent" is mentioned for the first one, and "not most frequent" for the second

```
the: + (grammatical word)
spring: yes - (here the meaning is the coiled device, not the
               most frequent one (season is)
has: + (grammatical word)
jumped: yes + (here the meaning is the physical movement, which
               is indeed the most frequent meaning)
out: yes - ("out" is a (part of) a compound preposition here,
            thus a grammatical word)
of: + (grammatical word)
the: + (grammatical word)
box: no - (here the meaning is the "container", which is indeed
           the most frequent one)
```

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.

The provided annotation guidelines are **not well defined** because they do not provide any exploitable definition of what should be considered as "the most frequent meaning".

Without additional information, the decisions of the annotators will be very subjective and the resulting annotation will probably lead to **very low inter-annotator agreement**.

For making the annotations more objective, for each of the non-grammatical words, a list of meanings may be provided in addition to the output produced by the system, so that the annotators take their decisions on a more common ground.

For example, for "*spring*", the output proposed to the annotators maybe:

```
spring: yes [season/device/source of water]
```

In addition, it is not realistic to believe that annotator can reasonably agree on what is the most frequent meaning of a word.

Again, this may entail a very low inter-annotator agreement, which will make the produced annotation virtually useless.

For example, for "*spring*", depending on the individual perceptions of each annotator, the most frequent meaning may be considered to be either the season of the device.

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.

$$r = \frac{P(\text{arou})}{P(\text{unds})} \frac{\sum_x P(\text{und}x)}{\sum_x P(\text{rou}x)}$$

Simplest justification:

$$P(\text{around}) = P(X_1 = \text{a}|X_2^6 = \text{round}) \cdot P(\text{round}) = P(X_1 = \text{a}|X_2^4 = \text{rou}) \cdot P(\text{round}) = \frac{P(\text{arou})}{P(\text{rou})}$$

$$P(\text{rounds}) = P(\text{round}) \cdot P(\text{s}|\text{round}) = P(\text{round}) \cdot P(\text{s}|\text{und}) = \frac{P(\text{unds})}{P(\text{und})}$$

For those who prefer to stick to parameters from the beginning on:

$$P(\text{around}) = P(\text{arou}) \cdot \frac{P(\text{roun})}{\sum_x P(\text{rou}x)} \cdot \frac{P(\text{ound})}{P(\text{oun})}$$

$$P(\text{rounds}) = P(\text{roun}) \cdot \frac{P(\text{ound})}{P(\text{oun})} \cdot \frac{P(\text{unds})}{\sum_x P(\text{und}x)}$$

Comment: Many students didn't go as far as using parameters (i.e. **4-grams** probabilities) only.

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.

Add special character for word boundary. Let's denote it here by '#', for simplicity. We thus now have:

$$P(\#\text{around}\#) = P(\#\text{aro}) \cdot \frac{P(\text{arou})}{P(\text{aro})} \cdot \frac{P(\text{roun})}{P(\text{rou})} \cdot \frac{P(\text{ound})}{P(\text{oun})} \cdot \frac{P(\text{und}\#)}{P(\text{und})}$$

$$P(\#\text{rounds}\#) = P(\#\text{rou}) \cdot \frac{P(\text{roun})}{P(\text{rou})} \cdot \frac{P(\text{ound})}{P(\text{oun})} \cdot \frac{P(\text{unds})}{P(\text{und})} \cdot \frac{P(\text{unds}\#)}{P(\text{nds})}$$

Thus more terms to make the difference (relevant).

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.

A corpus of 12'619'400 characters contains 12'619'397 4-grams.

This MLE is $\frac{5}{12'619'397}$

The total number of possible 4-grams with this alphabet is 128^4

Thus additive smoothing via Dirichlet prior is $\frac{5.003}{12'619'397+3\cdot10^{-3}\times128^4}$

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.

The fundamental difference lies in the conditionnal dependecies: direct in 2-gram language model and indirect through tags in order-1 HMM.

$$P(w_1, ..., w_n) = P(w_1)\prod_{i=2}^{n} P(w_i|w_{i-1})$$

$$P(w_1, ..., w_n) = \sum_{t_1,...,t_n} P(t_1)P(w_1|t_1)\prod_{i=2}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

Comment: Too many students do not seem to know marginalisation.

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`: $2\cdot10^{-4}$, `Adv`: $9\cdot10^{-4}$, `N`: $4\cdot10^{-4}$
fox: `N`: $2\cdot10^{-4}$, `V`: $8\cdot10^{-4}$
jumps: `N`: 10^{-4}, `V`: $3\cdot10^{-4}$
over: `Prep`
lazy: `Adj`
dog: `N`: $6\cdot10^{-4}$, `V`: $7\cdot10^{-4}$

	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.)

First notice that the above table contains transition probabilities from column tag to row tag (look at the sum of the first (or forth) row).
Furthermore, the $P(\texttt{Prep}|\texttt{Det})$ is zero; thus reading the matrix the transposed way would lead to a zero probability transition from `Prep` to `Det`, which makes this sentence a zero probability sequence...

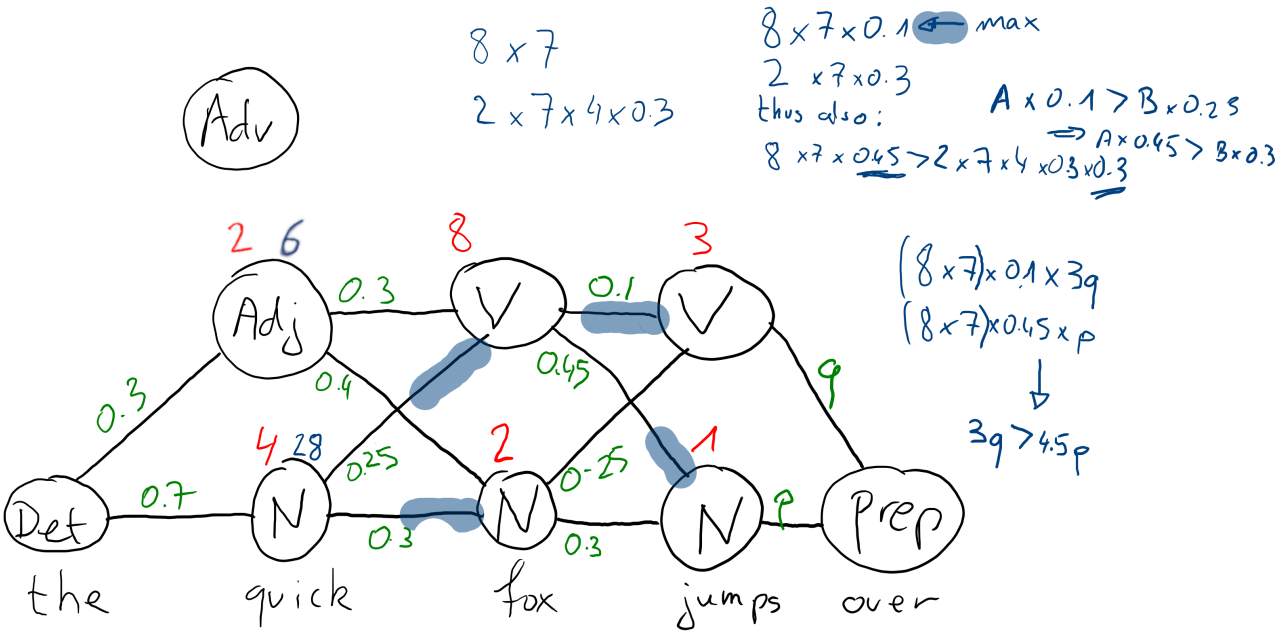
Second, we have quite some non-ambiguous words:

```
the  quick  fox  jumps  over  the  lazy  dog
Det                    Prep  Det  Adj
```

which makes the optimization of the two ambiguous parts *independent*.

Maximizing the second part ("*dog*") is not difficult and can be done in your head without any advanced algorithm: $P(\texttt{N}|\texttt{Adj}) \cdot P(\text{dog}|\texttt{N})$ compared to $P(\texttt{V}|\texttt{Adj}) \cdot P(\text{dog}|\texttt{V})$; i.e. 0.4×6 compared to 0.3×7, leading to N.

For the first part, use the Viterbi algorithm (rather than brute force):



The condition for "*jumps*" to be taged as V is thus:

$$3q > 4.5p \qquad (\text{and } p \leq 0.11, q \leq 0.2)$$

The most probable sequence is then:

the	quick	fox	jumps	over	the	lazy	dog
Det	N	V	V	Prep	Det	Adj	N

Comment: Too many students didn't pay enough attention to the transition probabilities and, among the ones who took the wrong way, almost all of them didn't care about P(Det|Prep) to be 0: some simply removed the second "*the*", others simply ignored the link (put an arrow without any number).

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.

Let C denote the random variable representing the class (e.g. binary for spam/ham), and w_i denote the i-th input token (maybe OoV spelling error).

We are looking for $\text{Argmax}_c\, P(C = c|w_1, ..., w_n)$ which following standard Naive Bayes approach reduces to $\text{Argmax}_c\, P(C = c) \prod_{i=1}^{n} P(w_i|C = c)$

Now, w_i might be a spelling error. Let's denote by t the proper words (in the lexicon $\mathcal{L}$). $P(w_i|C=c)$ is then simply:

$$P(w_i|C=c) = \sum_{t\in\mathcal{L}} P(w_i,t|C=c) = \sum_{t\in\mathcal{L}} P(w_i|t,C=c)P(t|C=c)$$

Where we can see:

(a) $P(t|C=c)$: the usual Naive Bayes classifier term

(b) $P(w_i|t,C=c)$: the error model **for the class** c

Comment: Several students suggest to take "the spelling errors as features" which does not make much sense: first of all, what are "the spelling errors"? (some, in fact many, words remain correct) and then most of all, this means than *any* string can be a feature: infinite feature set!

Also, again, almost nobody considers marginalization.

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.

(using log-10)

$$d_1 = [0 \quad 1\times 2 \quad 1\times 3 \quad 1\times 2 \quad 1\times 3 \quad 0 \quad 0]$$
$$d_s = [0 \quad 0 \quad 2\times 3 \quad 1\times 2 \quad 0 \quad 0 \quad 1\times 4]$$

$$\cos(d_1,d_2) = \frac{3\times 6 + 2\times 2}{\sqrt{4+9+4+9}\ \sqrt{36+4+16}} = \frac{22}{\sqrt{26}\ \sqrt{56}}$$

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.)

a) order does not matter, only how many relevant document over three

```
query 1: 3/3 = 1
query 2: 2/3
query 3: 3/3 = 1
```

b) We average: P@5 for q1, P@3 for q2 and P@4 for q3:

$$\text{R-Prec} = \frac{1}{3}\left(\frac{4}{5} + \frac{2}{3} + \frac{3}{4}\right) = \frac{48 + 40 + 45}{180} = \frac{133}{180}$$

(first formula is enough)

c) Average precision for each query is:

```
query 1: ( 1 + 1 + 1 + 4/5 + 5/6 ) / 5 = 139 / 150
query 2: ( 1 + 2/3 + 3/4 ) / 3         =  29 / 36
query 3: ( 1 + 1 + 1 + 4/5 ) / 4       =  19 / 20
```

(first formulas are enough), thus

$$\text{MAP} = \frac{1}{3}\left(\frac{139}{150} + \frac{29}{36} + \frac{19}{20}\right) = \frac{1207}{1350}$$

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.

Fewer than 10'000 and save the remaining for evaluation

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?

Map the remaining tokens to <UNK> token.

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.

CBOW: Learn to predict word from context words

Skip-gram: Learn to predict context words from chosen word

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.

$$\begin{aligned} U \times (\textit{what} + \textit{day} + \textit{the} + \textit{exam}) &= U \times \left[\ln 2 + \ln 0.5 + \ln 1.5 + \ln 2 \ , \ \ln 0.5 + \ln 2 + \ln 0.5 + \ln 2 \right] \\ &= U \times \left[\ln 3 \ , \ 0 \right] \\ &= \left[0 \ , \ \ln 3 \ , \ \ln 9 \ , \ \ln 3 \ , \ 0 \right] \end{aligned}$$

which through softmax thus leads to

$$\frac{1}{17}\left[1 \ , \ 3 \ , \ 9 \ , \ 3 \ , \ 1 \right]$$

Thus the answer is $\frac{9}{17}$.

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.

$$\begin{aligned} U \times (\textit{is} + \textit{exam}) &= U \times \left[\ln 2 + \ln 0.5 \ , \ \ln 0.5 + \ln 2 \right] \\ &= U \times \left[0 \ , \ 0 \right] \end{aligned}$$

which through softmax thus leads to

$$\frac{1}{5}\left[1 \ , \ 1 \ , \ 1 \ , \ 1 \ , \ 1 \right]$$

Thus the answer is $\frac{1}{5}$.

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:

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

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.

Notice that W^Q is simply $\sqrt{2}\ \ln 4$ times identity. Thus

$$\frac{(W^Q Q)\cdot(W^K K)}{\sqrt{2}} = (\ln 4)\ Q\cdot K = (\ln 4)\ \left[\ 1,\ 1\ \right]\cdot K$$

Noticing that $\left[\ 1,\ 1\ \right]\cdot K$ simply adds its components, you end up after softmax to:

$$\frac{\left[\ 4^{2.5}\ ,\ 4^{2.5}\ ,\ 4^{1}\ ,\ 4^{0}\ ,\ 4^{2}\ \right]}{\text{sum}} = \frac{\left[\ 32\ ,\ 32\ ,\ 4\ ,\ 1\ ,\ 16\ \right]}{85}$$

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.

The only thing that changes is thus the first vector which now becomes $[2-1=1,\ 0.5+0.5=1]$, the sum of the components of which is no longer 2.5 but 2, leading to:

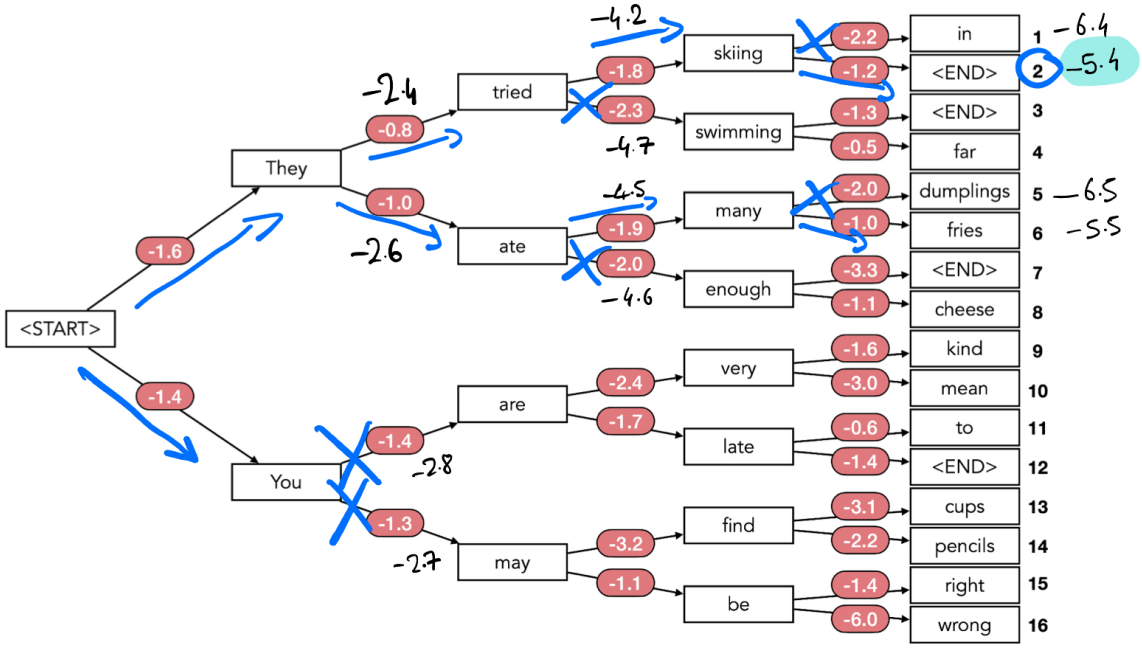
$$\frac{\left[\ 16\ ,\ 32\ ,\ 4\ ,\ 1\ ,\ 16\ \right]}{69}$$

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.

minimize $\sum_{t=1}^{T} -\log P(y_t|y_{<t}, x)$

= maximize $\sum_{t=1}^{T} \log P(y_t|y_{<t}, x)$

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.

11

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

16

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

2

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.

15

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.

2, 4, 15, 16

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.

a) TAs may sign their names or provide personal e-mails in their responses

b) TAs provide a wrong answer to a question from a student which the model subsequently trains on

c) Not too many examples. Fine to say none. Also fine to say the TAs might insult the students although unlikely.