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 (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Lexical Level
Solutions

J.-C. Chappelier M. Rajman

v. 20220928 – 1

QUESTION I **[4 pt]**

(adapted from Spring 2018 quiz 1)
For this question, *one or more* assertions can be correct. Tick only the correct assertion(s). There will be a penalty for wrong assertions ticked.

For a 3-grams of characters model, which of the following terms are *parameters* directly estimated from the learning corpus?

[✔] $P(\text{cat})$	[ ] $P(\text{at}\,|\,\text{c})$	[✔] $P(\text{cta})$	[ ] $P(\text{cats})$
[ ] $P(\text{c}\,|\,\text{at})$	[ ] $P(\text{t}\,|\,\text{ca})$	[✔] $P(\text{tac})$	[ ] $P(\text{ca})$

- Don't forget $P(\text{cta})$, nor $P(\text{tac})$: *all* 3-grams are estimated (even if the estimation is 0, which in this case may not even be the case: e.g. ***dicta****te*)

- Bigrams are not parameters; their estimation comes from the one of 3-grams (sum). For instance:
$$P(\text{ca}) = \sum_x P(\text{ca}x)$$

- $P(x|yz)$ are not parameters either. They are computed *from/with* the parameters. For instance:
$$P(\text{t}|\text{ca}) = \frac{P(\text{cat})}{\sum_x P(\text{ca}x)}$$

Barème et remarques pour la correction :
Right column (no tick outside 2nd col.): 1 pt; each correct tick: 1pt; wrong ticks: -0.5 each.

QUESTION II **[4 pt]**

Consider the following lexicon, which also indicates the probability of a word:

```
debt   0.04
deft   0.03
dust   0.04
exit   0.08
next   0.05
test   0.07
text   0.05
```

Using a simple probabilitic spelling error corrector (as simple as proposed in the lecture), order the candidates proposed to correct the OoV "*dext*".

First order by number of errors, then by decreasing word probability:
next, text (equal)
debt
deft
exit (at distance 2)
test
dust

QUESTION III **[5 pt]**

(from Fall 2018 quiz 1)
For this question, we ask you to tick *one and only one* of the proposed answers. If there is more than one single tick, your answers will not be considered at all.

In a language identification system using 4-grams Markov model, what is the probability of "*chats*" to be French (F), assuming that[1]:

$P(F \mid \text{chat}) = 2 \cdot 10^{-5}$ $P(F \mid \text{hats}) = 13 \cdot 10^{-4}$	$P(F \mid \text{cha}) = 3 \cdot 10^{-6}$ $P(F, \text{t} \mid \text{cha}) = 17 \cdot 10^{-7}$ $P(F, \text{s} \mid \text{hat}) = 5 \cdot 10^{-8}$	$P(\text{cha} \mid F) = 5 \cdot 10^{-5}$ $P(\text{t} \mid \text{cha}, F) = 19 \cdot 10^{-4}$ $P(\text{s} \mid \text{hat}, F) = 11 \cdot 10^{-3}$	$P(\text{hat} \mid \text{c}, F) = 7 \cdot 10^{-6}$ $P(\text{ats} \mid \text{h}, F) = 2 \cdot 10^{-7}$
$P(\text{ch} \mid F) = 11 \cdot 10^{-5}$	$P(\text{a} \mid \text{ch}, F) = 3 \cdot 10^{-4}$	$P(\text{t} \mid \text{ha}, F) = 7 \cdot 10^{-8}$	$P(\text{s} \mid \text{at}, F) = 13 \cdot 10^{-3}$

Answer:

[1]Most of those values are, of course, fake and incompatible.

[ ] $2 \times 13 \times 10^{-9}$	[ ] $19 \times 11 \times 10^{-7}$
[ ] $3 \times 2 \times 13 \times 10^{-15}$	[ ] $5 \times 7 \times 2 \times 10^{-18}$
[ ] $3 \times 17 \times 5 \times 10^{-21}$	[ ] $11 \times 3 \times 7 \times 13 \times 10^{-20}$
[✔] $5 \times 19 \times 11 \times 10^{-12}$	[ ] another value ( )

It's indeed $P(\text{chats} \mid F)$ we are talking about: indeed when one says "the probability of (some value) x ...", she indeed means $P(x)$, in the sense that the sum over the set of alternative values to x (including x itself) is 1.

So in this case: "the probability of *chats*..." means $P(\text{chats}...)$ in the sense that it has to sum up to 1 on all the alternatives to "chats". It's thus indeed $P(\text{chats} \mid F)$ and not $P(F \mid \text{chats})$ (the later does not at all sum up to one on alternatives of "*chats*"!)

$P(F \mid \text{chats})$ would be phrased something like "the probability of the writing language to be French when we read "*chats*".

Thus: $P(\text{chats} \mid F) = P(\text{cha} \mid F) \times P(\text{t} \mid \text{cha}, F) \times P(\text{s} \mid \text{hat}, F)$.

When done in exam, many students missed the initial $P(\text{cha} \mid F)$; some others didn't realize that $P(\text{chat} \mid F)/P(\text{cha} \mid F)$ is indeed $P(\text{t} \mid \text{cha}, F)$ (or similarly, some wanted to have $P(\text{chat} \mid F)$, which is indeed $P(\text{cha} \mid F) \times P(\text{t} \mid \text{cha}, F)$).

QUESTION IV **[5 pt]**

(from Spring 2019 quiz 1)

From a corpus of N occurences of m different tokens:

① What is the exact number of occurrences of 4-grams (of tokens) present in the corpus?

$$N - 3$$

(or if you want to be even more precise: 0 if $N < 4$)

② How many different 4-grams (values) could you possibly have?

$$m^4$$

(or if you want to be even more precise: $\min(m^4, N - 3)$)

③ Only G different 4-grams (values) are indeed observed. What is the probability of the others:

(a) using Maximum-Likelihood estimation?

$$0$$

(b) using "additive smoothing" with a Dirichlet prior with parameter $(\alpha, \cdots, \alpha)$, of appropriate dimension, where α is a real-number between 0 and 1?

$$\frac{\alpha}{N - 3 + \alpha\, m^4}$$

④ If a 4-gram has a probability estimated to be p with Maximum-Likelihood estimation, what would be its probability if estimated using "additive smoothing" with a Dirichlet prior with parameter $(\alpha, \cdots, \alpha)$?

$$\frac{(N-3)\, p + \alpha}{N - 3 + \alpha\, m^4}$$