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.

Introduction to Natural Language Processing

Solution and comments to Fall 2020, Graded Quiz #2

NAME: **SCIPER:**

Instructions:

You have 45 minutes (8:15–9:00) for this quiz, which consists of several questions of different weights. For each question the points are indicated and the total number of points is 30.
[...]

The solution comes in red. And some comments in blue.

QUESTION I [2 pt]

Indicate which of the following assertions are correct.

Using K. Oflazer 1996 spelling error correction algorithm presented in class, the correction of some input string X with a threshold $\theta = 3$ will visit *at least*:

(Select one or several answers; penalty for wrong ticks)

[ ] all sequences of characters that are at distance 4 from X

[ ] all sequences of characters that are at distance 3 from X

[ ] all sequences of characters that are at distance 1 from X

[ ] all sequences of characters

[ ] all the strings in the lexicon that are at distance 4 from X

[✓] all the strings in the lexicon that are at distance 3 from X

[✓] all the strings in the lexicon that are at distance 1 from X

[ ] all the strings in the lexicon

QUESTION II [5 pt]

What are the values (a), (b) and (c) in the highlighted cells when computing the edit distance betweeen "*compilation*" and "*compliance*" using insertion, deletion, substitution and transposition?

		c	o	m	p	l	i	a	n	c	e
c											
o											
m											
p											
i											
l											
a								**(b)**			
t											
i		**(a)**									
o											
n											**(c)**

$a =$8 $b =$1 $c =$5

Here is what you should directly be able to fill, simply from definition or direct application of the lecture:

		c	o	m	p	l	i	a	n	c	e
c		0									
o		1									
m		2									
p		3			0						
i		4									
l		5					1				
a		6						**1**			
t		7						2	2		
i		**8**						3		3	
o								4		≥ 4	4
n									4	≥ 4	**5**

Notice that there is absolutely no need to fill it completely to provide the answers.

QUESTION III **[2 pt]**

The edit distance between “piece” and “peace” is:
(Select one or several answers; penalty for wrong ticks)

[ ] 5.

[ ] 3.

[ ] 1, if considering insertion and deletion only.

[✔] 2, if considering insertion and deletion only.

[ ] 3, if considering insertion and deletion only.

[ ] 1, if considering insertion, deletion and substitution.

[✔] 2, if considering insertion, deletion and substitution.

[ ] 3, if considering insertion, deletion and substitution.

[ ] 1, if considering insertion, deletion, transposition and substitution.

[✔] 2, if considering insertion, deletion, transposition and substitution.

[ ] 3, if considering insertion, deletion, transposition and substitution.

QUESTION IV [3.5 pt]

Having the following edit distance chart:

		e	r	a	d	i
	0	1	2	3	4	5
d	1	1	2	3	3	4
i	2	2	2	3	4	3
c	3	3	3	3	4	4
s	4	4	4	4	4	5
o	5	5	5	5	5	5
v	6	6	6	6	6	6
e	7	6	7	7	7	7
r	8	7	6	7	8	8

What is the cut-off edit distance (as defined in K. Oflazer 1996 algorithm) for input string $X =$"*dicsover*" with respect to candidate solution $Y =$"*eradi*", for a correction threshold $\theta = 3$?

3: $I = 5 - 3 = 2$, $J = 5 + 3 = 8$, $D_c = \min\{3, 4, 5, 5, 6, 7, 8\} = \mathbf{3}$

QUESTION V [1.5 pt]

Considering K. Oflazer 1996 spelling error correction algorithm presented in class, the correction of the input string "huomr" with respect to candidate prefix Y ="pro", for a correction threshold $\theta = 3$, lead to a cut-off edit-distance of 3. Would the algorithm continue searching with prefix Y, for instance would is consider prefix "prof" (provided that "professor" is in the lexicon)?

[✔] Yes

[ ] No

QUESTION VI [1 pt]

What are the outputs produced by a morphological analyzer?
(Select one or several answers; penalty for wrong ticks)

[✔] canonical representations

[ ] surface forms

[ ] association between canonical representations and surface forms

[ ] association between surface forms and canonical representations

QUESTION VII [1 pt]

What *affixes* can one identify in the word "unbreakableness"?
(Select one or several answers; penalty for wrong ticks)

[ ] the suffix "leness"

[ ] the root "break"

[ ] the infix "break"

[ ] the prefix "able"

[✔] the suffix "able"

[ ] the circumfix "un ... ness"

[✔] the prefix "un"

[✔] the suffix "ness"

Note: the root is not an affix.

QUESTION VIII [1 pt]

What word could be associated to the following definition: "Related to a selection in advance"?
(Select one)

[ ] selectionally

[ ] preselection

[ ] retroselection

[✔] preselectional

QUESTION IX [1 pt]

The possibility of transforming a verbal form into a nominal form is a task using:
(Select one or several answers; penalty for wrong ticks)

[ ] a syntactic analyzer

[ ] inflectional morphology

[ ] a well trained PoS tagger

[✔] derivational morphology

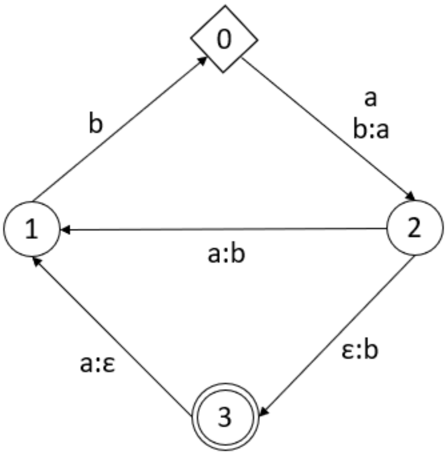
QUESTION X [1 pt]

What are possible morphological analyses of "drinks"?
(Select one or several answers; penalty for wrong ticks)

[ ] N+s
[ ] drink+VERB+p
[ ] drink+ADJ
[✔] drink+N+p
[✔] drink+VERB+3+Sing+Present+Indicative
[ ] drinks+N+p

QUESTION XI [4 pt]

Consider the following transducer T:



Which of the following statements are true for T, provided that strings are complemented only on the right by empty characters when necessary?
(Select one or several answers; penalty for wrong ticks)

[✔] T can recognize an infinite number of string associations
[ ] T recognizes the (baa, ab) association
[✔] T recognizes the (aabb, abbab) association
[✔] T associates a unique right string to the left string aaba
[ ] T associates a unique left string to the right string ab (**Note:** it as both a and b.)
[✔] there are left strings exclusively consisting of characters in a, b that cannot be associated to any right string by T (**Note:** for instance: aa.)

QUESTION XII [1 pt]

If we keep the same graphical description for the transducer, but change the "padding convention" by complementing strings by empty characters *only on the left* when necessary, would the new transducer recognize the same string associations?

[ ] Yes

[✓] No

Note: T always adds an extra 'b' at the end; if the padding convention is shifted on the right, that extra suffix 'b' has to correspond to some letter and there currently is no path leading to end state that associate some (left) character to (right) 'b'.

QUESTION XIII [3 pt]

Consider 3 regular expressions A, B, and C, such that:

- the sets of strings recognized by each of the regular expressions is non empty;
- the set of strings recognized by B is included in the set of strings recognized by A;
- some strings are recognized simultaneously by A and by C; and
- no string is recognized simultaneously by B and C.

Which of the following statements are true?
(where, for a regular expression X, (X) denotes the transducer which associate every string recognized by X to itself)

(Select one or several answers; penalty for wrong ticks)

[ ] Any string recognized by A but not by B is a left string in an association recognized by the transducer $(A)(C)$

[✓] Any string recognized by B is (at least) associated to itself by the transducer $(A) \otimes (B)$

[ ] $(A \otimes B) \circ (C)$ recognizes a non empty set of string associations

[✓] $(B \otimes A) \circ (C)$ recognizes a non empty set of string associations

QUESTION XIV [1 pt]

Consider a baseline transducer T used to model the plural of English nouns in the form: $T = T1 \circ T2 \circ T3$, where:

- $T1 = ([a-z]+)((\backslash+N\backslash+p) \otimes (\backslash+1))$
- $T2 = ([a-z]+)((\backslash+1) \otimes (Xs))$
- $T3 = ([a-z]+)((Xs) \otimes (s))$

What is(/are) the right string(s) associated by T to the left string "box+N+p"?

(Select one or several answers; penalty for wrong ticks)

[ ] box+1

[ ] boxXs

[✔] boxs

[ ] boxes

QUESTION XV [2 pt]

To take into account that:

- for nouns ending in "x", the plural form is produced by replacing the final "x" by "xes" (e.g. fox –> foxes),
- for nouns ending in "y", the plural form is produced by replacing the final "y" by "ies" (e.g. fly –> flies), and the transducer T3 is modified into (with a usual, non-lazy, $|$):

$$T3 = ([a-z]+)(((xXs) \otimes (xes))|((yXs) \otimes (ies))|(([^\wedge x])(Xs) \otimes (s)))$$

What is(/are) the right string(s) associated by T to the left string "fly+N+p"?

(Select one or several answers; penalty for wrong ticks)

[ ] fly+1

[ ] flyXs

[✔] flys

[✔] flies