Old Exam Questions for Practice

NLP

2021-2022

Introduction

This is a collection of old exam questions to help you practice for the NLP exam. This collection does not yet cover the newest topics from Edition 3 of the J&M book. Questions about vectors and neural networks are not included here, but you will find those in the practice exam of October 2020.

WARNING: This collection contains more questions than a real exam. A typical NLP exam has 16-18 multiple choice questions (5 points each) and 1 or 2 open questions (usually 10 points each). See the exam of October 2020 for a typical example of the size of the exam.

Answers to the questions will be provided in a separate file.

Multiple choice questions

1. Below you see the transition table of a finite state automaton. The initial state is 0; the final state is 4. $\emptyset$ denotes the *fail* state, where no successful transition is possible for the given symbol. Note that when encountering a b in state 2, two transitions are possible: we can either stay in state 2 or move on to state 3.

	Input		
State	a	b	c
0	1	$\emptyset$	$\emptyset$
1	$\emptyset$	2	$\emptyset$
2	$\emptyset$	2, 3	$\emptyset$
3	$\emptyset$	$\emptyset$	4
4:	$\emptyset$	$\emptyset$	3

   Which of the following regular expressions matches the FSA? (The regular expression has to match exactly the same set of strings that is accepted by the FSA: not more, not less.)

   (a) abb+c(cc)*

   (b) abbb*c+

   (c) abbb+c+c

   (d) ab*bb(cc)+

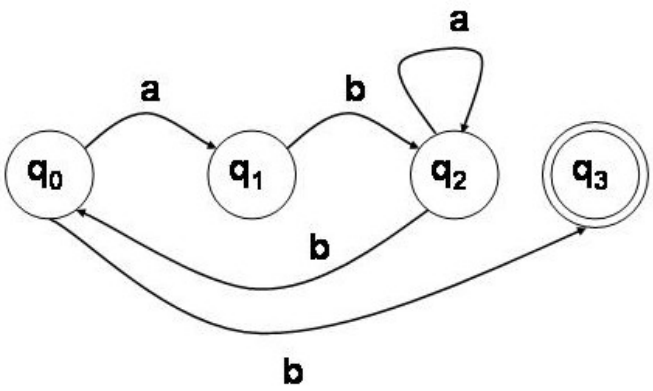
2. Which of the following regular expressions matches the automaton shown below?

   (a) aba+bb

   (b) aba+b*

   (c) (aba*b)*

   (d) (aba*b)*b



3. Consider the following two statements about inflection and derivation.

   i) In English tacking an s onto the end of an infinitive verb (for example, *sing* → *sings*) is a form of inflection.

   ii) In English adding the suffix *-ism* to an adjective (for example, *national* → *nationalism*) is a form of inflection.

   Which of these statements are true?

   (a) Only i) is true

   (b) Only ii) is true

   (c) Both are true

   (d) None of them is true

4. In his time, Shakespeare created many new words that are now commonly used in the English language. One of them is the word *unreal*. This word was used for the first time in Macbeth: *Hence, horrible shadow! Unreal mockery, hence!* (Macbeth, Act III, Scene IV).

   What kind(s) of word formation took place in the creation of *unreal*?

   (a) Compounding

   (b) Inflection

   (c) Derivation

   (d) Inflection and derivation

5. Which kind of word formation process has taken place to form the word *bookings*?

   (a) Inflection

   (b) Derivation

   (c) Both inflection and derivation

   (d) It depends on whether the stem *book* is a noun or a verb

6. Which of the following statements is **NOT** true?

   (a) Lemmatization takes part of speech into account; stemming does not

   (b) Lemmatization is aimed at producing the linguistic stem; stemming is not

   (c) Lemmatization can handle words formed through derivation; stemming cannot

   (d) Lemmatization can deal with irregular word forms; stemming cannot

7. Some natural language stemming algorithm has the following two properties: 1) the words *adhere* and *adhesion* remain distinct after stemming; 2) the words *experiment* and *experience* are reduced to the same stem.

   Which of the following statements is true?

   (a) Both 1) and 2) are errors of commission

   (b) Both 1) and 2) are errors of omission

   (c) 1) is an error of commission, 2) is an error of omission

   (d) 1) is an error of omission, 2) is an error of commission

8. When discussing the Porter stemmer in Section 2.4.4, J&M give the pair *organization-organ* as an example of an error of commission. However, the linguistic stem of *organization* is in fact *organ*: the word *organization* is formed in the same way as for example *fossilization* (fossil → fossilize → fossilization) and its morphological structure is organ+ize+ation.

   So, did J&M make a mistake with this example or are they right, and why?

   (a) J&M are right, because stemming is not aimed at finding the correct linguistic stem of a word

   (b) J&M are right, because from an Information Retrieval perspective, *organization* should not have been stemmed at all

   (c) J&M have made a mistake, because *organ* is the linguistic stem of *organization* and therefore *organization* has been stemmed correctly

   (d) J&M have made a mistake, because reducing *organization* to *organ* is not an error of commission but an error of omission.

9. Let $med(x, y)$ denote the minimum edit distance between words x and y. Consider the equality:

   $med(u + v, u + w) = med(v, w)$,

   where u, v and w are sequences of characters (words) and where $u + v$ denotes the concatenation of words u and v. Thus for example $tea + spoon$ equals $teaspoon$.

   The equality

   (a) is true for all u, v and w, no matter the cost functions of deletion, insertion and substitution.

   (b) is only true when the cost of substitution is the sum of the cost of insertion and deletion.

   (c) is only true when the minimum edit distance is a true distance function (i.e. satisfies the three requirements for distance functions δ: 1) $\delta(x, x) = 0$, 2) $\delta(x, y) = \delta(y, x)$ and 3) $\delta(x, y) <= \delta(x, z) + \delta(z, y)$.)

   (d) is only true when u is the empty string.

10. What is the *minimum edit distance* between the words *"seventy"* and *"seventeen"* when we use the Levenshtein distance in which *insertion* and *deletion* have a cost 1 and *substitution* has cost 2 ?

    (a) 3

    (b) 4

    (c) 5

    (d) 6

11. We are building a bigram model with explicit modelling of both the start and end of sentences. How many unique bigrams do we find in the sentence *to be or not to be*? (The question mark is NOT part of the sentence)

    (a) 4

    (b) 5

    (c) 6

    (d) 7

12. We are building a regularised unigram model using $k = 1$ on a training set of the following sentence: *a b b a a*, with a vocabulary of three words: "*a*", "*b*" and "*c*". What is the probability of the different unigrams?

    (a) $p(a) = \frac{3}{5} \quad p(b) = \frac{2}{5} \quad p(c) = \frac{0}{5}$

    (b) $p(a) = \frac{4}{5} \quad p(b) = \frac{3}{5} \quad p(c) = \frac{1}{5}$

    (c) $p(a) = \frac{4}{6} \quad p(b) = \frac{3}{6} \quad p(c) = \frac{1}{6}$

    (d) $p(a) = \frac{4}{8} \quad p(b) = \frac{3}{8} \quad p(c) = \frac{1}{8}$

13. We often use the logarithms of probabilities rather than probabilities, because:

    (a) We speed up computations and get the exact same results as with probabilities

    (b) We speed up computations and get sufficiently good approximations of the actual probabilities

    (c) This allows us to work around hardware limitations for numerical precision

    (d) Using the logarithms allows us to get more accurate results when dealing with large negative probabilities

14. What is a correct Part of Speech tagging for the sentence *that oil can burn*? (We use the following PoS tags: Det=determiner; Art=article; N=noun; V=verb; Pron=pronoun)

    (a) that/Art oil/N can/N burn/V

    (b) that/Det oil/N can/V burn/V

    (c) that/Pron oil/V can/V burn/V

    (d) that/Pron oil/N can/V burn/V

15. Assigning part-of-speech tags is hard because words can be ambiguous, in the sense that:

    (a) Most words may have a different POS tag depending on their context in a sentence

    (b) A relatively small number of very commonly used words may have a different POS tag depending on their context in the sentence

    (c) The same word may have multiple different tags at the same time

    (d) Different words may have the same tags in a given context

16. What is the best Part of Speech for the word *that* in the sentence *Did he go that far?*

    (a) Conjunctive

    (b) Adjective

    (c) Adverb

    (d) Pronoun

17. A language has 100 words. Every word w has equal probability of occurring in a sentence. For every word w_j, every word w_i also has equal probability of occurring after w_j. What are the values of the probabilities $P(w_j w_i)$, the probability that the bigram $w_j w_i$ occurs, and $P(w_i|w_j)$, the probability of word w_i if the preceding word is w_j ?

    (a) $P(w_j w_i) = 0.01$ and $P(w_i|w_j) = 0.01$

    (b) $P(w_j w_i) = 0.0001$ and $P(w_i|w_j) = 0.01$

    (c) $P(w_j w_i) = 0.01$ and $P(w_i|w_j) = 0.0001$

    (d) $P(w_j w_i) = 0.0001$ and $P(w_i|w_j) = 0.0001$

18. A language has 1000 words. In a sample of 100 words we count 9 words that occur 10 times, and 2 words that occur 5 times. The remaining 989 words of the language do not occur in our sample. We use *Add-One* smoothing to estimate probabilities of word occurrences. What is the *total probability* of the re-estimated probabilities for words that didn't occur in our sample?

    (a) 0.899091

    (b) 0.000909

    (c) 0.001010

    (d) 0.901010

19. Given the "document" *abcabcabcabc* over the three letter alphabet a, b, c, the value for the maximum likelihood estimate of the conditional probability $P_{mle}(c|ab)$ is:

    (a) 1.0

    (b) 0.0

    (c) 0.5

    (d) 0.75

20. Consider the following artificial co-occurrence counts.

	book	**text**	**tree**	**paper**
read	0	4	1	11
write	4	0	0	1
expose	2	1	0	0

    What is the cosine similarity between *write* and *expose*?

    (a) $\sqrt{4+1}/\sqrt{85}$

    (b) $8/\sqrt{138 \times 17 \times 5}$

    (c) $8/\sqrt{85}$

    (d) None of the above

21. Which of the following statements is true?

    (a) Skip-grams and Continuous Bag Of Words use a deep, non-linear, neural network to compute word embeddings.

    (b) Skip-grams are identical to a eigenvector and eigenvalue decomposition of the Pointwise Mutual Information matrix.

    (c) Skip-grams do not result in a single embedding, but in two word embeddings. Combined, these are used to reconstruct the context of words.

    (d) None of the above is correct.

22. Consider the sentence *she bought a potato and some carrots when she went to the corner store.* Which of the following lists of word sequences only contain constituents of this sentence?

    (a) "she bought", "a potato and some carrots"

    (b) "she", "to the corner store"

    (c) "the corner store", "bought a potato"

    (d) "potato", "she went to the corner store"

23. Consider the following grammar.

Rules:	Lexicon
S → NP VP	N → people | ideas | ...
NP → (Adj) N	Adj → young | furious | dangerous | green | ...
VP → V (Adv)	V → sleep | have | ...
VP → V NP	Adv → very | well | furiously | ...

Given this grammar, which of the following sentences is ungrammatical?

(a) People have dangerous ideas.

(b) Young people sleep very well.

(c) Green ideas sleep furiously.

(d) Young people sleep furious people.

24. Consider the following grammar.

Rules:	Lexicon:
1. S → AP Noun	Noun → house | book | ...
2. S → Noun	Adj → nice | red | angry | ...
3. AP → AP Adj	
4. AP → Adj	

Assume you are using a top-down depth-first parser that applies the grammar rules in the order given above. What will happen when trying to parse the input *nice angry house*?

(a) The parser will succeed and output one parse

(b) The parser will succeed and output more than one parse

(c) The parser will fail

(d) The parser will loop forever

25. Consider the following sentences. The sentences marked with * are not proper English.

the angry player is young
the young coach is angry
the former coach is angry at the player
*the coach is former
*the angry at the player coach is young
*the coach is young at the player

Now consider the following constituency grammars. Which of these grammars generates the first three sentences given above, but none of the sentences marked with * ?

Grammar (i)

Rules:		Lexicon:
S → NP VP	VP → Vbe AP	N → player | coach
NP → Det NP	AP → AP PP	Adj → young | angry | former
NP → AP N	AP → Adj	P → at | on
NP → N	PP → P NP	Vbe → is
		Det → the

Grammar (ii)

Rules:		Lexicon:
S → NP VP	NP → N	N → player | coach
VP → Vbe AP	AP1 → Adj1	Adj1 → angry
VP → Vbe AP3	AP2 → Adj2	Adj2 → former
NP → Det NP	AP3 → Adj3	Adj3 → young
NP → AP1 N	AP → AP1 PP	P → at | on
NP → AP2 N	PP → P NP	Vbe → is
		Det → the

Grammar (iii)

Rules:	Lexicon:
S → NP VP (AP1)	N → player | coach
VP → Vbe	Adj1 → angry
NP → N	Adj2 → former
NP → Det (AP2) N	Adj3 → young
AP1 → (Adj1 | Adj3) (PP)	P → at | on
AP2 → Adj1 | Adj2 | Adj3	Vbe → is
PP → P NP	Det → the

(a) Grammar (i)

(b) Grammar (ii)

(c) Grammar (iii)

(d) None of these grammars

26. Each of the two headlines below has two possible meanings: one is the intended one and one is a humorous (but unintended) meaning.

    - Headline 1: Fear of Bird Flu Grips Turkey
    - Headline 2: Sisters Reunited After 18 Years In Queue

    For which of the two headlines is the humorous meaning caused by syntactic ambiguity?

    (a) Headline 1

    (b) Headline 2

    (c) Both Headline 1 and Headline 2

    (d) Neither of the two headlines

27. Consider the following grammars. They all share the same lexicon.

    Grammar 1:
    S → NP VP
    VP → V NP NP
    NP → Det Nom
    NP → Det AP Nom
    AP → Adv AP
    AP → Adj

    Grammar 2:
    S → NP VP
    VP → VP NP
    NP → Det Nom
    Nom → AP Nom
    AP → Adv AP
    AP → nice

    Grammar 3:
    S → NP VP
    VP → likes NP
    NP → Det Nom
    Nom → AP Nom
    AP → very AP
    AP → nice

    Lexicon:
    V → give | like
    Det → a | the
    Adv → very
    Adj → nice
    Nom → girl | boy | book

    Which of these grammars is in Chomsky Normal Form?

    (a) Grammar 1.

    (b) Grammar 2.

    (c) Grammar 3.

    (d) None of the grammars.

28. Consider the following grammar.

Rules
S → NP VP
VP → run | like | eat
VP → V NP
VP → V PP
NP → horses | grass | sugar |cubes
NP → N N
NP → N PP
PP → P NP

Lexicon
N → horses | grass | sugar |cubes
V → run | like | eat
P → like | on

If we use this grammar to fill in a CKY parse table for the sentence *Horses like sugar cubes* (see Figure 1), what will the contents of cell [1, 3] be when the entire table is filled?

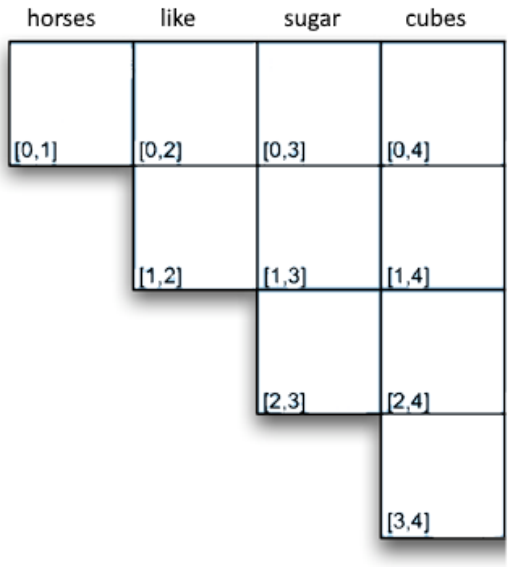


Figure 1: Empty CKY parse table for *Horses like sugar cubes*.

(a) VP, PP
(b) NP, VP, PP
(c) VP_1, VP_2, S
(d) The cell will be empty

29. Which of the following statements is **not** true?

    (a) Constituency grammars are more suitable than dependency grammars for capturing generalizations between how groups of words behave.

    (b) Constituency grammars are more suitable than dependency grammars for specifying constraints on word order.

    (c) Unlike Constituency grammars, dependency grammars do not make a distinction between terminals and non-terminals.

    (d) Unlike dependency grammars, Constituency grammars cannot express information about the head words in a sentence.

30. Which of the following is **not** a difference between the arc-standard and arc-eager transition systems in transition-based dependency parsing?

    (a) In the arc-eager system, the RightArc operator is always applied immediately; in the arc-standard system, its application may be postponed.

    (b) In the arc-standard system, the RightArc operator removes the word from the top of the stack; in the arc-eager system, it does not.

    (c) In the arc-standard system, Shift is the only operator that adds words to the stack; in the arc-eager system, it is not the only operator that does this.

    (d) In the arc-standard system, backtracking is not possible; in the arc-eager system, it is.

Open question

1. In his famous book *The Language Instinct* (Penguin Press, 1994), Steven Pinker quotes the following, accidentally ambiguous sentence: "Tonight's program discusses stress, exercise, nutrition and sex with Celtic forward Scott Wedman, Dr Ruth Westheimer, and Dick Cavett."

    (a) Write down a small context-free grammar that can be used to derive this sentence. Keep your grammar compact, but make sure it can assign at least two different tree structures to the sentence, illustrating the structural ambiguity. Proper names and *tonight's* may be treated as single words in the lexicon. Use only linguistic categories and phrases (Noun, Verb, NP, VP, etc.) in your grammar, and take care that your rules adequately reflect constituent structure.

    (b) How many different tree structures can your grammar assign to the sentence? Draw two of them and explain the ambiguity of the sentence in terms of the differences between the trees.