Solver: Wu Ziang

1)

a)

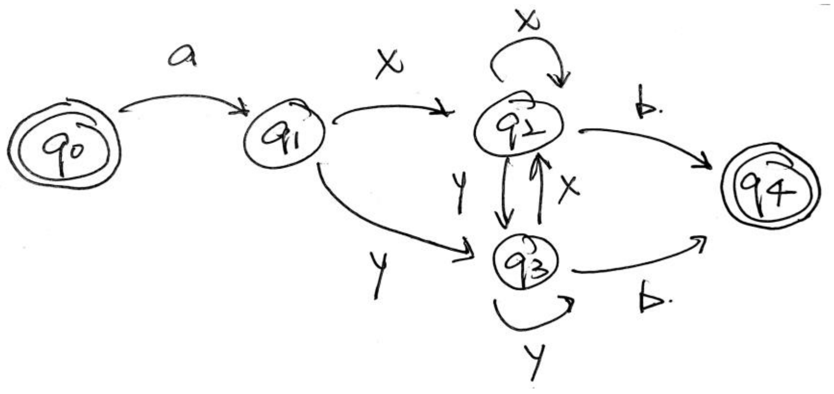
i) 1. The social media text is usually casual and does not follow the formal grammar, for example the sentence does not have the complete set of subject and object. The formal text follows the grammar strictly, for example accurate choice of word.
2. The social media text does not use the punctuation marks properly, adding more ambiguities to the sentence segmentation, for example "!!!" are used to express strong emotion. The formal text properly uses these punctuations.
3. The social media text is generally short and easy to understand, only contains some straightforward meaning in simple sentence structure. The formal text are generally more complex in terms of meaning and sentence structure.

ii) Tokenization is the process of identifying the tokens, words and symbols, in a text in preparation of further processing.
Tokenization of social media text is more ambiguous, for example "!!!" could be one compound to express strong emotion in social media text and should be tokenized as one single token, and should use different set of tokens, for example "XD" could be a smiley face in social media text but not a proper word in formal text.

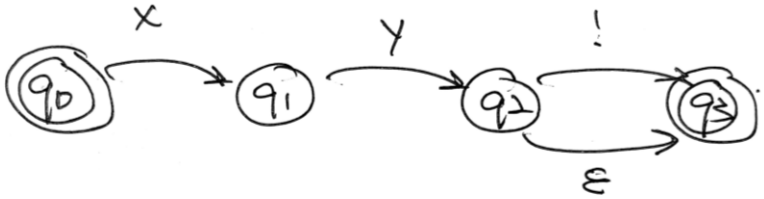
iii) Stemming is the process of identifying the stem of the word, the core meaning-bearing unit, which might not be a proper word, for example the stem of "computational" is "computation" and the stem of "computation" could be "compute" or "computa" depending on the stemmer.
It is not a good choice to apply stemming on social media text because they only use simple words with very few affixes attached to the stems.

b)

i)



ii)



The FSA in Q1(b)(i) is deterministic while the FSA in Q1(b)(ii) is non-deterministic.
The deterministic means that at each point in processing, there is always one unique choice, either transit to another state or stay at the current state.

2)

a)

i)

Word counts

we	study	review	on	yelp	a	research
2	2	3	2	3	1	1

Word probabilities

we	study	review	on	yelp	a	research
1/7	1/7	3/14	1/7	3/14	1/14	1/14

Bigram counts

	we	study	review	on	yelp	a	research
we		2					
study			1		1		
review				1			
on					2		
yelp			2				
a							1
research				1			

Bigram probabilities

	we	study	review	on	yelp	a	research
we	1/9	1/3	1/9	1/9	1/9	1/9	1/9
study	1/9	1/9	2/9	1/9	2/9	1/9	1/9
review	1/10	1/10	1/10	1/5	1/10	1/10	1/10
on	1/9	1/9	1/9	1/9	1/3	1/9	1/9
yelp	1/10	1/10	3/10	1/10	1/10	1/10	1/10
a	1/8	1/8	1/8	1/8	1/8	1/8	1/4
research	1/8	1/8	1/8	1/4	1/8	1/8	1/8

ii) $P(\text{a study on yelp}) = \frac{1}{14} \times \frac{1}{8} \times \frac{1}{9} \times \frac{1}{3} = \mathbf{\frac{1}{3024}}$

iii) The word "research" in this should be chosen and replaces by <UNK>, because it only appears once in the three sentences given, which means it is not very relevant in the given context and is less frequently used in the language compared to “a”.

b)

i) $P(\text{I have this watch}) = 0 + 0 + 0.0001 + 0 = 0.0001$
The detailed steps of computation based on the hidden Markov model shows the transition probabilities and the probability of each subsequence.

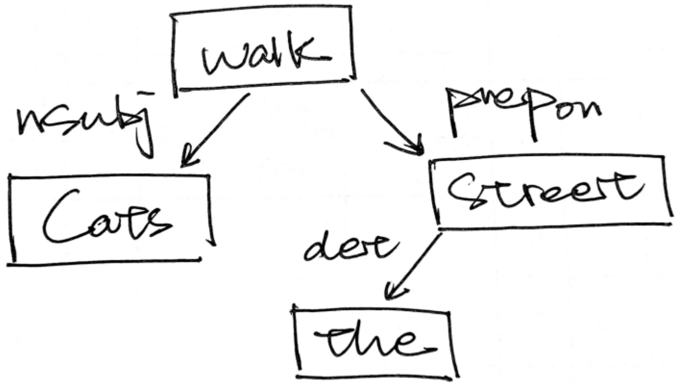
Hidden Markov Model (HMM)

		I		have		this		watch	
<S>	1	DT 0.5	0	DT 0 0 0.1 0.2	0	DT 0 0 0.1 0.2	0.0004	DT 0 0 0.1 0.2	0
		PRP 0.1	0.08	PRP 0 0 0 0.2	0	PRP 0 0 0 0.2	0	PRP 0 0 0 0.2	0
		NN 0.2	0	NN 0.5 0.1 0.4 0.2	0.0004	NN 0.5 0.1 0.4 0.2	0	NN 0.5 0.1 0.4 0.2	0.0001
		VB 0.1	0	VB 0 0.3 0.4 0	0.0048	VB 0 0.3 0.4 0	0	VB 0 0.3 0.4 0	0

ii) The backtrace pointer is not necessary in this computation, because the sum of all possible sequences is requested not the most probable sequence.

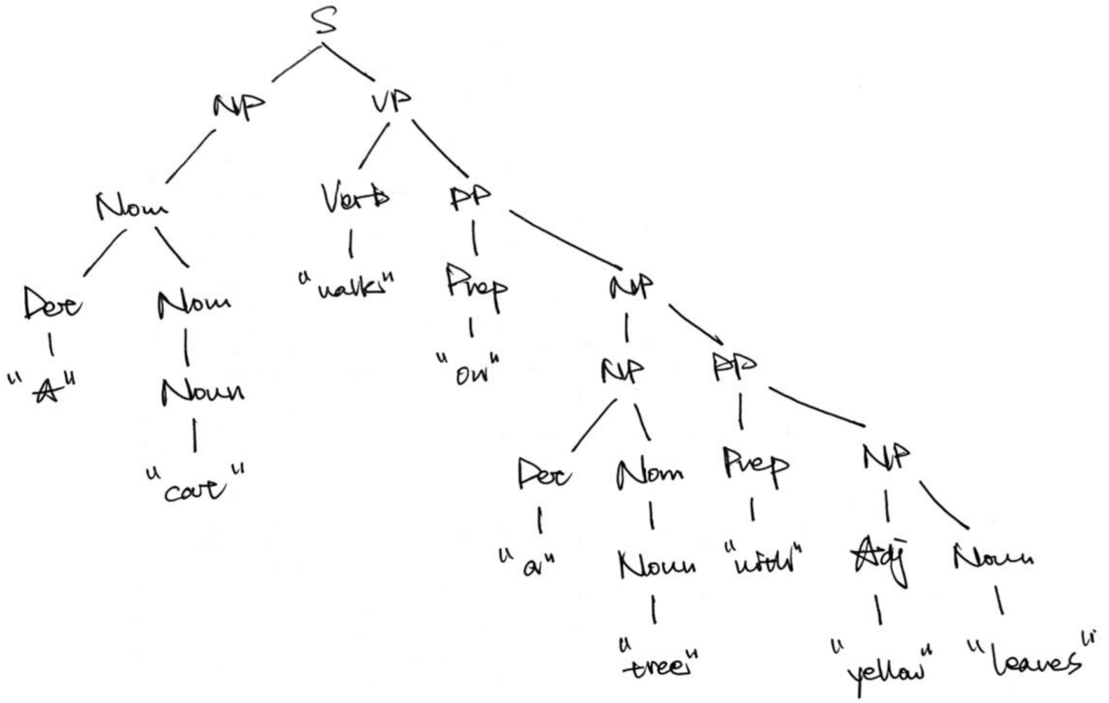
3)

a)

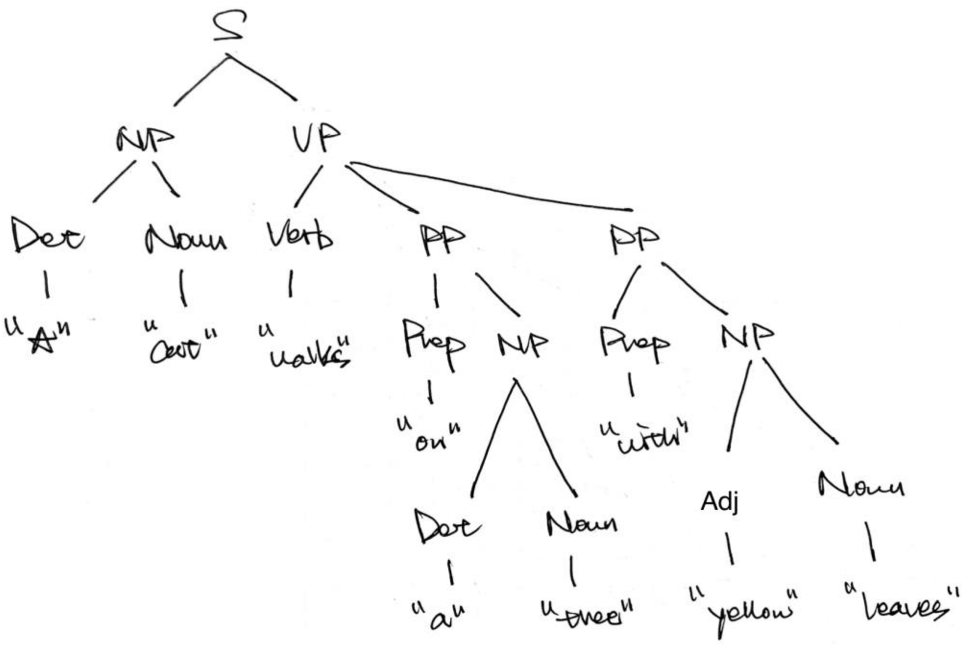


b) **Phase structure**: the constituents are specified, but the grammatical relations are implied
Dependency structure: the grammatical relations are specified, but the constituents are implied, a subtree corresponds to a constituent and the syntactic tag is implied by the subtree root.

c)



(1) PP "with yellow leaves" attaches to NP "a tree"

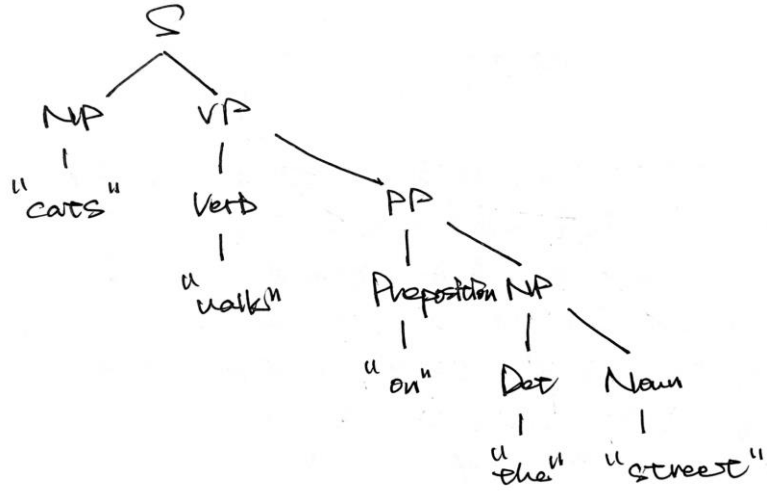


(2) PP "with yellow leaves" attaches to Verb "walk"

Two phrase structures of the sentence "A cat walks on a tree with yellow leaves" differs by the preposition phrase attachment. The first structure attaches "with yellow leaves" to the noun "tree" but the second attaches it to the verb "walk". This preposition phrase is more likely to be attached to tree and therefore the first phrase structure is more reasonable.

d)

Cats	**walk**	**on**	**the**	**street**
NP, Noun				S
	Noun, Verb			VP, NP
		Preposition		PP
			Det	NP
				Noun



e) $P(\text{Cats walk on the street}) = 0.8 \times 0.03 \times 0.2 \times 0.3 \times 0.5 \times 0.2 \times 0.6 \times 0.4 \times 0.2$
$= \mathbf{6.91 \times 10^{-6}}$

4)

a) **Information retrieval**: The input of information retrieval is a query for some information from the user, some search is performed, and the output is the requested information.
Information extraction: The input of information extraction is some desired categories, names of persons, organization and locations for example, and the output is the words requested. One application of information extraction is Named Entity Recognition.
Sentiment analysis: The input of sentiment analysis is a sentence expressing some opinions, the output from sentiment analysis is the polarity of the sentence, typically positive or negative.

b) The word sense **disambiguation** is the process of determining the proper meaning of the word. For example, the word "tank" has two word senses: "the tank is full of soldiers", where "tank" is a vehicle, and "the tank is full of nitrogen", where "tank" is a container.
The supervised learning is the process of training model with labelled data. The model learns from a large collection of the sentences and the word senses to determine the word senses given the context. After the training phase, the model could be applied to unseen sentence and determine the proper word sense of the word in the sentence.

c) The supervised learning model learns from the sentences and computes the probabilities of a sequence of the words being a valid name. After the training phase, the model could be applied to unseen sentence and determine the valid name sequence.
The **states** of the HMM model are the start boundary, the word component of the name, and the end boundary. The transition probabilities between the states are the probabilities where the name starts, continues and ends.
The **observations** of the HMM model are the specific words in the sequence. The observation probabilities are the sequences of a word being the valid name.

d) **Classification**: The classification model learns from the features of the labelled text to determine the class of that text. After the training phase, the model could be applied to unseen text to determine the class label of that.

Clustering analysis: The clustering analysis model learns from the characters of the unlabeled text to determine the similarity among the text and cluster based on that. After the training phase, the model could assign a cluster to the unseen text.

--End of Answers--