Department of Data Science, IIT Palakkad

DS5601: Natural Language Processing (Jan-May 2025)

08:30-09:30 **Mid Semester (07 March 2025)** **Marks : 20**

Instructions

1. Write your answers neatly in Blue/Black ink. Make sure your answers are legible.

2. If you have to make any assumptions about unspecified things, write them clearly with justification.

3. Write the question number clearly for each answer.

4. There will be partial markings for the questions, so even if you cannot solve the entire problem be sincere with the steps.

5. **Be precise**.

1. Write a regular expression to identify all the words in a text that start and end with a vowel. (2)

2. Identify "all" the entities, their types (you are free to define entities), and co-references from the following discourse: (5)

S0: I was traveling to New Delhi on the Rajdhani Express.

S1: The train connects Kerala’s capital, Thiruvananthapuram to India’s capital, New Delhi

S2: I saw the movie Delhi 6 on the train.

S3: Wikipedia says that the number 6 in Delhi 6 refers to the PIN code of the Chandni Chowk area of Old Delhi, a shortened form of 110006.

S4: A co-passenger asked me where am I going to get down.

S5: I said New Delhi railway station or as they simply say NDLS.

S6: After watching the movie, I read the Times of India.

S7: The Times said that New Delhi needs to improve its relations with Washington DC.

S8: Another news was ‘‘India signs an export deal with Australia’’.

S9: On the sports page, I read ‘‘India defeats Australia in a semi-final of Champions Trophy’’.

S10: The newspaper has too many ads.

S11: Next time I will take an Air India flight.

S12: Air India was acquired by the Tata Group.

S13: Yes...it returned to the group after 7 decades.

3. In a classification task, we know beforehand that some words clearly represent some classes. How can we use this information while training a naive Bayes classifier that will result in better classification performance? (3)

4. A 12th standard science student asks you "How does ChatGPT understand the meaning?". How will you answer the question? (*vague and verbose answers will be penalized.*) (4)

5. Solve the problem "Fan Fiction" described on the next two pages. (6)

(H) Fan Fiction (1/2)

MARY SU.0 is a fan-fiction writing robot. Fan fiction is a fiction written by people using another author's characters. Unfortunately, she's not very good at what she does. MARY writes fan-fiction by reading the text of a book (or series of books) and randomly generating new sentences based on the text. Her latest effort is fan-fiction based on the Harry Potter book series.

MARY SU.0 has a few different methods that she's able to use for generating sentences. The first class of methods are called *n-gram* methods. The simplest of these methods is the *unigram* method. In the unigram method, MARY chooses each token of the sentence completely randomly from the entire vocabulary of the book she read. (A token can also be a punctuation mark.) An example of a sentence generated using this method might look like this:

```
gave spiral the truly poisoned, Neville the shoulder Invisibility
```

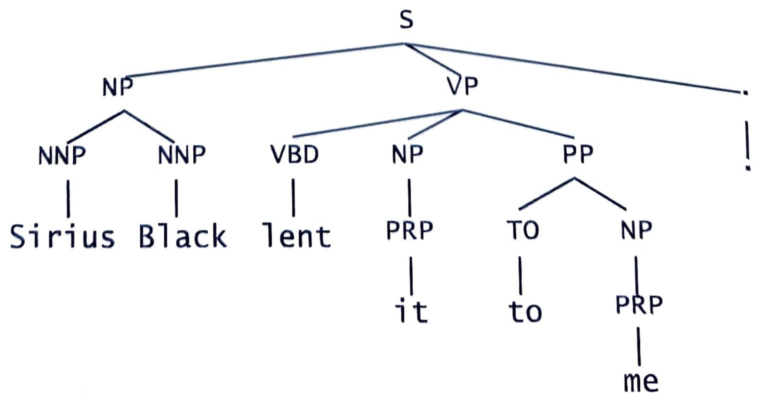
A second method is the *bigram* method. In this method, MARY first finds all the tokens that were used to start a sentence in the text and randomly chooses one of these to start the sentence. Then she builds the rest of the sentence by looking at the most recent token generated, finding all tokens that occur immediately after that token in the text, and randomly choosing one of these. For example, if the most recently generated token was "red", MARY would find all the tokens in the text that immediately follow "red", {"hair", "curtains", "as", ...} and randomly choose one of these to be the next word. A sentence generated using the bigram method might look like this:

```
Face your nose noisily after you saying stuff.
```

A third method is called the *trigram* method. This method is very similar to the bigram method, but uses the previous two tokens (instead of the previous one) to decide what the next token will be. A sentence generated using the trigram method might look like this:

```
But Harry hardly noticed that six extra chairs."
```

The last method that MARY can use to generate sentences is called the *Context Free* method. This method starts by taking each sentence in the text and generating a grammar tree, like the one below, for it.



(H) Fan Fiction (2/2)

The symbols that aren't words refer to labels of words or larger sequences. Some symbols refer to parts of speech, such as NNP for proper noun, PRP for personal pronoun, VBD for a verb in past tense, and TO for preposition. Other labels refer to sequences of words that form units, such as S for sentence, NP for a noun phrase, a sequence of one or more words that behaves like a noun (e.g. *dogs* or *the big dogs*), and VP for a verb phrase, which is a sequence of one or more words that behaves like a verb (e.g. *goes* or *went to the store*).

To generate a new sentence, she first generates an "S" which represents a sentence. Then she looks through her collection of grammar trees for all the sets of symbols ([NP VP .] for example) that occur immediately under an "S". She then repeats this process recursively for each of the new items generated until the tree has no more nodes that can be expanded (once a token is generated, it cannot be expanded). A sentence generated by this method might look like this:

`The next question will cast by Ron.`

H1. Below is a collection of sentences. Two of them are real sentences from the Harry Potter series. The rest were generated using one of the methods above; each method generated at least two sentences. Write either "u" for unigram, "b" for bigram, "t" for trigram, or "c" for context-free to indicate the method that most likely generated that sentence, or if you think the sentence was not automatically generated, write "r" for real.

a. Headmaster uninjured could that was Malfoy that badges
b. He bent over top of the water blushing furiously.
c. There were crouching in your bedroom.
d. He lived about a hundred wizards were closing.
e. Ron spooned iron bolts, keyholes, and a heavy wooden breadboard on to her back and picked up a fistful.
f. "What?" said Harry.
g. ‘Sorry!’ he said," said Mr. Malfoy's eyes.
h. Harry wasn't," said Dumbledore went slightly surprised.
i. years beginning at to annoyance spider!" just months Harry
j. You might have been an impostor.
k. They'll be the first to rise up in the Invisibility Cloak on," said Professor Flitwick pressed a box into his bag.
l. The broom gave them an enormous wink.