Department of Data Science
IIT Palakkad
CS5624 : Natural Language Processing

09:30-12:30 **End-Semester Examination (09 May 2022)** **Marks : 30**

Instructions

1. Write your answers neatly in Blue/ Black ink. Do not use pencil / Red ink. If your answer is not legible, you will not get any marks for that.

2. Doubts and questions will not be answered during the exam. If you have to make any assumption about unspecified things, write the assumption clearly with justification.

3. Answer all parts of a question together. If the parts of a single question are not together, then only the first part will be evaluated. Other parts will not get any marks.

4. Write your name and ID number at the top of the answer sheet. Save the pdf with the following naming convention: [roll_no]_end_sem_22.pdf and upload to the designated assignment in LMS. Do not email.

5. Write question number clearly for each answer. Draw a line after the answer.

6. No hard or soft material are permitted for consultation during the exam.

7. There will be partial markings for the questions, so even if you are not able to solve the entire problem be sincere with the steps.

8. **Be precise**.

9. **There are total 8 questions in this question paper**

1. Read the following excerpt of a news article published in "The Hindu" on 29th March 2022[1] (4)

Parliamentary panel grills Facebook on 'politically motivated' algorithm

by *Sobhana K. Nair and Yuthika Bhargava*
New Delhi: MARCH 29, 2022

The U.S.-headquartered firm was also asked about claims that hate content is rewarded on the platform, lack of language experts.

The Parliamentary Standing Committee on Information and Technology on Monday questioned Facebook officials over allegations that the algorithm used for its advertising platform unfairly promotes one political party in the country. The U.S.-headquartered social media giant was also questioned over claims that hate content is rewarded on the platform as well as the company's lack of Indian language experts for quality check.

The committee met on Monday on the subject —'Safeguarding citizens' rights and prevention of misuse of social/online news media platforms, including special emphasis on women security in the digital space'.

In their defence, Facebook representatives present at the meeting have said the algorithm has been designed to decide which advertisements do better than the others and there is no intervention by the company in this process, a source said. Facebook also informed the panel that the algorithm does not differentiate between political and non-political ads.

...
...
...

Facebook has been facing allegations of bias and inability to curb hate content on its platform in India for nearly two years.

...
...

-END-

[1] https://www.thehindu.com/news/national/house-panel-grills-facebook-on-politically-motivated-algorithm/article65271707.ece

Explain briefly which of the following entities should be held accountable to the government institutions for hate content on Facebook:
(Definition of 'accountable' : *If you are accountable to someone for something that you do, you are responsible for it and must be prepared to justify your actions to that person.*[2])

1. Executives of Facebook
2. Investors of Facebook
3. Facebook's advertisement algorithms
4. Scientists/Researchers/Designers/Developers/Testers of Facebook's advertisement algorithms
5. Advertisers that would benefit from hate speech
6. Users of Facebook
7. Those who do not use Facebook

Expected Answer: Answer to this question is subjective. Evaluation depends on the clarity of your arguments.

2. Linguist Wilhelm von Humboldt describes a language as a system which *"makes infinite use of finite means"*. In the context of NLP, what is the significance of this description? (3)

   Expected Answer: Abstraction (more specifically recursion) at different stages of language processing (from syntax to pragmatics) should be discussed.

3. Compute True Positives (TP), False Positives (FP), and False Negatives (FN) for each of the 4 entity types - PER (Person), ORG (Organization), LOC (Location), and GPE (Geopolitical entity). (3)

   Actual (gold-standard):

   ```
   Air/B-ORG India/I-ORG women/O pilots/O to/O fly/O over/O North/B-LOC Pole/I-LOC
   on/O world/O 's/O longest/O air/O route/O from/O Bengaluru/B-GPE to/O San/B-GPE
   Francisco/I-GPE ./O
   This/O will/O be/O the/O longest/O commercial/O flight/O in/O the/O world/O to/O
   be/O operated/O by/O Air/B-ORG India/I-ORG or/O any/O other/O airline/O in/O
   India/B-GPE ./O
   Air/B-ORG India/I-ORG further/O added/O that/O the/O direct/O distance/O between/O
   the/O two/O cities/O at/O the/O opposite/O end/O of/O the/O world/O is/O 13,993/O
   KM/O ./O
   ```

   Predicted:

[2] https://www.collinsdictionary.com/dictionary/english/accountable

```
Air/B-ORG India/I-ORG women/O pilots/O to/O fly/O over/O North/B-GPE Pole/B-LOC
on/O world/O 's/O longest/O air/O route/O from/O Bengaluru/B-GPE to/O San/B-PER
Francisco/I-PER ./O
This/O will/O be/O the/O longest/O commercial/O flight/O in/O the/O world/O to/O
be/O operated/O by/O Air/B-ORG India/I-ORG or/O any/O other/O airline/O in/O
India/B-GPE ./O
Air/O India/B-GPE further/O added/O that/O the/O direct/O distance/O between/O
the/O two/O cities/O at/O the/O opposite/O end/O of/O the/O world/B-LOC is/O
13,993/O KM/O ./O
```

Answer: Please read the Wikipedia article: *Inside–outside–beginning (tagging)* (https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging))

- **Person**
  - True positives: -
  - False positives: $\{\text{San Francisco}\}_{S1}$
  - False negatives: -
- **Organization**
  - True positives: $\{\text{Air India}\}_{S1}$, $\{\text{Air India}\}_{S2}$
  - False positives: -
  - False negatives: $\{\text{Air India}\}_{S3}$
- **Location**
  - True positives: -
  - False positives: $\{\text{Pole}\}_{S1}$, $\{\text{world}\}_{S3}$
  - False negatives: $\{\text{North Pole}\}_{S1}$
- **GPE**
  - True positives: $\{\text{Bengaluru}\}_{S1}$, $\{\text{India}\}_{S2}$
  - False positives: $\{\text{North}\}_{S1}$, $\{\text{India}\}_{S3}$
  - False negatives: $\{\text{San Francisco}\}_{S1}$

4. Rate the difficulty of each of the following machine translation exercises on a scale of 1 (very easy) to 4 (very difficult). (5)
   Briefly explain your ratings.
   (YNL represents 'your native language')

   1. Matrimonial advertisements (source: English, target: YNL)
      Expected Answer: 2
   2. Proceedings of Supreme court judgments (source: English, target: YNL)
      Expected Answer: 3 or 4 (adequacy is critical)
   3. Jokes (source: French, target: YNL)
      Expected Answer: 3 or 4
   4. Jokes (source: YNL, target: French)
      Expected Answer: 4
   5. Science fiction story (source: English, target: YNL)
      Expected Answer: 4

   Expected Answer: Evaluation depends on the clarity of your arguments.
   Translation of a joke *from YNL to French* is more difficult than translation of a joke *from French to YNL*. Think which translation is likely to be more fluent.

5. Assume that you have a large collection of documents related to k classes. Only some of the documents are labelled, and the rest are unlabelled. How can you use both the labelled and unlabelled documents to train a k-class classifier? (3)

   Expected Answer: Discussion on naive Bayes and Expectation-Maximization algorithm or any other semi-supervised learning approach is expected. This question is discussed in the reading assignment "The Naive Bayes Model, Maximum-Likelihood Estimation, and the EM Algorithm by Michael Collins".

6. Identify "all" the entities, their types and co-references from the following discourse: (4)

   ```
   S1:  I was traveling to Delhi on the Rajdhani Express.
   S2:  I saw the movie Delhi 6 on the train.
   S3:  Wikipedia says, 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 was saying that Delhi has not done enough to improve its relations
   with Beijing.
   S8:  Another news was ''India signs an export deal with Australia''.
   S9:  On the sports page, I read ''India defeats Australia in the boxing day test''.
   ```

```
S10:  The newspaper has too many ads.
S11:  The Rajdhani express takes a lot of time to reach India's Rajdhani from Kerala's
Rajdhani.  hahaha!!!
S12:  Next time I will take an Air India flight.
S13:  Air India now belongs to the Tatas.
S14:  Yes...it returned to the Tata Group after 7 decades.
```

Expected Answer: Entity types are not restricted to PERSON, LOCATION, ORGANIZATION, etc. I asked you to define your own entities. *Some* of the most confusing entities/co-references are listed below:

- *Delhi* in *Delhi 6* is not a location or city in the sentences S2 and S3
- In S7, *Delhi* (*Beijing*) represents the government of India (China) and not a city or location.
  Example: *Delhi, Beijing sign BRICS statement against terror*
  NEW DELHI : At the BRICS foreign ministers' meeting on Friday, India and China, notwithstanding the prevailing tension along the LAC, signed a joint statement on countering international terror and managing conflicts through dialogue. Delhi and Beijing utilised the meeting as confidence-building.....[3]
  In the following sentence 'White House' represents an official authority of the US government, and not a building[4].
  US Casualties From Covid Crosses One Million, Says White House
- In S8, *India* (*Australia*) represents the government of India (Australia)
- In S9, *India* (*Australia*) represents the cricket team of India (Australia) not the government of India (Australia)
- *Delhi* in S7 co-refers to *India* in S8 because they refer to the government of India.
- *India* in S8 **does not** co-refer to *India* in S9 as they have different entity types.
- *Delhi* in S1 **does not** co-refer to *Delhi* in S7. In S1, *Delhi* represents a city while in S7, *Delhi* represents the government of India.
- In S11, entities *India's Rajdhani* and *Kerala's Rajdhani* cannot be split into parts. Note the 's.
  Let us assume Gita is the only sister of Ram. In the sentence *S: Shyam loves Ram's sister*, *Shyam* loves *Gita* and not *Ram*.
  Kerala's Rajdhani refers to *Thiruvananthapuram* , so *Kerala's Rajdhani* represents one entity.
  India's Rajdhani co-refers to *Delhi* in S1.

7. How has your understanding of NLP changed as a result of this course? (2)

[3] https://economictimes.indiatimes.com/news/defence/delhi-beijing-sign-brics-statement-against-terror/articleshow/77942481.cms?from=mdr

[4] https://www.ndtv.com/world-news/us-covid-numbers-us-casualties-from-covid-crosses-one-million-says-white

Answer: Answer to this question is subjective. I was interested to know your views about the course.

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

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
```

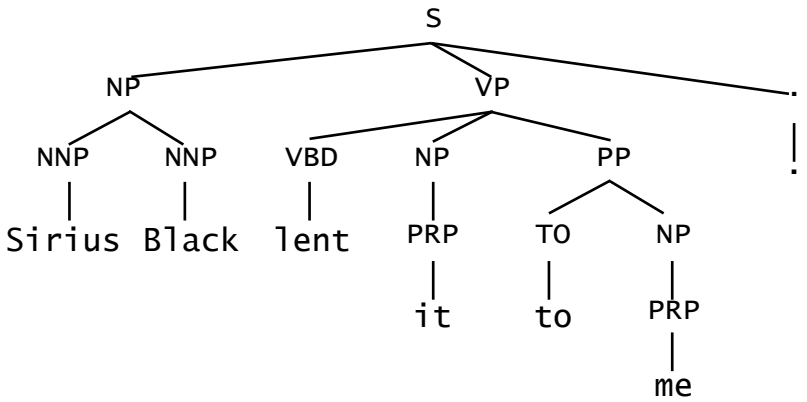
A second method is the *bigram* method. In this method, MARY first finds all the tokens that were used to start a sentence in the text and randomly chooses one of these to start the sentence. Then she builds the rest of the sentence by looking at the most recent token generated, finding all tokens that 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.



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.

a. ☐ b. ☐ c. ☐ d. ☐ e. ☐ f. ☐ g. ☐ h. ☐

i. ☐ j. ☐ k. ☐ l. ☐

Answer: Similar to Physics, Chemistry and Mathematics Olympiads, International Linguistics Olympiad is conducted. This problem (and the Z's Law problem in Quiz-1) was asked in North American Computational Linguistics Open Competition (NACLO[5]). You should explore the NACLO website for more resources related to the Linguistics Olympiad.
The solution to this problem is:

a. unigram
b. context-free
c. bigram
d. trigram
e. context-free
f. real
g. trigram
h. bigram
i. unigram
j. real
k. trigram
l. context-free

[5] https://nacloweb.org/