Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

NAME: **Bhuvi Ghosh** BATCH: **D22** SAP ID: **60009210191**

COURSE CODE: **DJ19DSL602** YEAR**: 2023-2024** DATE: **31/4/24**

COURSE NAME: **Computational Linguistics Laboratory** CLASS: **T.Y.B.Tech (D2)**

Experiment 5

Aim: - Implement CKY algorithm for Parsing the given string and Generate recursive set of sentences using Context Free Grammar

Theory:

Context Free Grammar

CFGS the backbone of many formal models of the syntax of natural language (and, for that matter, of computer languages). Context Free Grammar play a role in many computational applications, including grammar checking, semantic interpretation, dialogue understanding, and machine translation. It express sophisticated relations among the words in a sentence, yet computationally tractable enough that efficient algorithms exist for parsing sentences with them

Syntactic constituency

Syntactic constituency is the idea that **groups of words can behave as single units**, or constituents. For example, group of words come together to form a Noun Phrase. Part of developing a grammar involves building an inventory of the constituents in the language. Consider noun phrase the noun phrase, a sequence of words surrounding at least one noun. Here are some examples of noun phrases.

Harry the Horse	a high-class spot such as Mindy's
the Broadway coppers	the reason he comes into the Hot Box
they	three parties from Brooklyn

Context Free Grammar

It is formal system for modeling constituent structure in English and other natural languages is the Context-Free Grammar, or CFG.It is also called Phrase-Structure Grammars. A context-free grammar consists of a set of rules or productions. Which expresses the ways that symbols of the language can be grouped and ordered together, and a lexicon of words and symbols. Following is the Example of production rules of - CFG

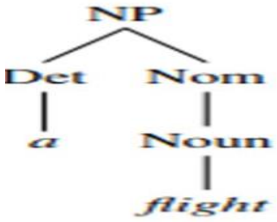
$$NP \rightarrow Det\ Nominal$$
$$NP \rightarrow ProperNoun$$
$$Nominal \rightarrow Noun \mid Nominal\ Noun$$

Context-free rules can be hierarchically embedded, so we can combine the previous rules with others, like the following, that express facts about the lexicon:

$$\begin{aligned} Det &\rightarrow a \\ Det &\rightarrow the \\ Noun &\rightarrow flight \end{aligned}$$

CFG as a generator

A CFG can be used to generate a set of strings. This sequence of rule expansions is called a derivation of the string of words. It is common to represent a derivation by a parse tree (commonly shown inverted with the root at the top). Figure below "A parse tree for "a flight"" shows the tree representation of this derivation



Formal Definition of Context-Free Grammar

A context-free grammar G is defined by four parameters: N, Σ, R, S (technically this is a "4-tuple").

N	a set of **non-terminal symbols** (or **variables**)
Σ	a set of **terminal symbols** (disjoint from N)
R	a set of **rules** or productions, each of the form $A \rightarrow \beta$,
	where A is a non-terminal,
	β is a string of symbols from the infinite set of strings $(\Sigma \cup N)^*$
S	a designated **start symbol** and a member of N

Constituency parsing

The problem of mapping from a string of words to its parse tree is called syntactic parsing; Syntactic parsing is the task of assigning a syntactic structure to a sentence. Parse trees can be used in applications such as grammar checking sentence that cannot be parsed may have grammatical errors (or at least be hard to read). Parse trees can be an intermediate stage of representation for semantic analysis .
The grammar for L0, with example phrases for each rule

Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	→	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles

Example of lexicons

Noun	→	*flights | breeze | trip | morning*
Verb	→	*is | prefer | like | need | want | fly*
Adjective	→	*cheapest | non-stop | first | latest | other | direct*
Pronoun	→	*me | I | you | it*
Proper-Noun	→	*Alaska | Baltimore | Los Angeles | Chicago | United | American*
Determiner	→	*the | a | an | this | these | that*
Preposition	→	*from | to | on | near*
Conjunction	→	*and | or | but*

Cocke-Younger-Kasami Algorithm

It is used to solves the membership problem using a dynamic programming approach. The algorithm is based on the principle that the solution to problem [i, j] can constructed from solution to sub problem [i, k] and solution to sub problem [k, j]. The algorithm requires the Grammar G to be in Chomsky Normal Form (CNF). Note that any Context-Free Grammar can be systematically converted to CNF. This restriction is employed so that each problem can only be divided into two sub problems and not more – to bound the time complexity.

How does the CYK Algorithm work?

For a string of length **N**, construct a table **T** of size **N x N**. Each cell in the table **T[i, j]** is the set of all constituents that can produce the substring spanning from position i to j. The process involves filling the table with the solutions to the sub problems encountered in the bottom-up parsing process. Therefore, cells will be filled from left to right and bottom to top.

	1	2	3	4	5
1	[1, 1]	[1, 2]	[1, 3]	[1, 4]	[1, 5]
2		[2, 2]	[2, 3]	[2, 4]	[2, 5]
3			[3, 3]	[3, 4]	[3, 5]
4				[4, 4]	[4, 5]
5					[5, 5]

In T[i, j], the row number i denotes the start index and the column number j denotes the end index. The algorithm considers every possible subsequence of letters and adds **K** to T[i, j] if the sequence of letters starting from i to j can be generated from the non-terminal K. For subsequence of length 2 and greater, it considers every possible partition of the subsequence into two parts, and checks if there is a rule of the form A -> BC in the grammar where B and C can generate the two parts respectively, based on already existing entries in T. The sentence can be produced by the grammar only if the entire string is matched by the start symbol, i.e, if S is a member of T[1, n].

Example: -

Consider a sample grammar in Chomsky Normal Form

NP --> Det | Nom
Nom --> AP | Nom
AP --> Adv | A
Det --> a | an
Adv --> very | extremely
AP --> heavy | orange | tall
A --> heavy | orange | tall | muscular
Nom --> book | orange | man

Now consider the phrase, “a very heavy orange book“:

a(1) very(2) heavy (3) orange(4) book(5)

Let us start filling up the table from left to right and bottom to top, according to the rules described above:

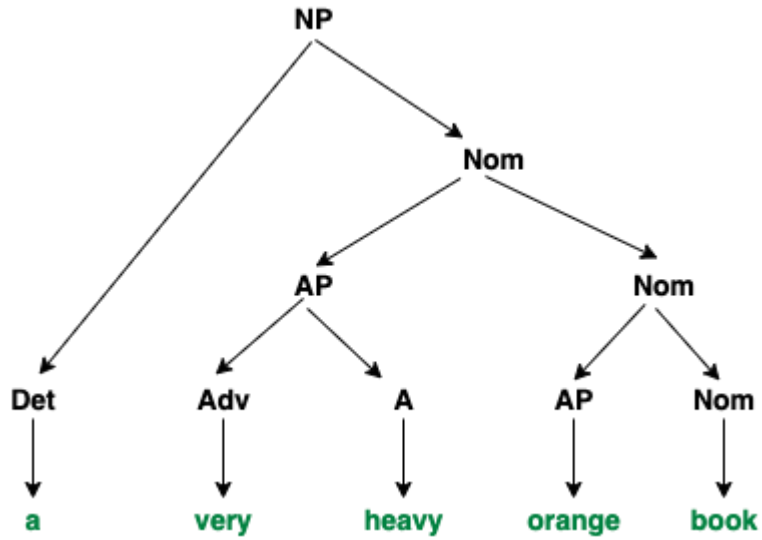
	1 a	2 very	3 heavy	4 orange	5 book
1 a	Det	–	–	NP	NP
2 very		Adv	AP	Nom	Nom
3 heavy			A, AP	Nom	Nom
4 orange				Nom, A, AP	Nom
5 book					Nom

The table is filled in the following manner:

1. T [1, 1] = {Det} as Det –> a is one of the rules of the grammar.
2. T [2, 2] = {Adv} as Adv –> very is one of the rules of the grammar.
3. T [1, 2] = {} as no matching rule is observed.
4. T [3, 3] = {A, AP} as A –> very and AP –> very are rules of the grammar.
5. T [2, 3] = {AP} as AP –> Adv (T [2, 2]) A (T [3, 3]) is a rule of the grammar.
6. T [1, 3] = {} as no matching rule is observed.
7. T [4, 4] = {Nom, A, AP} as Nom –> orange and A –> orange and AP –> orange are rules of the grammar.
8. T [3, 4] = {Nom} as Nom –> AP (T [3, 3]) Nom (T [3, 4]) is a rule of the grammar.
9. T [2, 4] = {Nom} as Nom –> AP (T [2, 3]) Nom (T [4, 4]) is a rule of the grammar.
10. T [1, 4] = {NP} as NP –> Det (T [1, 1]) Nom (T [2, 4]) is a rule of the grammar.
11. T [5, 5] = {Nom} as Nom –> book is a rule of the grammar.
12. T [4, 5] = {Nom} as Nom –> AP (T [4, 4]) Nom (T [5, 5]) is a rule of the grammar.
13. T [3, 5] = {Nom} as Nom –> AP (T [3, 3]) Nom (T [4, 5]) is a rule of the grammar.
14. T [2, 5] = {Nom} as Nom –> AP (T [2, 3]) Nom (T [4, 5]) is a rule of the grammar.
15. T [1, 5] = {NP} as NP –> Det (T [1, 1]) Nom (T [2, 5]) is a rule of the grammar.

We see that T [1][5] has **NP**, the start symbol, which means that this phrase is a member of the language of the grammar *G*.

The parse tree of this phrase would look like this:



Lab Experiment to be performed in this Session: -

Write a program to generate sentences using Context free grammar.
Write a program to implement CKY algorithm to parse a given string.

```
from collections import defaultdict
```

```
def parse(words, grammar):
    n = len(words)
    dp = defaultdict(list)
    for i, word in enumerate(words):
        for X in grammar:
            if word in grammar[X]:
                dp[i, i].append(X)
    for l in range(2, n + 1):
        for i in range(n - l + 1):
            j = i + l
            for k in range(i + 1, j):
                for X, Y in [(X, Y) for X in dp[i, k] for Y in dp[k, j]]:
                    if (X, Y) in grammar:
                        dp[i, j].extend(grammar[(X, Y)])
    start_symbol = next(iter(grammar))
    return start_symbol in dp[0, n - 1]
```

```
Start coding or generate with AI.
```

```
def generate_sentences(grammar, start_symbol, max_length=10):
    sentences = []
    def generate(symbol, sentence):
        if len(sentence) > max_length:
            return
        if symbol in grammar:
            for production in grammar[symbol]:
                if isinstance(production, str):
                    new_sentence = sentence + [production]
                    sentences.append(' '.join(new_sentence))
                else:
                    for symbol in production:
                        generate(symbol, sentence + [])
    generate(start_symbol, [])
    return sentences
```

```
grammar = {
    'S': [('NP', 'VP')],
    'NP': [('Det', 'N'), ('Pronoun')],
    'VP': [('V', 'NP'), ('V')],
    'Det': ['the', 'a'],
    'N': ['book', 'flight', 'meal', 'money'],
    'Pronoun': ['she', 'he', 'it'],
    'V': ['book', 'include', 'prefer']
}
```

```
grammar = {
     "NP": [["Det", "Nom"]],
     "Nom": [["AP", "Nom"], ["book"],
             ["orange"], ["man"]],
     "AP": [["Adv", "A"], ["heavy"],
            ["orange"], ["tall"]],
     "Det": [["a"]],
     "Adv": [["very"], ["extremely"]],
     "A": [["heavy"], ["orange"], ["tall"],
           ["muscular"]]
    }
```

```
words = ['a', 'very', 'heavy', 'orange', 'book']
if parse(words, grammar):
    print(f"The string '{' '.join(words)}' is in the language generated by the grammar.")
else:
    print(f"The string '{' '.join(words)}' is not in the language generated by the grammar.")
```

```
 The string 'a very heavy orange book' is not in the language generated by the grammar.
```

```
start_symbol = 'S'
sentences = generate_sentences(grammar, start_symbol)
print("Generated sentences:")
for sentence in sentences:
    print(sentence)
```

```
Generated sentences:
the
a
book
flight
meal
money
Pronoun
book
include
prefer
the
a
book
flight
meal
money
Pronoun
V
```