AUM

American University Of The Middle East

CE 368 Data Structures

Section : F1-F2-F3-F4-F5-M1-M2-M3

Encoding/Decoding Using Burrows–Wheeler transform

Course Project (30%)

Project Deliverable 1 (15%)-W11

Project Deliverable 2 (15%)-W14

Semester: Spring 2024

2023- 2024

Contents

I. INTRODUCTION

Encoding is process of converting data from one format into another by putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage. Decoding is the opposite process which is used to convert an encoded message back into the original sequence of characters. In addition, Encoding/Decoding algorithms are generally useful to compress the data in which there are frequently occurring characters. There are many different encoding schemes or methods that can be used to encode data. Some schemes include: Burrows–Wheeler transform, Huffman Code, Run-length Encoding, etc. In this Project, we will be exploring and implementing Burrows–Wheeler transform. We use ASCII (which stands for American Standard Code for Information Interchange) coding to represent characters. In ASCII coding, every character is represented with 8 bits. **(See Appendix A)**

II. LEARNING OUTCOMES

1. Make an informed choice of implementation method for the solution of a given problem. Implement and document a structured program to meet a given specification. (1, 2)
2. Select and apply appropriate data structures and algorithms to a given problem. (1, 2)
3. An ability to apply knowledge of mathematics, science, and engineering. (1)
4. An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution. (1)
5. An ability to use current techniques, skills, and tools necessary for computing practice. (1, 2)

III. PROJECT FILES

- Project Description File.

IV. PROJECT SUMMARY

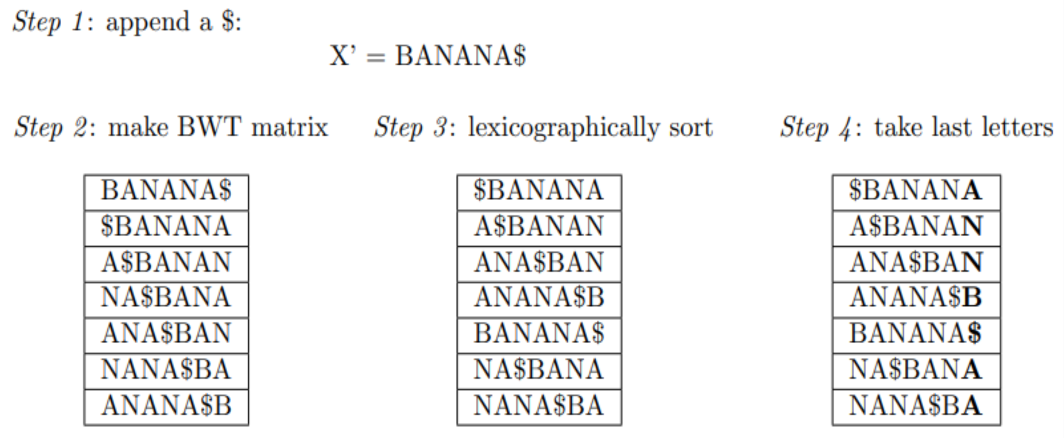
Burrows–Wheeler Transform (BWT) was invented by Michael Burrows and David Wheeler in 1994 while Burrows was working at DEC Systems Research Center in Palo Alto, California. BWT algorithm works by reading a sequence of symbols, grouping them into a specialized format for efficient transmission or storage, and decoding them by converting the encoded message back into the original sequence of characters. **BWT(S)** is a permutation of the letters of S and an end of file character.

The pseudocode below shows the BWT steps:

```
Function FindBWT(S)
    Define the Stop Character.
    Define X as "S + the end of file character"
    Construct the BWT matrix of X by listing of all possible distinct rotations of X
    Sort the BWT matrix lexicographically (Lexical Order)
    Take last column (Last letters)
    Return (BWT(S))
Function InverseBWT (string s)
    Repeat length(s) times
        Insert s as a column of table before first column of the table
        Sort rows of the table alphabetically
    Once Done, Then, the row with the "end of file" character at the end is the original text.
    Return (row that ends with the 'EOF' character)
```

Example-1-:

Initial String S: BNANANA. Construct BWT(BANANA) and the EOF character is $



Step 1: append a $:

X' = BANANA$

Step 2: make BWT matrix	*Step 3*: lexicographically sort	*Step 4*: take last letters
BANANA$	$BANANA	$BANAN**A**
$BANANA	A$BANAN	A$BANA**N**
A$BANAN	ANA$BAN	ANA$BA**N**
NA$BANA	ANANA$B	ANANA$**B**
ANA$BAN	BANANA$	BANANA**$**
NANA$BA	NA$BANA	NA$BAN**A**
ANANA$B	NANA$BA	NANA$B**A**

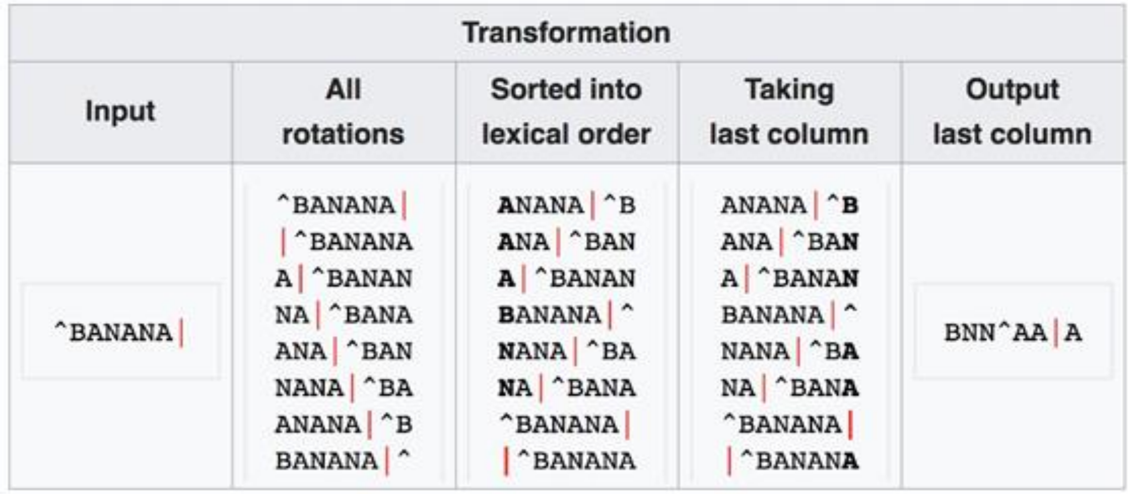
Table 1

The Encoded message is: ANNB$AA

Example-2-:

Initial String S: ^BANANA and the EOF Character is |

Construct the BWT matrix of S:



Transformation				
Input	All rotations	Sorted into lexical order	Taking last column	Output last column
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	ANANA|^B ANA|^BAN A|^BANAN BANANA|^ NANA|^BA NA|^BANA ^BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|**^** NANA|^B**A** NA|^BAN**A** ^BANANA**|** |^BANAN**A**	BNN^AA|A

Table 2

The Encoded message is: BNN^AA|A

Reconstructing S from BWT(S) for example 1:

Inverse transformation							
Input: ANNB$AA							
Append	Sort	Append	Sort	Append	Sort	…	…
A	$	A$	$B	A$B	$BA	…	…
N	A	NA	A$	NA$	A$B	…	…
N	A	NA	AN	NAN	ANA	…	…
B	A	BA	AN	BAN	ANA	…	…
$	B	$B	BA	$BA	BAN	…	…
A	N	AN	NA	ANA	NA$	…	…
A	N	AN	NA	ANA	NAN	…	…

Table 3

Continuing in this manner, you can reconstruct the entire list. Then, the row with the "end of file" character at the end is the original text.

Read Me:

The program implements **the tree nodes data structures**.

The C program must be menu driven. Upon running the program, the user is given a menu with the functions that can be performed. The user should be able to encode/decode any string. The program must adhere to the project requirements below.

- The C program should work correctly without any error.
- Functions should take into consideration all special cases.
- The program should be well documented and proper comments must be provided.
- The program should not fail for a typical input and minor special cases (proper testing should be done before submitting the full code).
- Students should be prepared for answering any of the questions related to all that have been covered in the project (this includes both theoretical and practical parts).

Deliverable 1-(100 points)   15% (Week 11):

- Students of each class need to form project groups **(Group of 3 students)**
- Deliverable 1 must be submitted through Moodle.
- **Do not try to copy exact flowchart from the websites. Do not try to copy from other groups. Do not use ChatGPT.**
- In Deliverable 1, A Power Point Presentation (PPT) with **voice over** must be **prepared** and submitted **(Recording & Adding your Audio to slides in your PowerPoint. All Students should participate. Do not exceed 5 minutes)**. The PPT should include the following:

- Using the above pseudocode and explanation, draw a flowchart for the BWT algorithm along with a description of the different stages you have.
- Test the BWT algorithm for the strings in appendix B. Show and explain the results and conclude. **Solve this question by hand and Do not write a C Code. The C code is only required for Deliverable 2.**
- Show the corresponding tables **with all entries** (table 2, and table 3) for each testing string. Show and explain the results. **Solve this question by hand and Do not write a C Code. The C code is only required for Deliverable 2.**
- Find the number of Sort and Append operations for each testing string. Show and explain the results and conclude. **Solve this question by hand and Do not write a C Code. The C code is only required for Deliverable 2.**

Deliverable 2-(100 points) 15% (Week 14):

- Students need to form project groups **(Same Group of students as in deliverable 1).**
- **In Deliverable two, a scientific report must be written and an interview will be held.**
- **Do not try to copy exact code from the websites. Do not try to copy from other groups. Do not use ChatGPT**

Scientific Report [7.5%]:

- You have to answer the following:
    - Write a C code to implement the BWT algorithm (Encoding/decoding). The program must be menu driven. Upon running the program, the user is given a menu with the functions that can be performed. **The user should be able to encode/decode any string.**
    - Test your code for the strings in appendix C. Show the result and conclude.
    - Show the corresponding tables **with all entries** (table 2, table 3) for each testing string.
    - Find the Encoded message for each testing string and conclude and the number of sort and append operations for each testing string and conclude.
    - Test your code for the strings in appendix B. Compare results from PD1 with results from PD2 and conclude.

- The scientific report must be **written** and submitted. **The report must be submitted through Moodle** and should include the following:
    - Full analysis of the above questions
    - The written C code along with comments must be included in the report. (**Source code and comments** to describe your code).

- Screen shots of the all outputs must be provided in the report. The screen shots should demonstrate the functionality of the program showing all menus and showing the correct execution of all the functions.

Interview/demo [7.5%]:

- An interview will be held for PD1 and PD2. Grading will be individual based on answering questions regarding the program code and all details related to the project including PD1 and PD2. **In other words, a group of 3 people can have three different grades based on the oral exam.**
- The written C code must not include comments during the interview/demo. (**Source code with no comments** to describe your code).

V. DELIVERABLES & PROJECT MANAGEMENT

Due Date	Task	How to Submit
Week 11	**Deliverable 1 (15 %)**	**Submit power pointpresentation with voice over through Moodle.**
Week 14	**Deliverable 2 (15 %)**	**Submit scientific report (7.5%) + Oral Demo (7.5%)**

VI. Plagiarism

- Upon suspicion and doubt of the authenticity of the work submitted, the instructor has the right to ask the student to verify her/his work. This can be done through, but not limited to, oral examination or discussion, or any other action deemed necessary. If the student fails to prove the authenticity of the work, then the Instructor will apply the academic misconduct rules as mentioned in the AUM Student Handbook which may include awarding the work a zero grade.

- For a detailed description of academic misconduct please refer to the undergraduate AUM Student Handbook.

VII. Copyrights

Students are expected to adhere to copyright practices, ***refer to the undergraduate AUM Student Handbook.***

VIII. Project and team-based work

The Project component of the course, if exist, is essential to passing this course. The project shows competency in understanding and applying the course objectives and achieving the learning outcomes. The project should allow the student to investigate, apply, research, and practice real-life business situations. It is expected that each student to fully and actively participate in the project as an effective team member. A project document will be distributed later in the semester with details about the project.

Inclusion in a team is ensuring that every member has a sense of belonging, trust, and support, member can therefore participate fully and authentically in the project. When a team is inclusive, all team members contribute, show interest in the work, feel respected and involved, and are able to bring their own unique strengths and skills to their roles. Inclusivity in a project resolve in a team that perform better, solve problems faster, and achieve more.

How to render your team more inclusive:

1- *Be yourself, and show up at times to your meeting*
2- *Treat all your teammates with dignity, respect, fairness, and without discrimination*
3- *Speak up about your own thoughts and show initiative*
4- *Communicate nicely with the right approach without affronting your teammates*
5- *Accept criticism and Involve and empower your team members in the project*
6- *Challenge stereotypes and biases and accept your teammates and their differences*

Things that hinder your inclusiveness:

1- *Avoiding face-to-face communication and overreliance on email or messaging*
2- *Creating communication barriers based on differences*
3- *Creating biases based on stereotypes and prejudice*
4- *Inconsistent response to mistakes*
5- *Lack of motivation or demotivating your team members*
6- *Leaving team members ideas unheard and disregarding members contribution*

For all group related work, the ***entire team is responsible for the team outcome and the deliverables****, except for the specific parts of the project that may be graded individually depending on the project's requirement and as communicated in the project document.*

IX. Grading Rubrics

PD1 Grading Rubric: (Presentation with voiceover)

	Unsatisfactory (0-59%)	**Developing (60%-74%)**	**Satisfactory (75%-87%)**	**Excellent (>88%)**
Slides Content (20% - Group)	The slides fail to meet any of the following criteria: 1) Information presented is accurate and relevant to the presentation topic 2) Content is well-organized, all key points are covered and clearly articulated. 3) An adequate level of detail is provided to cover and analyse the topic comprehensively	The slides meet only one of the following criteria: 1) Information presented is accurate and relevant to the presentation topic 2) Content is well-organized, all key points are covered and clearly articulated. 3) An adequate level of detail is provided to cover and analyse the topic comprehensively	The slides meet only two of the following criteria: 1) Information presented is accurate and relevant to the presentation topic 2) Content is well-organized, all key points are covered and clearly articulated. 3) An adequate level of detail is provided to cover and analyse the topic comprehensively	The slides meet all of the following criteria: 1) Information presented is accurate and relevant to the presentation topic 2) Content is well-organized, all key points are covered and clearly articulated. 3) An adequate level of detail is provided to cover and analyse the topic comprehensively
Slides Design (10% - Group)	The slides fail to meet any of the following criteria: 1) Video is of very good quality, visually engaging, and balanced in terms of text and figures. 2) Effective use of font and colors to emphasize key points and maintain visual appeal. 3) Consistent use of fonts, colors, and design elements throughout the presentation.	The slides meet only one of the following criteria: 1) Video is of very good quality, visually engaging, and balanced in terms of text and figures. 2) Effective use of font and colors to emphasize key points and maintain visual appeal. 3) Consistent use of fonts, colors, and design elements throughout the presentation.	The slides meet only two of the following criteria: 1) Video is of very good quality, visually engaging, and balanced in terms of text and figures. 2) Effective use of font and colors to emphasize key points and maintain visual appeal. 3) Consistent use of fonts, colors, and design elements throughout the presentation.	The slides meet all of the following criteria: 1) Video is of very good quality, visually engaging, and balanced in terms of text and figures. 2) Effective use of font and colors to emphasize key points and maintain visual appeal. 3) Consistent use of fonts, colors, and design elements throughout the presentation.

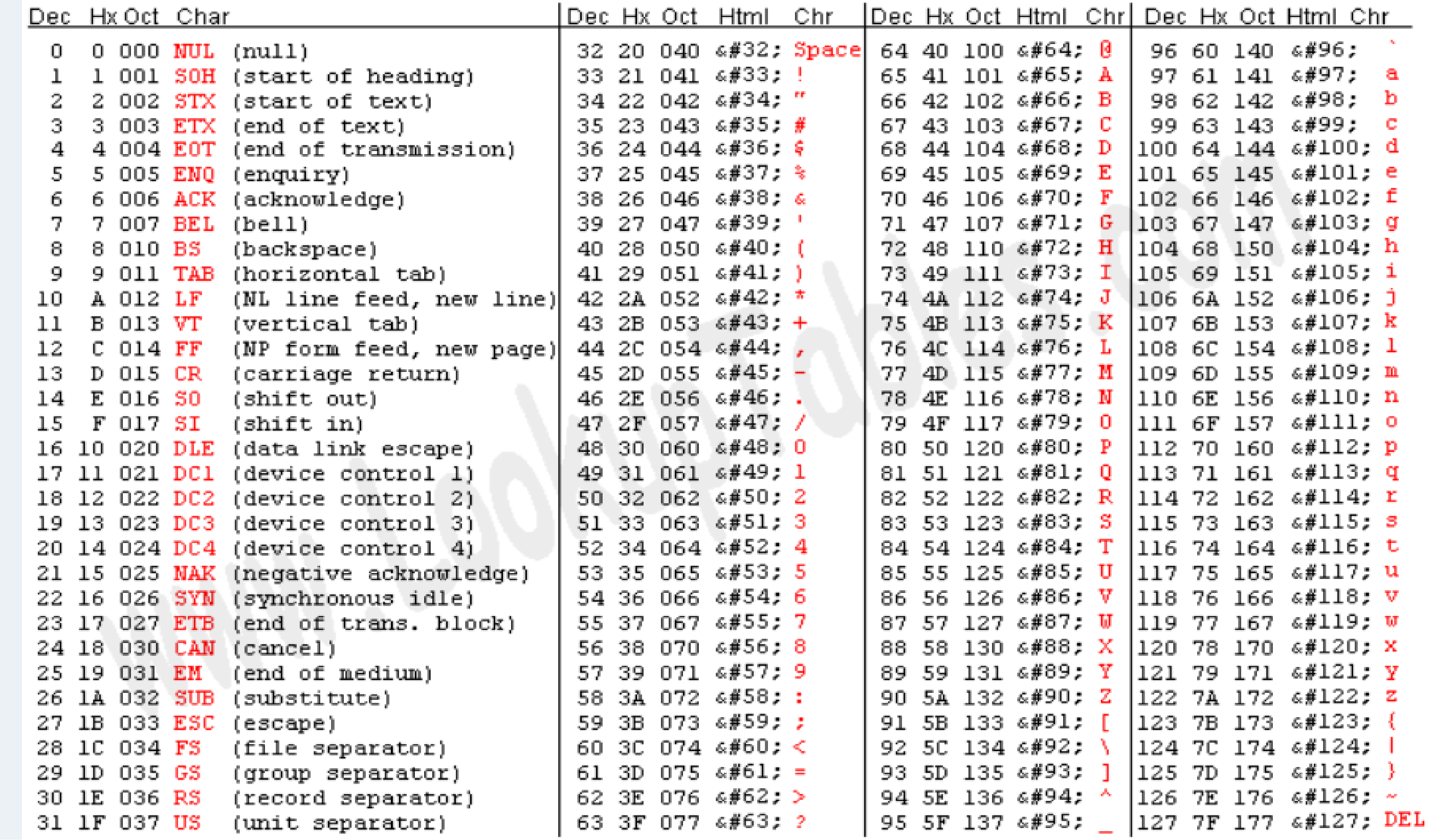
Voiceover Delivery (20% - Individual)	The voiceover fails to meet any of the following criteria: 1) Clear and audible voiceover, appropriate use of English language, and terminologies. 2) The student is explaining, not just reading, the presented information effectively. 3) Effective integration between the voiceover and slide content with minimal stumbles, filler words, or awkward pauses.	The voiceover meets only one of the following criteria: 1) Clear and audible voiceover, appropriate use of English language, and terminologies. 2) The student is explaining, not just reading, the presented information effectively. 3) Effective integration between the voiceover and slide content with minimal stumbles, filler words, or awkward pauses.	The voiceover meets only two of the following criteria: 1) Clear and audible voiceover, appropriate use of English language, and terminologies. 2) The student is explaining, not just reading, the presented information effectively. 3) Effective integration between the voiceover and slide content with minimal stumbles, filler words, or awkward pauses.	The voiceover meets all of the following criteria: 1) Clear and audible voiceover, appropriate use of English language, and terminologies. 2) The student is explaining, not just reading, the presented information effectively. 3) Effective integration between the voiceover and slide content with minimal stumbles, filler words, or awkward pauses.
Voiceover Delivery (10% - Group)	The team voiceover fails to meet any of the following criteria: 1) The time allocation between team members is fair and provides equal opportunities for each member to contribute 2) The presentation is delivered within the exact provided time. 3) Smooth transitions between team members' segments.	The team voiceover meets only one of the following criteria: 1) The time allocation between team members is fair and provides equal opportunities for each member to contribute 2) The presentation is delivered within the exact provided time. 3) Smooth transitions between team members' segments	The team voiceover meets only two of the following criteria: 1) The time allocation between team members is fair and provides equal opportunities for each member to contribute 2) The presentation is delivered within the exact provided time. 3) Smooth transitions between team members' segments	The team voiceover meets all of the following criteria: 1) The time allocation between team members is fair and provides equal opportunities for each member to contribute 2) The presentation is delivered within the exact provided time. 3) Smooth transitions between team members' segments
Proposed Design(s) (40% - group)	The proposed design(s) fails to meet any of the following criteria: 1) Well-developed and detailed with no mistakes 2) Clear and understandable explanation of the design concept 3) Adequate consideration of relevant and realistic constraints	The proposed design(s) meets only one of the following criteria: 1) Well-developed and detailed with no mistakes 2) Clear and understandable explanation of the design concept 3) Adequate consideration of relevant and realistic constraints	The proposed design(s) meets only two of the following criteria: 1) Well-developed and detailed with no mistakes 2) Clear and understandable explanation of the design concept 3) Adequate consideration of relevant and realistic constraints	The proposed design(s) meets all of the following criteria: 1) Well-developed and detailed with no mistakes 2) Clear and understandable explanation of the design concept 3) Adequate consideration of relevant and realistic constraints

PD2 Grading Rubric: (Prototype with Q&A)

	Unsatisfactory (0-59%)	**Developing (60%-74%)**	**Satisfactory (75%-87%)**	**Excellent (>88%)**
Proposed Design/Prototype (50% - group)	1) The design is not well-developed through the needed approaches or may be incomplete or has numerous mistakes. 2) Prototype/model/simulation is poorly executed, lacking functionality or demonstrating significant flaws 3) Prototype/model/simulation results are not or incorrectly analyzed	1) The design is partially developed through the needed approaches , has some mistakes, but may lack detail or cohesiveness 2) Prototype/model/simulation shows some functionality, but improvements are needed to enhance its quality and effectiveness 3) Prototype/model/simulation results are partially analyzed while not showing students understanding of the results	1) The design well-developed through the needed approaches and detailed with minimal mistakes 2) Prototype/model/simulation is of good quality, demonstrating functionality and providing a realistic representation of the proposed design 3) Prototype/model/simulation results are briefly analyzed and discussed showing students understanding of the results	1) The design is exceptionally well-developed and detailed through the needed approaches and has no mistakes, with a clear and cohesive design that demonstrates an exceptional level of thought and creativity 2) Prototype/model/simulation is of exceptional quality, accurately representing the proposed design and showcasing advanced functionality. 3) Prototype/model/simulation results are carefully analyzed and discussed
Questions and Answers (50% - Individual)	1) Provides incorrect or incomplete answers. 2) Does not exhibit any confidence in answering questions.	1)Provides partially correct or vague answers. 2) Shows limited confidence in answering questions and often appears hesitant/unsure.	1) Provides mostly correct answers with minor mistakes or limited details. 2) Shows confidence in answering questions but occasionally appears hesitant/unsure.	1) Provides correct and detailed answers. 2) Consistently demonstrates a high level of confidence while answering questions.

Appendices

Appendix A: ASCII Table



Dec	Hx	Oct	Char		Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL	(null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH	(start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX	(start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX	(end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT	(end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ	(enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK	(acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL	(bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS	(backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB	(horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF	(NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT	(vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF	(NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR	(carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO	(shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI	(shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE	(data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1	(device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2	(device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3	(device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4	(device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK	(negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN	(synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB	(end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN	(cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM	(end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB	(substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC	(escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS	(file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS	(group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS	(record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US	(unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Appendix B: Strings for Testing Using BWT

S1: acaacgc with end of file character $

S2: ^BANANA with end of file character $

S3: ^BANANA with end of file character |

S4: ^mississippi with end of file character |

Appendix C: Strings for Testing Using BWT

S5: ^BANANANANA with end of file character $

S6: SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES with end of file character $

S7: ^SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES with end of file character |

S8: ARARBRBRBRABAABAABAAAA with end of file character $