Machine Learning Final

Name _____

Student ID number _____

Directions: For this test, you will need to self-partition the class into 6 teams, with each team having an approximately equal number of members. At the scheduled day and time of the final exam, your team will present your results to the class. Each team will have a 2-part exam as discussed below.

1. ($66\frac{2}{3}$ points)

   Number your teams from 1 to 6. Each team is assigned the following machine learning technique.

team	1	2	3	4	5	6
ML	HMM	SVM	k-means	k-NN	RF	ANN

   Using your assigned ML technique, determine how well you can distinguish between standard English text of length 136 as compared to "word salad" text (i.e., text that is composed of randomly selected words), also of length 136. You are only to consider 26 symbols, lower-case a through z (no spaces). For this to be a fair comparison, every team must use the same text corpus, and the "word salad" text can only contain words found in your English text corpus. You must analyze 2 different features (you may also considered the combined features) and all teams must use the same 2 features. Your team can consider more than 2 features, but the same 2 "core" features must be considered by all teams. One obvious feature is letters, and there are many other possible features to consider.

   This part of the exam is cooperative, in the sense that anyone on any team can help any other team. Everyone in the class will receive the same grade for this part of the test, based on the average performance of the 6 teams. It is therefore in your best interest to make sure that your team does a thorough analysis, and that all of the other teams also do a good job.

2. ($33\frac{1}{3}$ points) This part of the exam is competitive between the teams—the teams will be evaluated and scored independent of each other. Prior to the final exam presentations, you are *not* to discuss *any* of your work on this part of the test with *anyone* outside of your own team. You are free to use any technique to attempt to solve this problem.

   A solution to the Zodiac 340 cipher (Z340) was recently proposed. This putative solution treats most of the cipher as filler. Here, we will concentrate on the first 8 lines (136 symbols) of the cipher, which makes up the vast majority of the proposed solution. The "solution" is based on the assumption that the Z340 is a homophonic substitution, with the modification that "reverse homophones" occur. A reverse homophone is a ciphertext symbols that can decrypt to more than one plaintext symbol.

   For this part of the exam, your goal is to construct a putative solution to the first 8 lines of the Z340 cipher that is "better" than the putative solution mentioned above (ideally, we would like to construct many such solutions). The way that we will quantify "goodness" is based on how many reverse homophones you use—the fewer reverse homophones, the better. Equivalently, solutions that come closer to solving the homophonic substitution problem are better. There are many ways to view the problem of constructing such solutions; here we consider it as a type of constrained optimization.

   Denote a putative solution as $(p_0, p_1, p_2, \ldots, p_{135})$, where each $p_i \in \{\mathrm{a}, \mathrm{b}, \ldots, \mathrm{z}\}$. Table 1 lists the constraints, when viewing the cipher as a homophonic substitution. To read this table, we begin at $i = 0$ and continue to $i = 135$. If row i has a "—" in the "match" column, the ciphertext symbol in that position has not yet appeared and hence there is no constraint on p_i, which implies that any letter can be selected. On the other hand, if a number (or numbers) appear in the match column of row i, then p_i must match all of the specified positions. For example, at $i = 91$, the match column is $62, 23$, which means that for this to be a valid solution to the homophonic substitution, we must have $p_{91} = p_{62} = p_{23}$. In other words, the ciphertext symbol is the same at all of these positions.

   However, if we allow for reverse homophones, the problem changes from "and" to "or." In this case, a "—" still means that there is no constraint. But for the constrained positions, ciphertext symbols might not have been assigned uniquely to plaintext letters, and hence we can meet a constraint by matching any one of the specified positions. For example, again considering $i = 91$, the constraint is met if either $p_{91} = p_{62}$ or $p_{91} = p_{23}$. If p_{91} fails to match either of these, then it has failed to meet the constraint, and we will view it as a new reverse homophone. This is the version of the problem that you will consider.

From Table 1, we see that there are 60 positions with no constraint and 76 constrained positions. Given any putative solution $(p_0, p_1, p_2, \ldots, p_{135})$, we can count the number of these 76 constraints that are satisfied using the method outlined in the previous paragraph.

The "word salad" 136-letter text below was generated by selecting words from the Zodiac's vocabulary (including misspellings), as found in the file `ZodiacWords.txt`, without any consideration of grammar. This particular example satisfies 61 of the 76 constraints.

```
HER WHRITE NAME SPOT THINGAMAJIG NINE CONTROL IT I DIG YEAR
HIT THE I RUN TUT MI ARE OIL CIRCUT ARE EATS NOT UP HER MINT
ONE WONT A RILE ABOT OK MT ARE AT DOO I RAM A IT GRAVEL
```

Note that spaces are not part of the plaintext and have only been included for readability.

As another example, the pseudo-English 136-letter text below was generated using letter 7-grams (with an overlap of 6 consecutive letters) based on the file `alice_oz.txt`. This example satisfies 33 of the 76 constraints.

```
ALICE HAD GIVEN IT TO HER CHILD FOR HE HAD BROUGHT AS THE
COURAGE HE GAVE HER HERE IN THE PRIZES BUT WHO THERE IN THE
DOOR IT WAS HIGH ADDED THAT DOROTHY DID NOT HURT YOU OH
```

A randomly selected 136-letter segment of English text would be expected to satisfy only about 5 constraints.

Any solution that satisfies 48 or more of the 76 constraints will be clearly better (by our measure) than the proposed solution mentioned above.[1] However, it is not enough to simply satisfy the constraints, as your solution must also be somewhat "English-like." For example, if we simply map every ciphertext symbol to "`I`" then all of the constraints are satisfied. Even the Zodiac is not likely to have encrypted "`I`" 136 consecutive times—and if he did, there would be no hope of decrypting it. But, it is plausible that the plaintext may be only semi-grammatical, and misspellings are common in any Zodiac writing.

An ideal solution would accept a dictionary as input, and produce semi-grammatical solutions with at least 48 of the 76 constraints satisfied. Even more ideally, this would work with a relatively small dictionary of Zodiac-like words, such as `ZodiacWords.txt`.

[1] The proposed solution actually only satisfies 46 of the constraints, and also includes several additional degrees of freedom that we do not account for in our approach. Therefore, any solution that satisfies 48 of the constraints as specified here will be unquestionably superior to the proposed solution.

Table 1: Constraints

i	match	i	match	i	match	i	match
0	—	34	20	68	—	102	47,22
1	—	35	—	69	—	103	77
2	—	36	—	70	38	104	80,71,64,63,39,19
3	—	37	—	71	64,63,39,19	105	66,31
4	—	38	—	72	48	106	—
5	—	39	19	73	—	107	24
6	—	40	—	74	—	108	—
7	—	41	—	75	16	109	99,67,15
8	—	42	14	76	10	110	40
9	—	43	26	77	—	111	37
10	—	44	21	78	—	112	—
11	—	45	33	79	8	113	42,14
12	—	46	12	80	71,64,63,39,19	114	59,7
13	—	47	22	81	—	115	28
14	—	48	—	82	—	116	72,48
15	—	49	0	83	9	117	46,12
16	—	50	—	84	—	118	76,10
17	—	51	—	85	61,53,52,18,4	119	101,44,21
18	4	52	18,4	86	57	120	113,42,14
19	—	53	52,18,4	87	65,2	121	109,99,67,15
20	—	54	—	88	55,6	122	50
21	—	55	6	89	78	123	32
22	—	56	5	90	56,5	124	74
23	—	57	—	91	62,23	125	102,47,22
24	—	58	30	92	—	126	91,62,23
25	—	59	7	93	58,30	127	104,80,71,64,63,39,19
26	—	60	—	94	75,16	128	68
27	—	61	53,52,18,4	95	—	129	17
28	—	62	23	96	83,9	130	27
29	—	63	39,19	97	89,78	131	116,72,48
30	—	64	63,39,19	98	3	132	127,104,80,71,64,63,39,19
31	—	65	2	99	67,15	133	—
32	—	66	31	100	25	134	117,46,12
33	—	67	15	101	44,21	135	69