NLP: **Exam 1**

Total points - 100

Due: Monday, October 2, 2023

* No late submission allowed. (remember this is exam not assignment)
* Failure to submit on time will automatically **grade zero**.
* No coding allowed.

1. [Points 10] Answer the following MCQ questions.
   1. [**points 5**] Which of the followings are correct for the given regular expression [\w.-][+@[\w.-](mailto:+@[\w.-)]+
      1. al.ice@yahoo.com
      2. bob.alice-com
      3. 1.Abc@kennesaw.edu1
      4. @ksu.edu
      5. All of the above

Correct expressions are:

[i. al.ice@yahoo.com](mailto:i.%09al.ice@yahoo.com)

iii. 1.Abc@kennesaw.edu1

* 1. Which of the followings are correct? Given regular expression ^[^A-Z0-6]+
     1. ^abc 10
     2. abc9
     3. 1 "Hello"
     4. ^abc9
     5. None of the above

Correct expressions are:

i[i. abc9](mailto:i.%09al.ice@yahoo.com)

iv. ^abc9

1. [**points 10**] Differences between Naïve Bayes and logistic regression classifier. Can we use naïve bayes to generate text? If so, how we can generate text, explain the details, and provide examples.
2. [**points 15**]
   1. What is sequence labeling? How would you build your parts of speech baseline model? [**points 3**]
   2. What are name entities? What is entity recognition task? [**point 3**]
   3. Why is Name Entity Recognition (NER) a difficult task? [**point 3]**
   4. Design a NER model using Hidden Markov Model and discuss transition probability, emission probability and sequence likelihood estimation with examples. **[point 6**]
3. [**Points 15**] Please formulate your digit-language model for the following text. Consider each digit as a single word to formulate your language model. Show the details of your LM formulation. Please follow lecture slides to answer these questions.

**Training Text:** “112324567890104455668892100228899774410102997744333536373839300102030304050607”

Test text: 1234321567890

1. Unigram language model [Points 8]
2. Compute perplexity of your model [Point 7]
3. [**points 35**] Assume that we are in an alien world and their languages are different and only contains the following vocabulary [a, b, c, d]. Their NER tags are given as [X,Y, O, O].

Tags are assigned as followed:

‘a’ 🡪 X,

‘b’ 🡪 Y

‘d’🡪 O

‘c’🡪 O

If multiple similar sequence exists, the tags should follow BIO notation. E.g., for a given sequence ‘a a a’, the tag sequence would be B-X, I-X, I-X; similarly, ‘b b’ sequence, tag will be B-Y, I-Y; ‘d c’ sequence tag sequence would look like B-O, I-O etc.

Given the following sentences as training examples.

Sentence 1: a a a a b b b a a b c b c d a b a d b d d c d b

Sentence 2: b b b c c c d d c d c b c a c b c b c a c a b a d

Sentence 3: d b d b d b d c d a d a d c d d d b d c

Sentence 4: c d c b c b c a c c a c b c d c a a a b

Sentence 5: c d c d d d b b a a b a d

**Test Sentence**: a b a c d d b b d a d b

You are given a task to assign NER Tag using Hidden Markov Model for the given test sentence. Please answer the following questions. Hints: follow lecture slides step by step.

* 1. Please calculate transition probabilities [**points 10**]
  2. Calculate emission probabilities [**points 10**]
  3. How many possible tag sequences (paths) can be generated for the given test sentence, show detailed computations? [**points 5**]
  4. Assign possible tags for the given test sentence following maximum likelihood calculation. Please show details likelihood calculation for all the optimized tag sequences (follow Viterbi algorithm approach and show at least two sequence computation). [**points 10**]

1. [**points 15**] A restaurant chain wants to see whether customers like their foods or not. They hire you as an NLP scientist to do this task. You crawled their website, collect and annotate all the reviews. Your task is to conduct the experiment for this text data.
   1. How many classes would you identify for this food review and why? Please list down the classes name as your preferences. [**points 3**]
   2. You train two classifiers – (i) logistic regression/softmax classifier, and (ii) a two-layer feed forward neural network. Please mention differences between these two classifiers in restaurant review classification. [**points 5**]
   3. Now you are deciding to pick only one classifier for the deployment environment. How would you design your experimental setup to evaluate these two classifiers? How would you decide to pick which classifier for the deployment environment? Please mention the details reason behind your decision. [**points 7**]

**Submission Instructions:**

**Important.**

No coding allowed. You must have to write answer for the given question.

**You should not zip your submission. – you may submit individual files –for question answering.**

**Late submission or Extension:** No late submisison allowed. Late submission will not be accepted and automatically graded zero.

**Grading Policy/Rule:** Copying/cheating/plagiarism is strictly prohibited as mentioned in our introductory lectures and syllabus. This policy holds for each assignment/homework/exam. In case of copying/cheating/plagiarism etc. you will be graded zero for the assignment as well as ‘F’ for the subject. Note that the first incident of cheating will result in the student getting a final grade of ‘F’ for the course. The second incident, by CCSE rules, will result in a semester suspension from the College.