Mid-Term Exam

Total Points 100

Due: Tuesday, October 5, 2021

1. [**points 5**] Which of the following are correct? Given a regular expression [^A-Z|abc]+
    a. NLP is an interesting topics
    b. Regular expressions is easy.
    c. i like nlp
    d. Negation operation is fun
    e. All of the above.
2. [**points 5**] Which of the followings are correct? Given regular expression ksu.*edu
    a. KSU is a great college
    b. Ksu is an Edu
    c. Ksu&Edu
    d. ksu@edu
    e. None of the above
3. [**points 5**] Write down the differences among naïve bayes, logistic regression and softmax classifier.
4. [**points 5**] Write down the differences of different activation functions - sigmoid, tanh, relu.
5. [**points 5**] What is sequence labeling? What is POS tagging? How would you build your parts of speech baseline model? Write down differences between parts-of-speech tagging and Name entity recognition.

6. [**points 7**] Please write down the differences among micro and macro-average for precision, recall and f1 metrics. Please give an example of each of these metrics.

7. [**points 8**] What is word embedding? How does word2vec model work? Please explain how neural language model works while training word embedding together.

8. [**points 20**] Given the following equations.

    $a = 2x – y$
    $b = az$
    $L = a+2b$

    i) Please draw computational graph (circuit diagram) for the given equations above. [**points 5**]
    ii) Show forward pass values on the diagram, for the given values of $x=1$, $y = 4$, $z = -3$. [**points 5**]
    iii) Show a complete backpropagation circuit diagram with corresponding gradient values. [**points 10**]

9. [**points 25**] Please build your character-gram (char-gram) language models for the given training set. Please assume that you experiment will only have the following characters – [a, b, c, d, f, h] that exists in the training set.

Training set:
b a b a d a f f
a c h a d f f a h
f b a a h c f h d d f
a b f f c c d f h
h h a a c a c d d d

Test set:
h d c d f
d b b c c a

Task:

i) Build char-unigram language model [**points 5**]

ii) Build char-bigram language model [**points 5**]

iii) Compute joint probability for the given test set using char-unigram. [**points 5**]

iv) Compute perplexity of your models (char-unigram, char-bigram) for the given test set and compare which model is better for each of these test case. [**points 10**]

10. [**points 15**] Assume that we are in an alien world and their languages are different and only contains vocabulary [*delta, gamma, alpha, beta, sigma, derivative, summation*]. Their given parts-of-speech tags are [*A, B, C, D*]. You are given a task to assign tags using Hidden Markov Model for them. Given the following sentences as training examples.

Training Sentences:

Sentence1: delta gamma sigma summation
Tags: A, B, C, A

Sentence 2: alpha, sigma, beta derivative
Tags: A, C, D, A

Sentence 3: derivative gamma delta beta
Tags: A, B, B, D

Sentence 4: sigma summation beta alpha
Tags: C, B, C, D

Sentence 5: alpha beta sigma derivative
Tags: A B C A

Test Sentence: *gamma beta alpha sigma*

a. Calculate transition probability matrix [**points 7**]
b. Calculate emission probability matrix [**points 8**]