**Homework and Programming assignment 2**

**Total Points - 100**

**Due Date: October 01, 2022**

**1.** You are given sample documents with corresponding classes. Documents are annotated as A, or B. These documents are splitted as training and testing set shown in the below table. Assign the most probable class to the test sentence given below using naïve bayes classification approach. Please mention each step clearly (i.e., prior probability, conditional probability, etc.). [**Points 25**]

|  |  |  |  |
| --- | --- | --- | --- |
| Doc # | words | Class |  |
| d1 | Chinese Beijing Chinese | B | Training |
| d2 | Chinese Chinese Shanghai | B | Training |
| d3 | Tokyo Japan Chinese | A | Training |
| d4 | Chinese Macao | B | Training |
| **d5** | **Chinese Chinese Chinese Tokyo Japan** | ? | **Testing** |

**2.** What is cross-validation? Would you please give an example and explain how cross-validation work? When to use cross-validation during an experiment? [**Points 15**]

**3.** Explain how gradient descent algorithm works? Please explain the effect of learning rate on the learning algorithm while updating parameters using the following equations.

wt+1 = wt -

Please show differences among different types of gradient descent – mini-batch, batch, and stochastic gradient descent. [**Points 15**]

**4.** You are given the following text documents. Please answer the following questions. [**Total Points 45**]

**Text:**

*It is going to rain today.*

*Today I am not going outside.*

*NLP is an interesting topic.*

*NLP includes ML, DL topics too.*

*I am going to complete NLP homework, today.*

1. Would you please calculate the TF-IDF vector for each of the tokens? Please show each step (i.e., tokenization, vocabulary, TF, IDF, etc.). [**points 15**]
2. Please calculate term-term co-occurrence matrix with a given context window 3. [**points 15**]
3. Please find the most similar words ( a pair of words) from their vector representations for both TF-IDF and co-occurrence matrix based vector representation cases. To compute most similarity score, you may use cosine similarity score. Show details similarity calculation for both cases – (a) TF-IDF and (b) co-occurrences-based representations. [**points 15**]

**Submission Instructions:**

Please note that rephrasing or rename variables, functions names does not mean you are not copying/cheating. For this similar situation, you will be graded zero.

**You should not zip your submission.**

**Late submission or Extension:** Late HomeWorks/assignment will not be accepted unless an extension is approved by me in advance. Requests for extensions must be made at least three days before the due date with valid reason. **3 points** will be deducted for each day after the submission deadline from your grade even if you are approved for extension. For details, please see the **Homework and Exam Policies** section of your syllabus for more details.

**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.