BINUS University

Academic Career: ***Undergraduate / ~~Master~~ / ~~Doctoral~~ *)***	**Class Program:** ***~~International~~/Regular/~~Smart Program~~/~~Global Class~~*)***
☑ Mid Exam **□ Final Exam** **□ Short Term Exam** **□ Others Exam : ____________**	**Term : ~~Odd~~/Even/~~Short~~ *)**
☑ Kemanggisan **☑ Alam Sutera** **☑ Bekasi** **□ Senayan** **□ Bandung** **□ Malang**	**Academic Year :** **2021 / 2022**
Faculty / Dept. : School of Computer Science	Deadline — Day / Date : Wednesday/ April 27th, 2022 Time : 13:00
Code - Course : COMP6576001 - Natural Language Processing	Class : All Classes
Lecturer : Team	Exam Type : Online
*) *Strikethrough the unnecessary items*	
The penalty for CHEATING is DROP OUT!!!	

Learning Outcomes:

LO1: Describe what is Natural Language Processing and its components
LO2: Explain fundamental concepts of how to work with Natural Language Processing
LO3: Apply Natural Language Processing concepts in certain real-world applications

I. Case Study (100%)

1. **[LO 1,LO 2, LO 3, 25 point]**Jolly is planning to scrape websites to collect dataset for his NLP project. The plan is to search particular keywords using their pattern. Jolly is planning to use Regular Expression in his project, your task is to help jolly to build the Regular Expression strings to extract:
   a. **[LO1, LO2, 5 points]** Explain how would the Regular Expression help Jolly to solve the problem, in your opinion, is there any better solution to help Jolly with his NLP project? Please explain thoroughly!
   b. **[LO3, 5 points]** Indonesian, Singaporean and Malaysian Phone Number pattern (examples: +62215262578, +60358577, +60388888000). The phone number is always started with character "+" and followed by the country code (62 for Indonesia, 65 for Singapore, 60 for Malaysia). The length of the phone number is varied.
   c. **[LO3, 5 points]** A set of email address for academia (examples: jolly.wanda@bee.ac.id, jojo@bee.ac.uk, jolly.wanda@bee.edu, jojo@bee.edu.au).
   d. **[LO3, 10 points]** Finally, implement the Regular Expression from 1b. and 1c. with any programming language, libraries of your choice. Explain the process!

2. **[LO 3, 25 points]** As the programmer in Jojo AI Studio, your next project is to annotate the following sentences with Penn Treebank Tagset:
   a. **[LO3, 5 points]** Social Signal Processing can be considered as blue-sky research in the Artificial Intelligence area.
   b. **[LO3, 5 points]** Please book me a flight to Indonesia. I am planning to visit Jojo Academy this week.
   c. **[LO3, 5 points]** I think therefore I am the first principle of Rene Descartes's philosophy.

Verified by,

Hidayaturrahman (D6423) and sent to Program on April 1st, 2022

d. **[LO3, 10 points]** Finally, implement the 1a, 1b and 1c. with any programming language, libraries of your choice. Explain the process!

3. **[LO 2, LO 3, 25 points]** Jolly as a programmer is planning to build a language model using N-Gram from this following corpus:
<s> Saya suka dengan makanan laut </s>
<s> Saya suka dengan minuman yang manis </s>
<s> Kemarin saya makan makanan laut </s>
<s> Hari ini saya makan makanan khas Sunda </s>
<s> Besok saya berencana makan makanan khas Betawi </s>
<s> Kemarin saya makan roti </s>
<s> Hari ini saya makan Pizza </s>
<s> Besok saya berencana makan Burger </s>
<s> Hari ini saya minum teh tawar </s>
<s> Kemarin saya minum Coca Cola </s>
<s> Besok saya berencana minum kopi </s>
<s> Lusa saya berencana minum teh manis </s>

The <s> token indicates the start of the sentence and </s> indicate the end of the sentence. Your task is to calculate the language model probability (using Bi-Gram) of following:
   a. **[LO2, 5 points]** P(<s> saya suka dengan makanan laut </s>)
   b. **[LO2, 5 points]** P(<s> Besok saya berencana makan Pizza </s>)
   c. **[LO2, 5 points]** P(<s> Saya makan Pizza </s>)
   d. **[LO2, 5 points]** P(<s> Saya minum kopi </s>)
   e. **[LO2, 5 points]** P(<s> saya suka </s>)

4. **[LO 2, LO 3, 25 points]** Using the corpus from problem no 3:
   a. **[LO2, 10 points]** Which features/language model/word representation model (e.g. NGram, TF-IDF, Bag of Words, Word2Vec) that can represent the words well to solve a classification problem? Please explain your opinion comprehensively!
   b. **[LO3, 15 points]** Implement the features/language model/word representation model (e.g. NGram, TF-IDF, Bag of Words, Word2Vec) that you have chosen in 4a. with any programming language, libraries of your choice. Explain the process!

-- Selamat Mengerjakan --

Verified by,

Hidayaturrahman (D6423) and sent to Program on April 1st, 2022