ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐH CÔNG NGHỆ

"KHAI PHÁ DỮ LIỆU HƯỚNG ỨNG DỤNG"
HKII 2018-2019

ĐỀ THI CUỐI KỲ
(90 phút, không sử dụng tài liệu)

Cho 04 văn bản đơn giản sau:

v1: "*Học sâu đang dẫn đầu trong lĩnh vực học máy.*"

v2: "*Triết lý nhân văn của ông đã làm nên một tác phẩm văn học sâu sắc.*"

v3: "*Thuật toán học máy đã tạo ra các tác phẩm văn học.*"

v4: "*Tác phẩm học máy tạo ra vẫn còn cứng nhắc.*"

Câu 1: Giả sử bạn chỉ làm việc với 02 văn bản v1 và v3, hãy tìm biểu diễn của **v1** trong không gian vector sử dụng phương pháp tf-idf?

Câu 2: Bạn đang cần xây dựng 1 mô hình phân lớp văn bản Naïve Bayes có thuộc lớp "**Văn học**" hay không. Giả sử rằng bạn chỉ có 03 văn bản **v1, v2, v3** ở trên làm ví dụ huấn luyện mô hình. Hãy xây dựng mô hình đó và sử dụng nó để quyết định xem văn bản **v4** có thuộc *lớp "Văn học" không?*.

Câu 3: *Trình bày công thức tính và ý nghĩa của các độ đo Precision (Chính xác), Recall/Sensitivity (Hồi tưởng/Nhạy) và F1. Kiểm định chéo k-fold là gì? Hãy trình bày cách thực hiện kiểm định chéo k-fold.*

Hết!