Kuis Data Mining

1. Representasi seperti apa yang dapat digunakan untuk mengolah data berupa teks? Apa yang menjadi kelemahan dari representasi tersebut?
2. Dengan menggunakan Jaccard similarity for bags, hitung nilai similarity dari dua pengguna layanan *streaming* film, A dan B, jika
   a) A memberikan film X 3* dan film Y 1*
   b) B memberikan film X 2*, film Y 2*, dan film Z 1*
3. Dalam kasus Naive Bayes, apa yang harus dilakukan untuk menangani *missing values*? Mengapa seperti itu?

4. Bagaimana proses dilakukannya *pre-pruning* dan *post-pruning* dalam *decision trees*? Mengapa kita melakukan hal ini?
5. Apa yang membedakan antara cara kerja *random forest* dan *decision trees*?
6. Jelaskan apa yang dimaksud dengan prinsip Occam's razor.
7. Dalam *supervised learning*, sebutkan masing-masing kegunaan dari data latih, data validasi, dan data uji.
8. Jelaskan cara kerja n-fold cross validation.