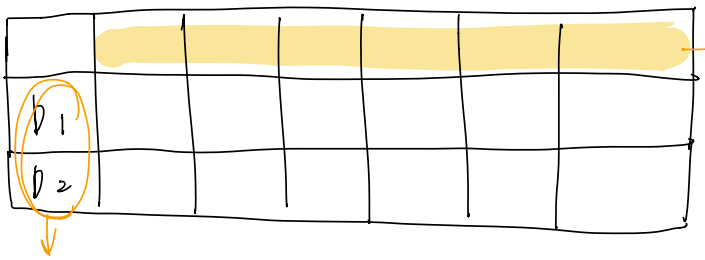
text mining

表是一个矩阵 (matrix)



→ term vector

如果字很多，維度會太大

可透過下2种方法來處理

Document vector = Profile

主成份分析法
奇異值分解法 ⇒ 找特徵值

① words become features

資料 → 以什麼型式的文件讀入 → word segmentation →

去除停用字、保留詞幹 (stemming)

stop word (英) walked, walking 雖形態不同 but same meaning

爬資料 ex: 小五