Literature Review 1(LT#1):
Q1.
一，根據論文所述，我們要解決的問題主要敘述如下:
1. User-based 的巨量數據流的排序演算法，該算法本身包含包回歸模型構成的 model 外(權重 g 表示)，和一個根據使用者不同的變量(權重以 w 表示)
2. 非監督式學習，因為目標是在沒有用戶明確標記的情況下進行排名。
3. 該演算法必須必須可以於分散式系統中運行 (使用 Big Table) ，不影響郵件收發速度，並具備高容錯空間。

二，演算法:

$$s = \sum_{i=1}^{n} f_i g_i + \sum_{i=1}^{n+k} f_i w_i, \qquad p = \frac{1}{1+\exp^{-s}}.$$

1. 簡單線性回歸 (解釋性高、擴展性高) 用來當全域的模型框架，利用遷移學習(把訓練好的 model 參數拿過來 fit 不同 user based preference 的資料)來達成用戶模型
- S 前半部是指由 1 到 n 個特徵 乘上各自變量的權重組成的特徵集，後半部是指全部的特徵再加上額外 k 個 user based 的特徵乘上各自權重(wi) 所構成的 user based model，我們的模型會用後半部已 user 為單位對模型進行快速的調整。
- p 是預測用戶在時間區段 T 中對該信件互動的機率，公式如上，以及大概會納入參考的變量種類敘述如下:

(1). $p = Pr(a \in A, t \in (Tmin, Tmax)|f, s)$
- a: 是對郵件執行的操作，包含收件人對該信件的行動，包含在 A 集合（例如，A ={打開、手動更正、刪除…}）
- t: 代表寄件者寄出到收件者對該信有所行動的時間差，介於域值 T(min , Max)，簡單來說低於 min 的行為不會拿去 fitting, 又 Max 上限減少我們需要納入回歸中考量的變量(以免回歸模型變得太巨大，解釋性低落)
- f,s 為給定之條件機率變數，f 是特徵向量集合(上述)，s 表示 用戶有機會看到郵件(ex, 該時間區段收到 10 封信，該封信被看到的 baseline 是 1/10)
(2). 特徵種類包含:
1. 社交特徵 e.g., 收寄人行為(有向關係)
2. 內容特徵 e.g., 主題術語，於 data preprocessing 階段當作丟到 PA 裡面 tuning classifier

而因為使用者行為會改變，我們要不斷更新 classifier，所以論文中利用線上的算法中的 Passive Aggressive 演算法來調整 weight， 類似是使用一個 slide window 區隔資料，根據新抵達的一批資料是否屬於同一個數據生成分布，對既有模型選擇持續學習或是拋棄，公式不做詳述，但是重點在該模型 C 值可以修正 global model。

在使用者手動標籤上上，對該模型影響很大，簡單來說，若我們標籤的行為與重要郵件排序順位呈現高度正相關，會加深該造成該排序的權重，反之，則會降低。

最後則是巨量數據，同前述，我們使用 bigtable 來進行線上的平行運算，加速排序，而 big table 也就是透過複製郵件所需 訊息，分析後，排序該 user:message-id， 並在多個應用上串聯，同時又不斷刪除完成排序的資料來維持其空間與資安。

Q2.
建立一個 line、message 訊息串聯的訊息管理程式，藉由納入特徵如下

特徵:

- 是否被 tag
- 使用者回復個別聊天室窗的頻率
- 根據聊天內容動態調整該聊天室窗的性質(工作議題/ 閒聊/感情種類)
- 藉由 TFIDF 強化對議題的分類
- 社會網絡分析:

  使用者於各群組間的關聯性，因為時常在同一群組的人可能現實生活關係更緊密，使用者對對方的訊息也更感興趣