# 語料前處理

每個語料都整理成以句子為單位的格式，依據計算需求有三種模式：

1. 斷詞+pos tag

碰到(VC) 這(Nep) 種(Nf) 情形(Na) ，真(D) 是(SHI) 會(D) 令(VL) 人(Na) 又(Caa) 氣憤(VH) 又(Caa) 苦惱(VH) 。

1. 斷詞，無標點符號

['碰到', '這', '種', '情形', '真', '是', '會', '令', '人', '又', '氣憤', '又', '苦惱']

1. 完整句子，含標點符號

碰到這種情形，真是會令人又氣憤又苦惱。

# 參數計算

* 句子長度：算出每一個句子中含有幾個詞彙
* 詞頻：參考 COCT 詞頻資料，以詞頻 1200 為標準，回傳每一個句子中含有幾個高頻詞、幾個低頻詞，還有句子中每一個詞的詞頻
* 詞彙等級：參考國教院詞語分級表，回傳每一個句子中，各級別的詞彙分別有幾個
* 詞彙長度：算出每一個句子中有幾個「長度大於三個字」的詞彙
* 完整的句子：
  + 以。？！結尾
  + 至少含有一個動詞
* 完整的語境：用 `Caa` `Cab` `Cba` `Cbb` 判定
* Black list
  + 連續八個以上數字或數詞
  + 十個以上包含 `.` 的非中文字串 (網址)
* Grey list：用 `Nb` 判定