WSM_Project3 ：

首先，這次的資料大致分為兩種，圖片跟文字，之後主要就這兩個面向討論。

關於實作方向跟目標：

這次project分成書面五十分、報告五十分、跟額外五十分三個部分。

先就額外的部分來說，其實最後排名分數只有二十，重點反而是：

1. 課堂沒教的架構(10分)
2. RNN or LSTM (10分)
3. 視覺化(10分)

由於這些部分，加上我、喻能、跟亮緯做得比較多的都是Deep Learning的部分，我想我們就先從Deep Learning的做法下手，用keras實作，一些課堂教的模型可以在最後一週寫報告時作為比較。

關於圖片

說是說這個課程沒教，不過我、喻能、跟亮緯處理圖片可能比文字更熟，可能反而是我們的優勢。

圖片做法大致分成兩種： 1. 預訓練模型 2. 自己練

而關於圖片和廣告吸引力的關係，可能就要請大家問問認識的廣告系同學或是老師，看看影響因素有哪些(內容物、模糊或清晰、圖片大小etc)

先說預訓練的部分，假如我們發現廣告吸引力的重點在內容物之類的，使用預訓練模型就會非常棒，可以省下大量的訓練時間，結果也會精準得多。

但就要考慮到圖片大小的問題，因為預訓練模型的圖片大小都是固定的，可能需要一些切割或前處理，對結果也有未知的影響。

如果重點是其他的，比如說清晰程度跟大小，那可能就要使用自行訓練的模型，用一些條件判斷的方式(有些kernel有教)，把這些東西量化之後當成input送進去之類的。

所以調查所謂「廣告的吸引力」跟「廣告裡的圖片」有哪些關聯因素就很重要(當然大家如果時間很多我們也可以都做XD)。

另外，因為圖片檔真的瞎雞巴大，我們百分之百需要使用keras generator分段載入、處理跟訓練。

關於文字

嗯，這個我不太熟XD

我們的文字資料包含廣告內文，以及一些其他像地區資料的條件，總之是文字這樣。那不管如何，總之向量化(vectorization，或者說indexing)是很重要的。

一種是整理好直接交給indri開幹，但那個維度應該會高到炸開，training會做到死，改進的方法是送給indri index完，先跑個簡單的Dense model，再用keras的輔助功能篩選重要的權重(關鍵字)，之後就只focus那幾個。

另外就是用word2vec，我是完全不會，聽說index的方法跟詞義有關，比較合理，維度應該也比較小(?)，google應該有API，但是要花時間研究。

還有，上述的方法都是做「英文」的處理，但這次資料是俄文.....。 (老師似乎有請助教找找有沒有相關的俄文套件啦)

進度：

這個我沒有太多意見，我是覺得我們可以先分成三週(下週一開始) 第一週主要完成模型跟處理問題，確保它可以動。 第二週開始校正結果，搞一些奇奇怪怪的東西跟做法（我們有很多人可以問)。 第三週就輕鬆地把報告完成，把圖畫出來之類的。

這週末就先想辦法把資料讀進去(generator)，以及做完前處理(indexing, 圖片裁剪)好了。