D15 Term Frequency - Inverted Document Frequency (TF-IDF)

簡報閱讀 範例與作業 問題討論

重要知識點

- 了解 TF-IDF 的原理與應用。

Word Embedding

前一個章節我們介紹了 bag of words 的原理與實作，我們可以發現，雖然 **Bag of words 能表達單字的存在**，但並不能凸顯每一個單詞對整句話或整段文章的重要性。

TF-IDF

TF-IDF 的全名為 Term Frequency - Inverted Document Frequency，可以看出其實是由兩個部分組成。分別為『 **Term Frequency(詞頻)** 』、『 **Inverted Document Frequency (倒文件頻率)** 』。

舉個例子，假設有一個訓練集包含 1000 個評論，我們要預測評論是正向或負面：

- 其中有一個單詞在**每個評論中幾乎都有出現**，此時這個詞有很大的機會對我們的預測沒有顯著的影響(ex. the)。
- 另外又有一個單詞，他在 1000 個評論中出現的頻率較低，然而卻**在特定幾個評論中頻繁的出現**，此時這個單詞就可能有較高的機率對判定評論的正負面有影響(ex. excellent)。

由上述例子我們可以發現，不同單詞對一句話或一段文章的重要性並不相同，而TF-IDF的出現就是為了凸顯不同單詞的重要性。

詞頻(TF)：

一個單詞出現在一個文件的次數/該文件中所有單詞的數量

因此當一個單詞集中出現在數個評論中時，這個單詞在這些文件中的TF值就會較高。

$$w_{x,y} = \boxed{tf_{x,y}} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

資料來源：用在NLP 的 TF-IDF

IDF：

Log (所有文件的數目/包含這個單詞的文件數目)

IDF： log(N/dfx) = logN - logdfx

當一個單詞集中出現在數個評論中時，此時雖然 logN 值對於所有單詞都是相同的，但這個詞的 logdfx 值會較低，因此 IDF 值也會較高。

$$w_{x,y} = tf_{x,y} \times \boxed{\log\left(\frac{N}{df_x}\right)}$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

資料來源：用在NLP 的 TF-IDF

綜合上述，當一個單詞的 TF * IDF 值越大時，代表**這個單詞對整段文章的重要性也越大**，我們可以歸納出：

- 不同單詞在同一個文章中獲得的 TFIDF 值可能不相同，值的高低代表了單詞對整段文章的重要性。
- 同一個單詞在不同文章所得到的 TFIDF 值也可能不同。

資料來源：用在NLP 的 TF-IDF

TF-IDF 示範程式碼

本次課程有搭配的實作程式碼，示範如何使用 TfidfVectorizer 得到 word embedding 的結果。

補充資料：用在NLP 的 TF-IDF

參考資料

網站：TFIDF原理與實作教學

下一步：閱讀範例與完成作業