D14 文字預處理

簡報閱讀 範例與作業 問題討論

重要知識點

- 了解訓練資料需要經過哪些步驟，才能輸入 NLP 模型。

下方為本日會使用到的技術，部分內容前面章節可能提過，這裡會將前處理所需技巧串起，我們會依照下方順序逐步清理文字資料集。

課程練習

本章節請直接使用 .ipynb 練習，程式碼都有註解和說明。

參考資料

網站：[Text Preprocessing]

Text Preprocessing in Python: Steps, Tools, and Examples

Data Monsters Follow
Oct 16, 2018 · 7 min read

by Olga Davydova, Data Monsters

In this paper, we will talk about the basic steps of text preprocessing. These steps are needed for transferring text from human language to machine-readable format for further processing. We will also discuss text preprocessing tools.

After a text is obtained, we start with text normalization. Text normalization includes:

- converting all letters to lower or upper case
- converting numbers into words or removing numbers
- removing punctuations, accent marks and other diacritics
- removing white spaces
- expanding abbreviations
- removing stop words, sparse terms, and particular words

下一步：閱讀範例與完成作業