數學、語言
以及那些你誤會多年的東西

統計與機率

統計模型：

- Jieba
- CKIP-classic
- TF-IDF

偽. 機率模型：

- CKIP/Stanford CoreNLP
- Transformer models (BERT, GPT)

機率

假設有一公正硬幣，

請問投擲 100 次的「正面」機率為何？

請問投擲 100 次的「反面」機率為何？

Q. 你為什麼沒有硬幣，沒有真的投擲經驗，沒有任何投擲硬幣記錄下來的資料，就能得出前述的「機率」？

假設有一外星硬幣，為不公正 6 面方體。

其中：

1, 6 的面積最大，佔總面積 70%

2, 5 的面積次之，佔總面積 20%

3, 4 的面積最小，佔總面積 10%

請問，投出結果為 1 的機率為多少？

請問，投出結果為 1 之後，再投出 1 的機率是多少？

真. 機率模型

特點:

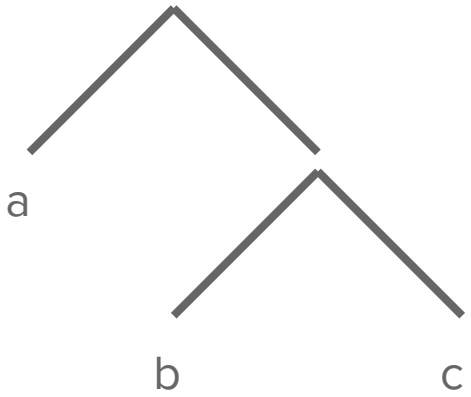
1. 需要對「操作目標」的結構有所瞭解
2. 不需要「操作資料」的記錄即可計算
3. 演算法比資料的統計結果重要！

偽. 機率 (真. 統計)

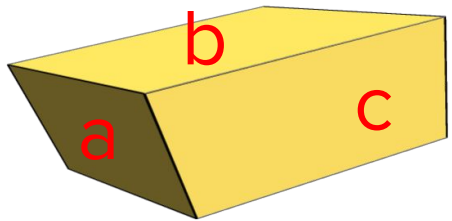
特點:

1. 不需要對「操作目標」的結構有所瞭解
2. 需要「操作資料」的記錄才可計算
3. 資料的統計結果, 被拿來當做下一次操作的演算法

X-bar 結構



外星硬幣再思考



假設有一外星硬幣，為不公正 6 面方體。

其中：

1, 6 的面積最大，佔總面積 50% 為一**平面**

2, 5 的面積次之，佔總面積 34% 且**外凸**

3, 4 的面積最小，佔總面積 16% 為一**斜面**且**外凸**

請問，投出結果為 1 的機率為多少？

結論：

條件愈多，表示愈不容易落入這一面！表示這一面為集合的元素「很少」

a = {a1, a2, a3...ai}

b = {b1, b2, b3, b4, b5, ...bj}

c = {c1, c2, c3, c4, ...ck}

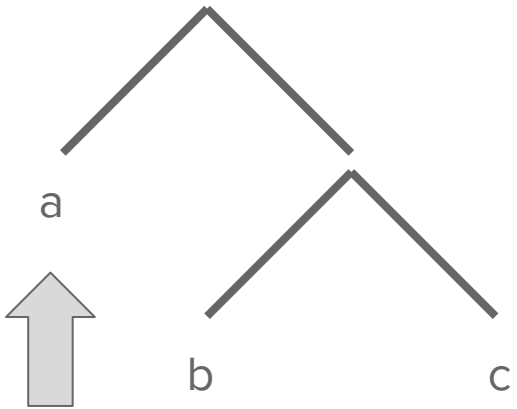
結論:

條件愈多，表示愈不容易落入這一面！表示這一面為集合的元素「很少」

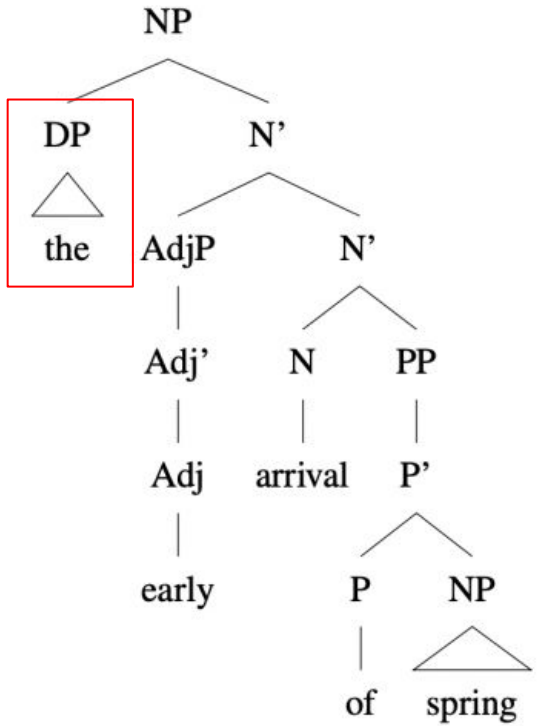
a = {a1, a2, a3...ai}

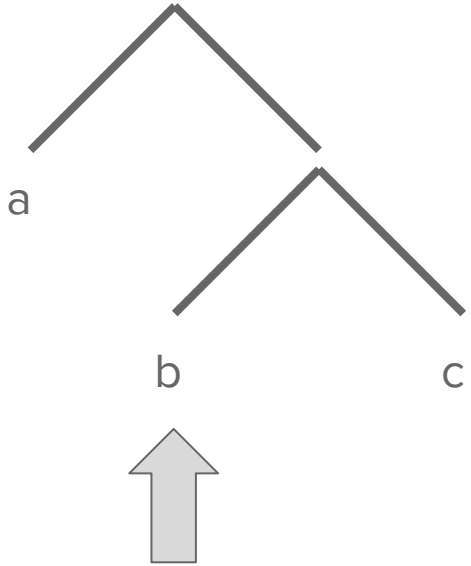
b = {b1, b2, b3, b4, b5, ...bj}

c = {c1, c2, c3, c4, ...ck}

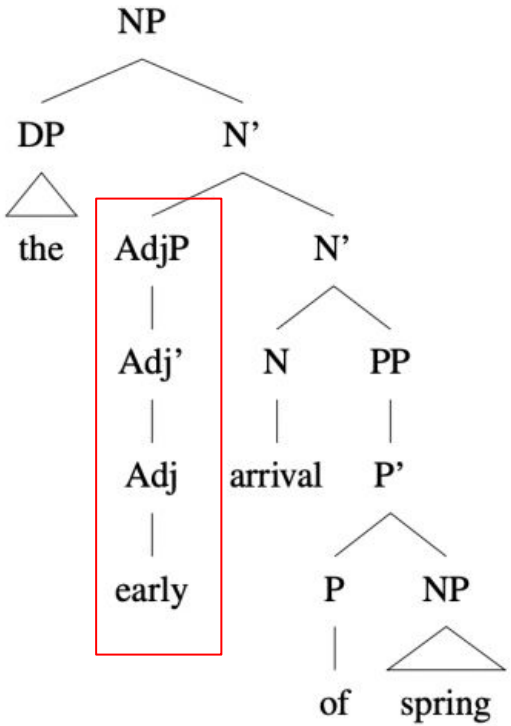


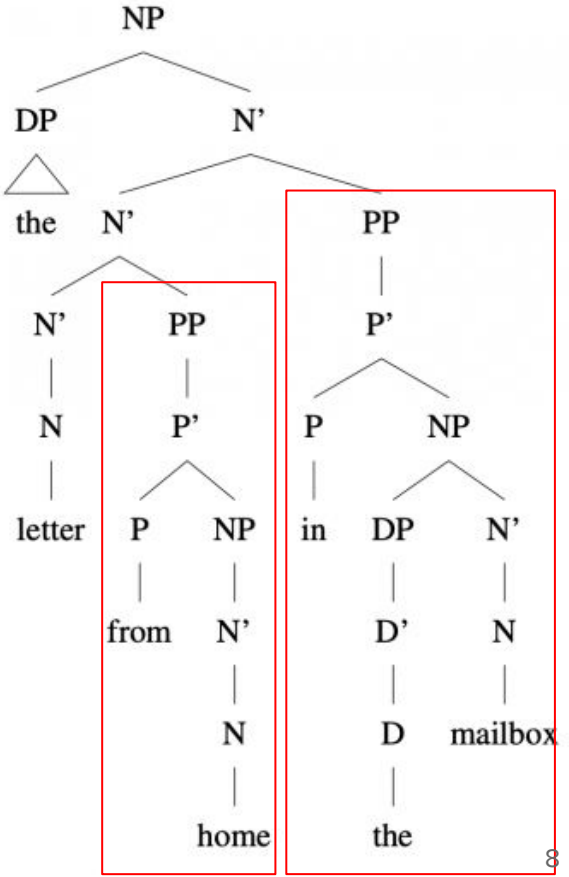
這個數量很少，那在一個語言裡，
它一定是常常出現的高頻詞！
(因為從 a1 ～ ai 沒有幾個字可以用)

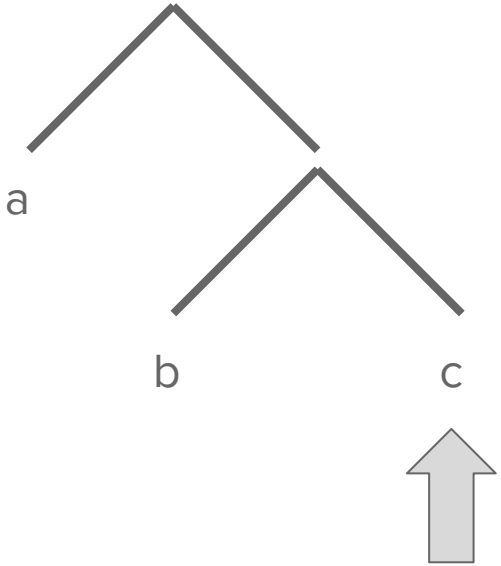




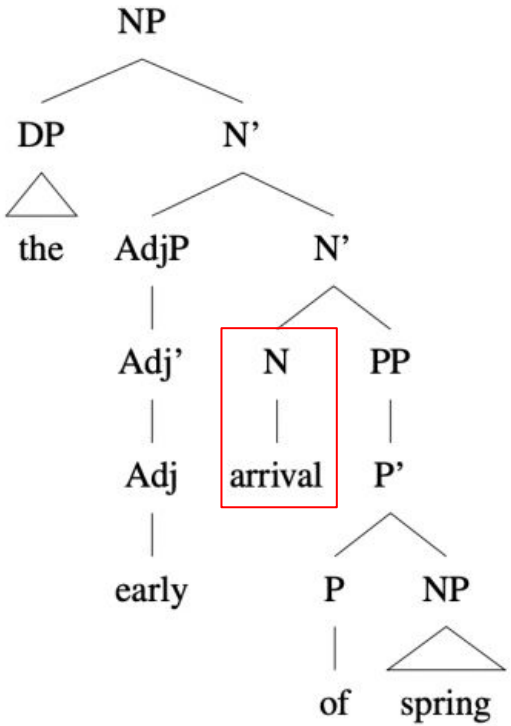
這個數量很多，且在一個語言裡，它和許多東西都能搭配，那它一定是 adjunct.

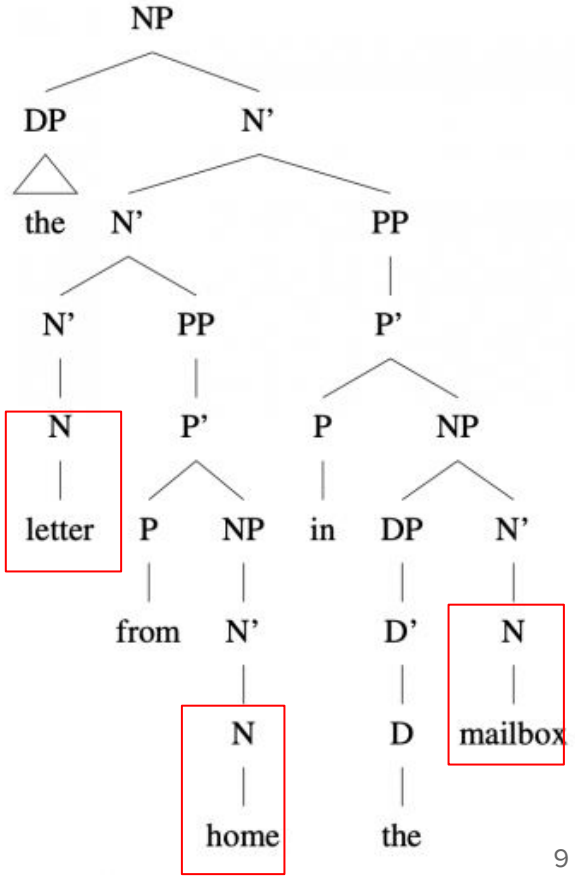






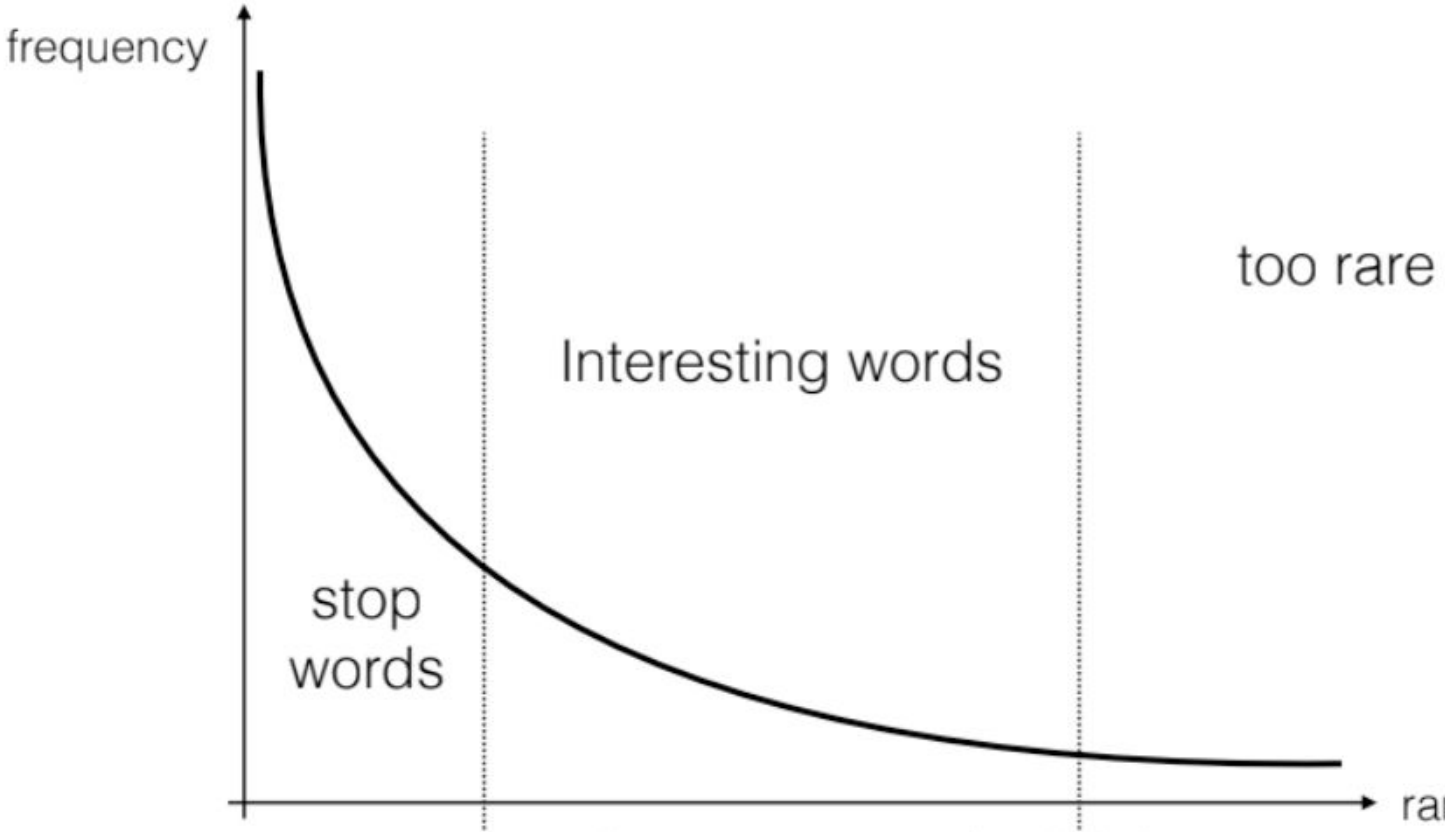
剩下的東西在這裡！
(剛好都是中心語呢！
真巧吶！)





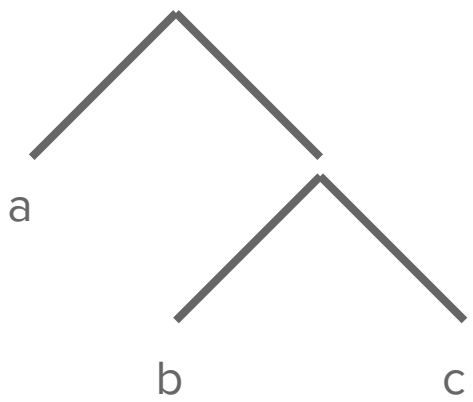
Zipf's Law

since 1935



https://www.ngcm.soton.ac.uk/2021/07/07/natural-language-processing-in-python/

用數學描述 X-bar



Definition of X-bar elements

For elements in an X-bar structure:

a. **Specifier**: Zipf's law's high-freq words
b. **Adjunct**: freqComf(b, x) > freqComf(b, c)
c. **Head**: setX - set(a) - set(b)

Assignments

分類任務:

請設計一個分類任務，並將你要分辨的兩類文本X, Y 分別放到 Whoiswho.py 中的 TPP_LeeLIST, NTU_LeeLIST 中 (可視需求改名)，做為產生模型使用。

收集 X, Y 兩類文本，分別放入 test_tpp 和 test_ntu 中 (可視需求改名)。執行看看「詞彙模型」的預測和正解是否一致。

回答以下兩個問題:

1. 預測一致的時候，你抓取哪個詞類做為特徵？
2. 請解釋為什麼這個詞類可以做為你的模型特徵。

OR

1. 若你的模型預測不一致，請說明你抓取哪個詞類做為特徵？
2. 請解釋你覺得為什麼此時詞彙模型無法做好你設計的分類工作？