D08 基礎語言模型：N-Gram

簡報閱讀 範例與作業 問題討論

重要知識點
什麼是語言模型？
語言模型機率
N-Gram模型

重要知識點

- 了解何謂語言模型
- 了解 N-Gram 語言模型的原理以及應用

什麼是語言模型？

house
Malt
This is the ...
did
mouse

當我們看到 this is the 的詞的時候知道後面接 house 比接 did 來得合理。

可口可樂 ⟶ 百事可樂
皮卡丘 ⟶ 寶可夢
天使獸 ⟶ 數碼寶貝

同理當我們想到可口可樂的時候，可能會"聯想"到百事可樂(而不是珍珠奶茶)，人們長期在學習累積語言後有判斷合不合理與聯想的能力。
那 NLP 有沒有這樣判斷合不合理與聯想的能力呢？
N-gram 就是基於這樣聯想的語言模型。

語言模型如何做到聯想或判斷句子的合理與不合理？
人類根據其他字詞(this is the)來判斷當前字(house)出現合理與否，而語言模型也是一樣的！ 根據其他字詞來判斷當前自出現的機率，因此便可以藉由機率大小的方式來判斷句子出現可能性。

p(house|This is the) > p(did|This is the)

(house：當前字；This is the：前序字詞)

語言模型機率

corpus(文本資料)：
This is the house that Jack built.
This is the malt.
That lay in the house that Jack built.
This is the rat.
That ate the malt.
That lay in the house that Jack built.
This is the cat.
That killed the rat.
That ate the malt.

根據貝氏定理：

$$p(\text{house}|\text{This is the}) = \frac{\text{count(this is the house)}}{\text{count(this is the...)}} = \frac{1}{4}$$

由上述的說明可以知道，若想得到一個句子出現的機率，如：P(This is a house)，可轉化為下列表示。

取的下列句子機率：
W = (W1W2W3W4....Wm)
P(W1,W2,W3,W4,...,Wm) = ?

根據鏈鎖率：
P(W1,W2,W3,W4,...,Wm) = P(W1)*P(W2|W1)*P(W3|W1,W2)*...*P(Wm|W1,...,Wm-1)

因此若想知道 "我喜歡深度學習更喜歡NLP" 合不合理(機率大不大)可以藉由下面方法得到

P(我喜歡深度學習更喜歡NLP) = P(我)*P(喜|我)*P(歡|我,喜)*...*P(P|我喜歡深度學習更喜歡NL)

為了簡化問題，我們可以假設當前的字出現只跟前一個字有關

P(我喜歡深度學習更喜歡NLP) = P(我)*P(喜|我)*P(歡|喜)*...*P(P|L)

這樣的假設算法即是 N-Gram 模型中 n=2 的情形，又稱作 Bigram

PS：試著思考為何需要做這樣的簡化呢？(詳細請見作業)

N-Gram模型

N-Gram 語言模型是基於統計的語言模型算法，主要是將文本中的內容取最靠近的 N 個字當作條件機率計算的先驗條件，形成長度是 N 的字詞片段序列。每個字詞片段及稱為 gram。 常見的 N-Gram 模型有 Unigram(1-gram)，Bigram(2-gram)，Trigram(3-gram)。

(句子出現的合理性(機率) ＝ 左式；字詞出現的機率 ＝ 右式中的項)

Unigram: $$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i)$$

Bigram: $$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-1})$$

Trigram: $$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-2}w_{i-1})$$

Bigram語言模型

我們知道了 Bigram 語言模型，現在我們來看看 Bigram 模型如何判斷句子的合理性，為了考慮字詞在開頭與結尾的情境，可以加入 start 的 token 與 end 的 token 的在文本中。

Bigram (n=2)：
P(W1,W2,W3,W4,...,Wm) = P(W1|<start>)*P(W2|W1)*P(W3|W2)*...*P(Wm|Wm-1)*P(<end>|Wm)

corpus：
That is the house that Jack built.
This is the malt.
It is a fancy house.

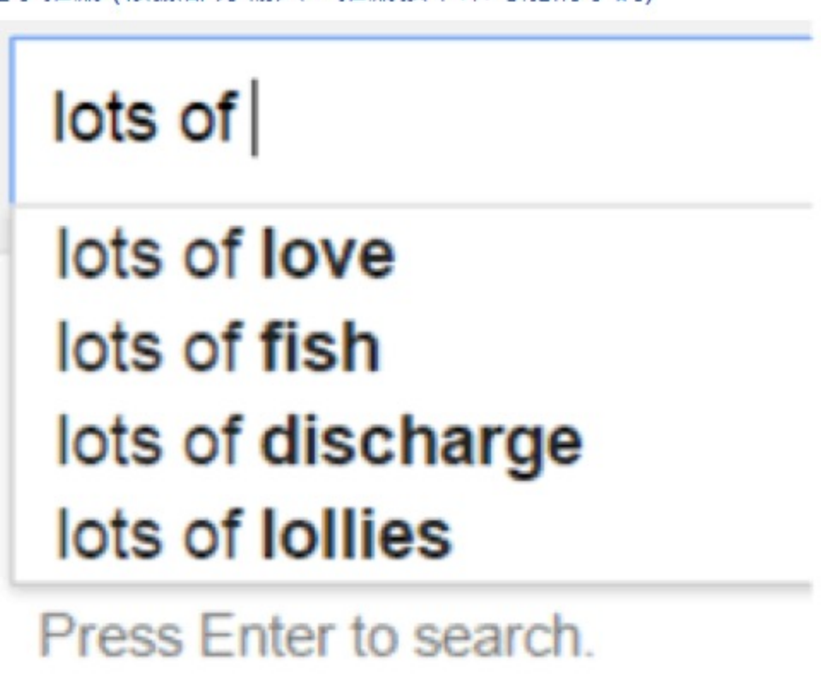
Probabilities：

P(this is the house) = P(this|<start>) * P(is|this) * P(the|is) * P(house|the)*P(<end>|house)
= (1/3) * (1) * (2/3) * (1/2)*(1/2)

N-Gram應用場景

語言模型可以應用的場景很多，N-Gram 常用的應用場景像是"選字推薦"、"錯字勘正"、"分詞系統"等

選字推薦 (根據部分輸入，推薦接下來可能的字詞)



分詞系統(目前分詞算法還是以斷詞課程中介紹的為大宗)

$$P(Y_1) = p(\text{我})p(\text{喜欢}|\text{我})p(\text{自然语言处理}|\text{喜欢})$$

$$P(Y_2) = p(\text{我})p(\text{喜}|\text{我})p(\text{欢}|\text{喜})p(\text{自然}|\text{欢})p(\text{语言}|\text{自然})p(\text{处理}|\text{语言})$$

$$P(Y_3) = p(\text{我})p(\text{喜}|\text{我})p(\text{欢}|\text{喜})p(\text{喜}|\text{自})p(\text{然语言}|\text{自})p(\text{处理}|\text{然语言})$$

錯字勘正

今天天氣狠好，適合出門玩

P(天|今)*P(天|天)*P(氣|天)*P(狠|好) < threshold

N 的大小

當N值較大：對字詞的約束性更高，具有更高的辨識力，複雜度較高
當N值較小：字詞在文本出現的次數較多，更可靠的統計結果，但對字詞的約束性較小

你好嗎
→ Unigram: 你, 好, 嗎
→ Bigram: 你你 ,你好, 你嗎, 好好, 好嗎, 好你, 嗎嗎, 嗎好, 嗎你

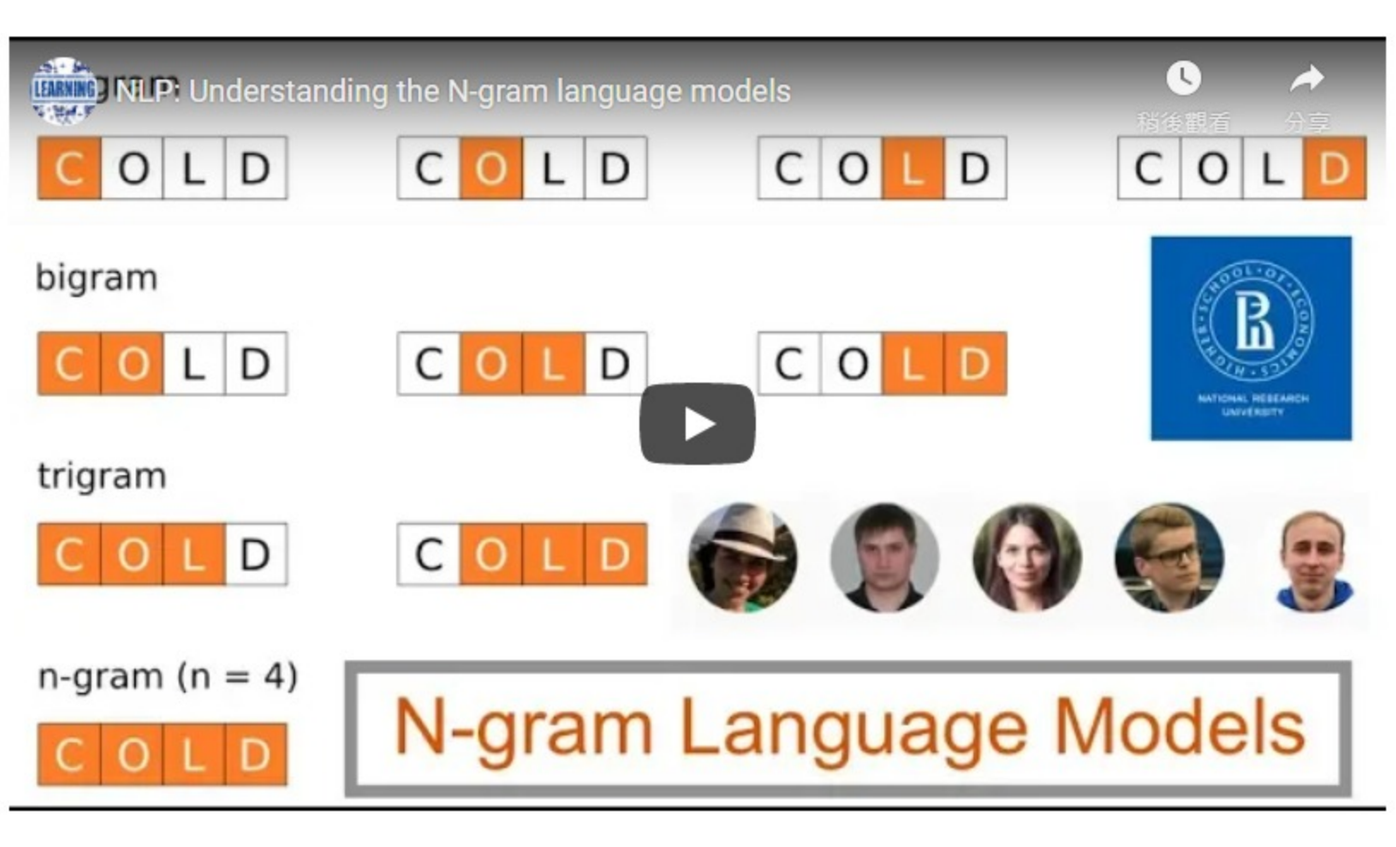
※ Bigram 的可能的 gram 數較多

知識點回顧

在這章節我們學習到了
- 何謂語言模型
- 了解 N-Gram 語言模型的原理以及應用

延伸閱讀

網站：基礎語言模型介紹
基礎語言介紹 YouTube 影片



網站：語言模型得應用
介紹常見的語言模型應用

網站：N-Gram語言模型補充介紹
文中提到語言模型機率計算的挑戰

INTRODUCTION

Language Models: N-Gram

A step into statistical language modeling

Shashank Kapadia Follow
Mar 26, 2019 · 4 min read

Introduction

Statistical language models, in its essence, are the type of models that assign probabilities to the sequences of words. In this article, we'll understand the simplest model that assigns probabilities to sentences and sequences of words, the ***n-gram***

You can think of an N-gram as the sequence of N words, by that notion, a 2-gram (or bigram) is a two-word sequence of words like "please turn", "turn your", or "your homework", and a 3-gram (or trigram) is a three-word sequence of words like "please turn your", or "turn your homework"

Intuitive Formulation

Let's start with equation P(w|h), the probability of word *w*, given some history, *h*. For example,

$$P(the \mid its\ water\ is\ so\ transperant\ that)$$

網站：N-Gram中文錯字檢查
使用 N-gram 實踐中文錯字檢查

中文拼寫糾正

最基本的思想，將所有的常見錯別字整理為字典。
但是這個字典的數量實際上非常有限，所以還是要藉助算法。
本文簡單地講解如何使用n-gram模型結合漢字拼音來作中文錯別字糾錯，然後介紹最短編輯距離在中文搜索糾錯方面的應用；
最後從依賴樹入手講解如何作文本長距離糾錯（語法糾錯），並從該方法中得到一種啟示，利用依賴樹的特點結合ESA算法來做同義詞的查找。

n-gram模型

在中文錯別字查錯情景中，我們判斷一個句子是否合法可以通過計算它的概率來得到，假設一個句子S = {w1, w2, ..., wn}，則問題可以轉換成如下形式：

```
[plaintext]
P(s) = P(w1, w2, ..., wn) = P(w1) * P(w2|w1) * P(w3|w2,w1) *....*P(wn|wn-1,wn-2,...,w2,w1)
```

P(S) 被稱為語言模型，即用來計算一個句子合法概率的模型。
但是使用上式會出現很多問題，參數空間過大，信息矩陣嚴重稀疏，這時就有了n-gram模型，它基於馬爾科夫模型假設，一個詞的出現概率僅依賴於該詞的前1個詞或前幾個詞，則有

下一步：閱讀範例與完成作業