筆記

機器翻譯:

不同語言間的文本自動翻譯

原始語言和目標語言使用詞彙和結構差異

原始語句文具分析，目標語言文具生成

日本準備東奧:

如何招待全世界的選手、訪客。準備照護各區域的方言翻譯

機器問題:

問題>>文件檢索>>答案選擇

在文件檢索中，有機會遇到代名詞，而要讓電腦認知到代名詞是在指哪一位人名。

文本分析:

對各類型來源的數位資料，進行文件探勘

1市場產品資訊

2政治意見追蹤

3社群網路分析

4熱門議題分析

2009年市場分析:

以台積電來做分析，以新聞和評論作分析

多標的輿情分析:

機器人和人辯論賽:

歧義解析:

1詞彙層次

2語法層次

3語意層次

如何讓電腦知道在當下的情況還使用一個單字的形容詞狀態還是名詞狀態?

主體在外面還是裡面，被動還是主動。

容錯力:

打錯字

語法錯字

不合文法

如何讓電腦能夠接受這些非正式的單字

強健性:

領域改變

網路語言:俗語、外來語、縮寫、諧音、符號合併

表情符號:能夠讓一個句子的情感表現更豐富

概念表示:

貓、CAT、ネコ

以一個符號來表示某個概念，容易產生歧義

蝙蝠>>蝙蝠

>>bat>>球棒

>>球拍

必須了解符號在某情況下的意涵

語言單位:

1詞素 們

2字元 台

3詞彙 台灣大學

4n-連詞 台灣

5多詞表達 網路紅人>網紅

6命名實體 台灣積體電路股份有限公司 台積電

7子句(clause)

8句子(sentence)

9段落(paragraph)

10段落(passage)

11文件(document)

12多文件(multi-document)

一個文件:

一個文件有三個段落，每個段落有多個句子。

要抓出句子的邊界(標點符號)

分類:掌握共通性

詞性類別:

靠詞性斷詞和標記

NA普通名詞 NE地方詞 VK狀態句賓動詞 VH狀態不及物動詞 D副詞 DFA動詞前程度副詞

語意類別:

同意詞>>夕陽、斜陽、殘陽、落日

相依類別:

我喜歡小吃店

4個分類

言談類別:

1時序類

2因果類

3轉折類

4推展類

意見類別:

1正面>>我喜歡、很好、舒服

2負面>>古老、不大

情感類別:

1憤怒

2厭惡

3恐懼

4喜悅

5悲傷

6驚訝

立場類別:

1贊成

2反對

自然語言處理一些任務:

1詞素分析

2詞彙 中文斷詞

3命名實體辨識

4詞性標記

5句法類別 句法剖析

6相依類別 相依剖析

7語義類別 語義關係標記

8言談關係 言談剖析

9意見探勘

10情感分析

11立場偵測

符號計算:

任何詞彙、類別都以符號呈現。放進電腦比對

夕陽!=落日>>匹配失敗，需要外部資源

分布式表示:

意思產生來自使用

語境(上下文):

關門、把門關上、打烊、停業

上下文範圍界定:

篇章>>段落>>句子>>子句

共同出現:

語境(上下文)>>以句子為單位

宣布關門

關門歇業

關門、關門...(ry)

以向量(c1,c2…..cn)， n表示語境的個數，ci是wi出現在對應語境的次數

向量計算:

利用高維度向量表示

例如cosine用來計算兩個詞彙的關係

夾角越小，詞彙關係越緊密

1優點>>可計算語義關聯程度

2缺點>>維度太高，太稀疏

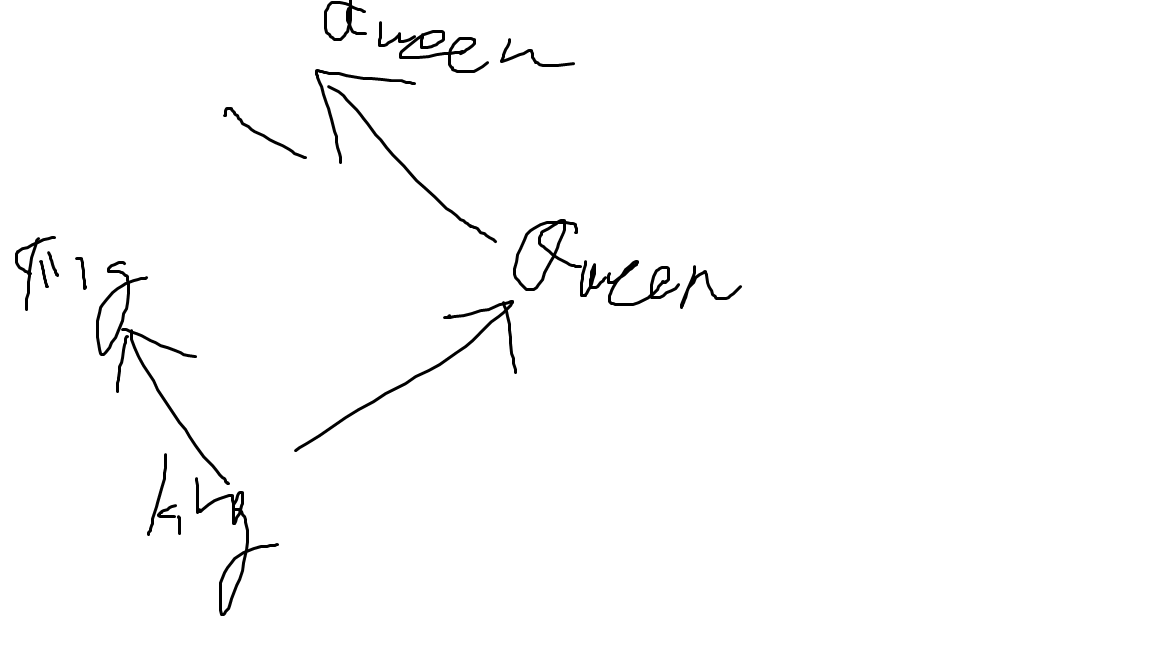
3解決>>進行降維處理

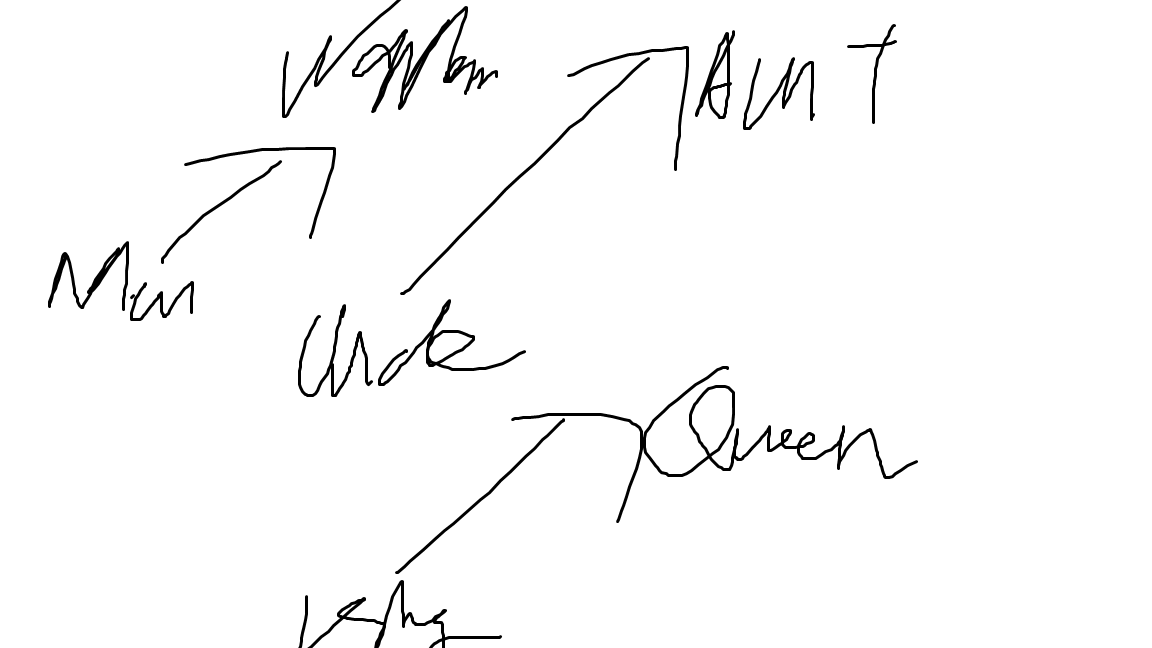
分散式表示:

1用上下文的詞來預測目標詞

2以當前的詞來預測上下文的詞

將詞會轉換成低維度稠密向量:

1語法關係 

2語義關係

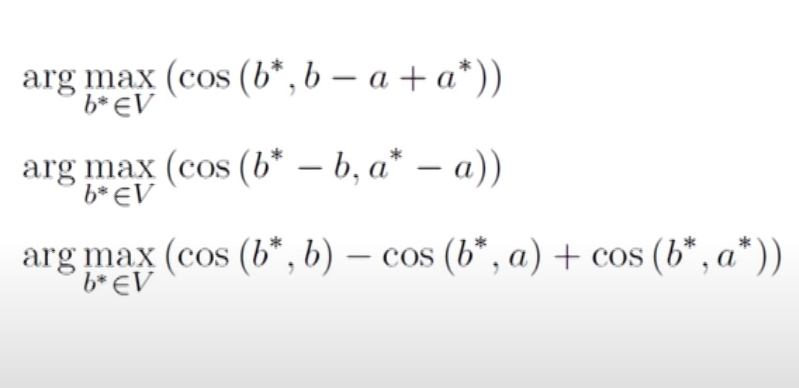
類比運算:

1語法關係>>QUEENS，單複數關係

2語義關係>>QUEEN，性別關係；華盛頓=印度，首都OR甘地(國父)

向量運算:

a:a\*=b:b\*

多種計算方式>>

延伸:

1語言成分表示>>詞向量、句子向量、段落向量、文件向量

2成分類別表示>>詞性向量、語義向量、句法類別向量

3用戶資料表示>>用戶向量

相似關係計算:

1詞彙分群

2句子分群

3段落分群

4文件分群

5問答系統答案選擇

6文件搜尋

言談剖析器結構:

