歡迎學習“模組 6：自然語言處理簡介”。

本模組包含以下小節：

- 自然語言處理 (NLP) 簡介。本簡介部分將介紹使用 NLP 時面臨的主要挑戰，以及 NLP 應用程式的整體開發過程。
- 概述了五項可用于加速開發基於 NLP 的應用程式的 AWS 服務。

本模組還包括 <錄製的/講師主導的> 演示，將向您展示如何通過 AWS 管理主控台使用 Amazon Comprehend、 Amazon Polly 和 Amazon Translate。

此外，本模組還包括一個引導式動手實驗，教您如何使用 Amazon Lex 開發一個用於安排預約的機器人。

最後，您需要完成一個知識測驗，以測試您對本模組中涵蓋的關鍵概念的理解程度。

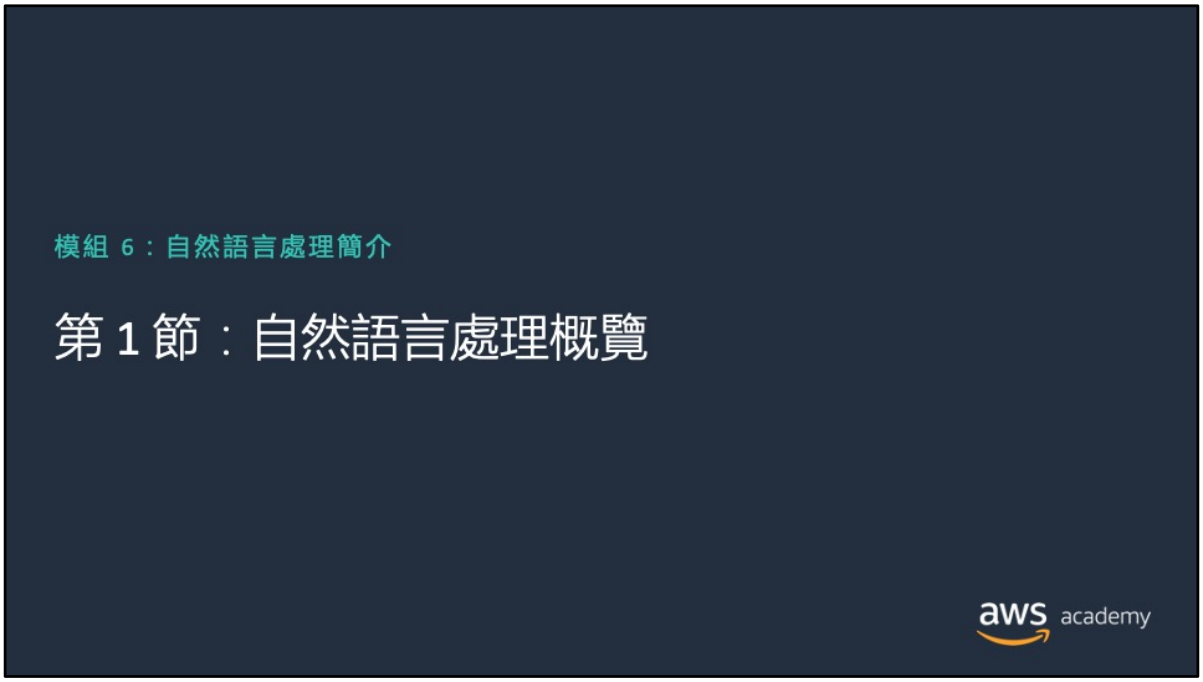
模組目標

學完本模組後，您應該能夠：

- 描述使用託管 Amazon ML 服務解決的自然語言處理 (NLP) 使用案例
- 描述適用於 NLP 的託管 Amazon ML 服務
- 使用託管 Amazon ML 服務

完成本模組後，您應能夠：

- 描述使用託管 Amazon ML 服務解決的自然語言處理 (NLP) 使用案例
- 描述適用於 NLP 的託管 Amazon ML 服務
- 使用託管 Amazon ML 服務



介紹“第 1 節：自然語言處理概覽”。在本節中，您將瞭解自然語言處理的含義。

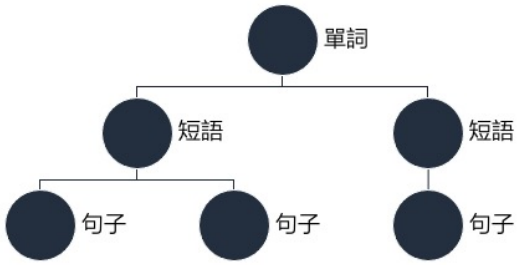
在瞭解自然語言處理 (NLP) 的含義之前，我們先來看一個使用 Amazon Alexa 的 NLP 示例。

1. 使用 Amazon 設備（如 Echo）錄音。錄音會發送到 Amazon 伺服器，以進行更加高效的分析。
2. Amazon 將您的話語分解為單獨的語音。然後，它連接到包含各個單詞發音的資料庫，找到最接近各聲音組合的單詞。
3. 它通過識別重要的單詞來理解任務內容並執行相應的功能。例如，如果 Alexa 聽到*外部*或*溫度*之類的單詞，它將打開天氣應用程式。
4. Amazon 伺服器將資訊發送回您的設備後，Alexa 可能會開始講話。

NLP 是什麼？

NLP 開發了可以自動分析和表示人類語言的計算演算法。

機器學習系統可以通過評估語言結構來處理大量單詞、短語和句子。

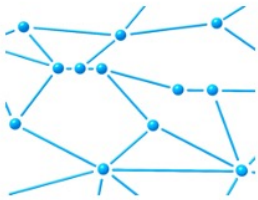


NLP 是一個廣義術語，用於描述可以通過機器學習 (ML) 解決的一系列常規業務或計算問題。NLP 系統要早於 ML。您的舊手機和螢幕閱讀器上的語音轉文本即是兩個有力證據。現在，許多 NLP 系統都會使用某種形式的機器學習。NLP 會考慮語言的分層結構。單詞是分層次結構的最低層。一組單詞構成一個短語。下一級是由短語組成的句子；最後由句子來表達想法。

NLP 系統面臨幾個重大挑戰，接下來將向您講解。

NLP 面臨的挑戰

不夠精確



許多複雜的依賴項

基於上下文的含義

缺乏結構

語言並不精確。一個單詞可能具有不同的含義，這取決於和它一起組合的其他單詞（即上下文）。通常，同一個單詞或短語可能具有多種含義。例如，*weather* 一詞。如果是 *under the weather*，則表示身體不舒服。但是，如果說 *there is wonderful weather outside*，則表示外面的天氣很好。對於短語 "*Oh, really?*" ，由於上下文和語調的不同，可能會表達驚訝、不同意或許多其他含義。

NLP 面臨以下一些主要挑戰：

- 識別文本結構 – 任何 NLP 應用程式的首要任務之一都是將文本分解為有意義的單位（例如單詞、短語和句子）。
- 為資料添加標籤 – 系統將文本轉換為資料後，下一個挑戰就是應用代表各種詞性的標籤。每種語言都需要使用不同的標記方案來匹配該語言的語法。
- 呈現上下文 – 由於單詞意義取決於上下文，因此任何 NLP 系統都需要一種呈現上下文的方法。由於上下文眾多，因此這是一個巨大的挑戰。將上下文轉換為電腦可以理解的形式非常困難。
- 應用語法 – 儘管語法定義了語言的結構，但語法的應用場合幾乎是無窮無盡的。人類使用語言的方式千差萬別，如何處理這種差異是 NLP 系統面臨的主要挑戰。如果可以解決這一挑戰，機器學習將能夠產生巨大影響。

自然語言處理使用案例

搜索應用程式

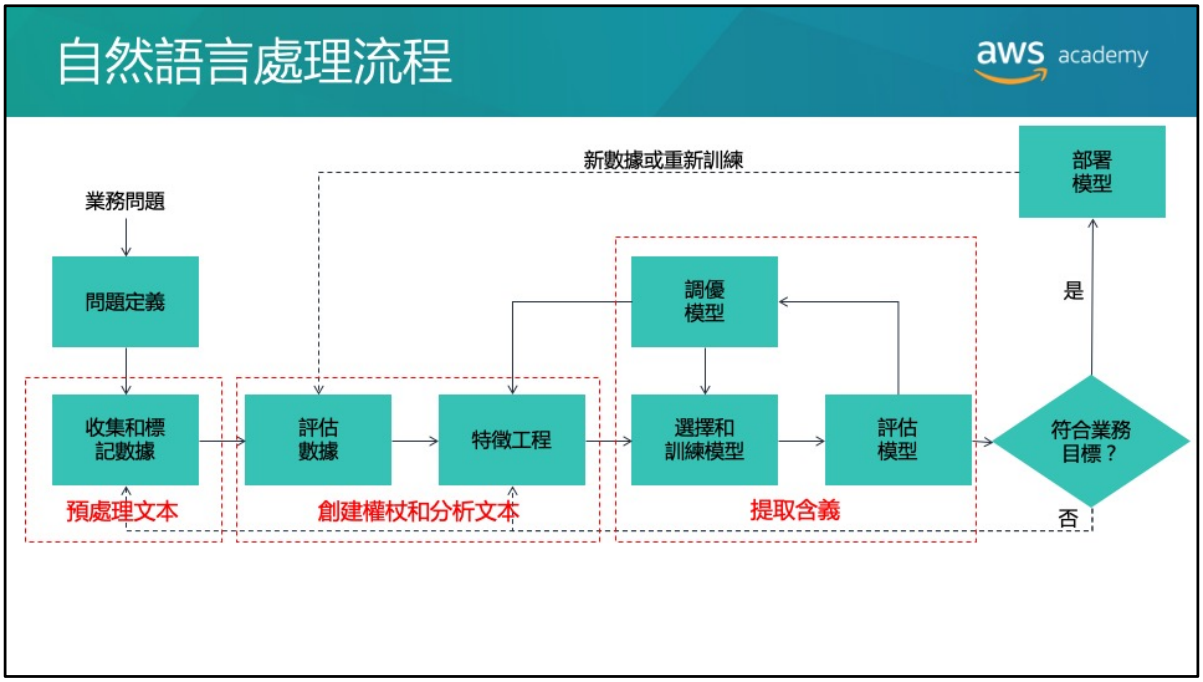
市場和社會研究

人機界面

Chatbot

您可以將 NLP 應用於各種問題。一些比較常見的應用包括：

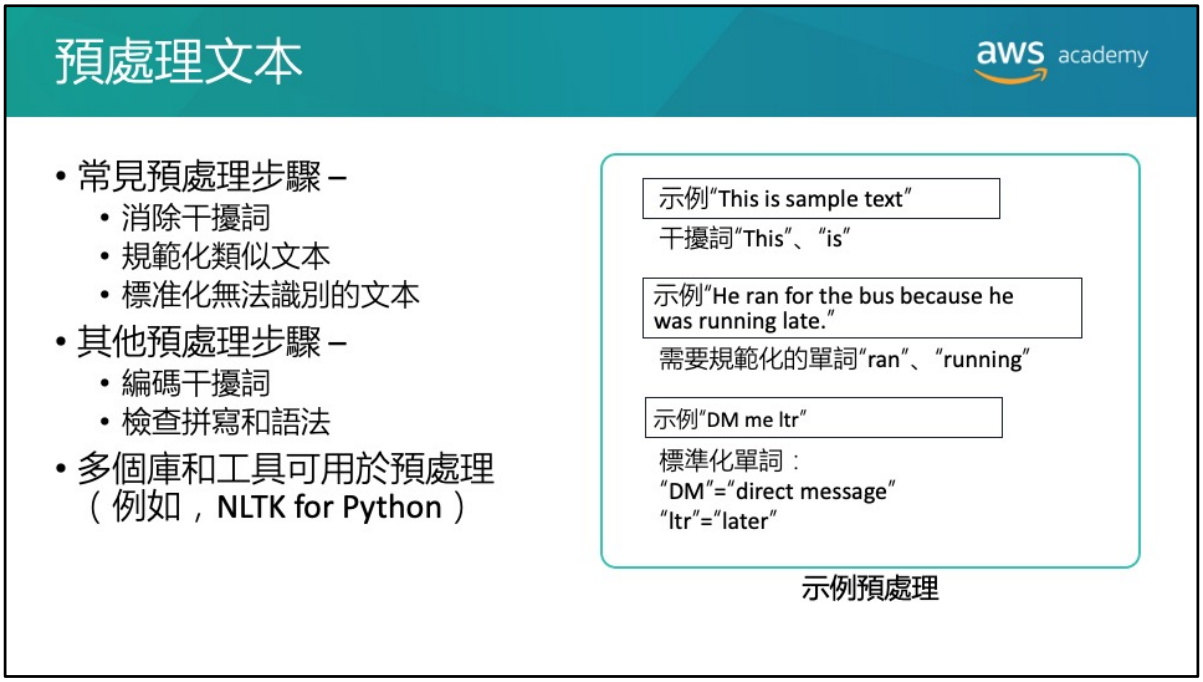
- 搜索應用程式（如 Google 和 Bing）
- 人機介面（如 Alexa）
- 市場行銷或政治運動的情緒分析
- 基於媒體分析的社會研究
- 在應用程式中模仿人類語音的 Chatbot



在開發 NLP 解決方案時，您可以應用在本課程中學習的 ML 開發管道。首要任務是定義一個問題，然後收集和標記資料。

對 NLP 來說，收集資料包括將文本分成有意義的子集以及標記集合。特徵工程是 NLP 應用程式的一個重要組成部分。使用不規則或非結構化文本時，這一過程將變得更加複雜。例如，如果您構建一個應用程式來對文檔進行分類，必須能夠區分同字不同義的單詞。

在 NLP 領域中為資料添加標籤有時也稱為*標記*。在添加標籤的過程中，必須為將各文本字串分配到不同的詞性。您可以使用專用工具來說明進行 NLP 標記。

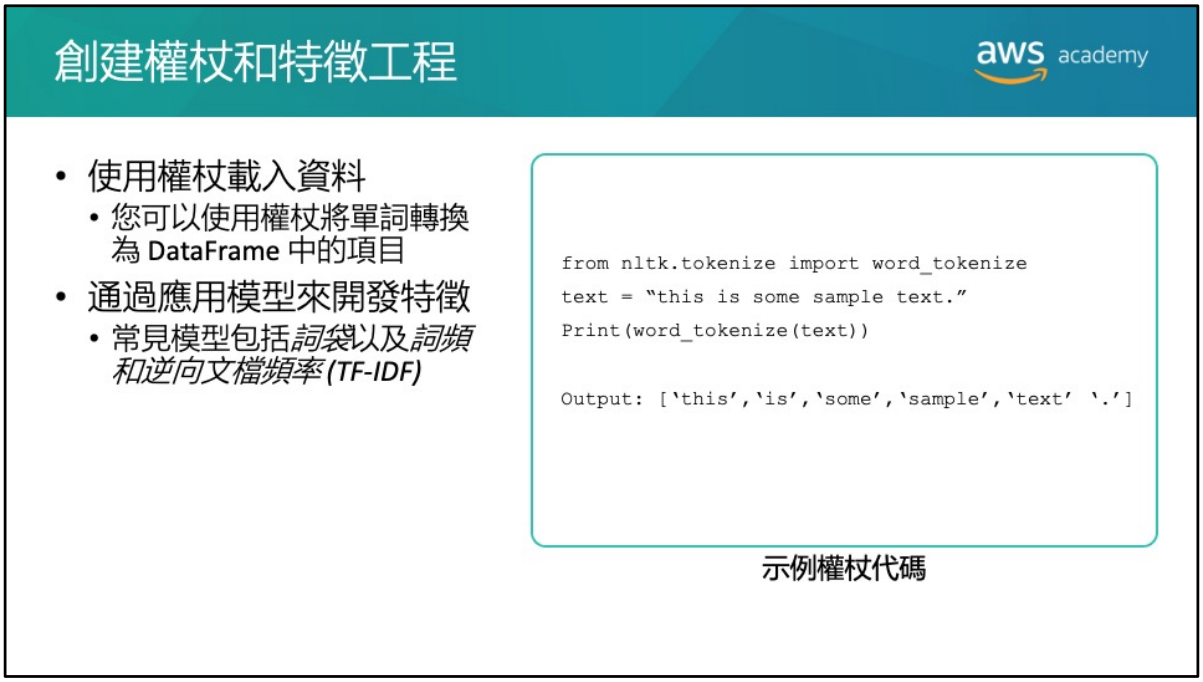


NLP 應用程式的首要任務是將文本轉換為資料，以便對其進行分析。您可以從輸入文本中刪除分析不需要的單詞來完成文本轉換。在該示例中，單詞 *This* 和 *is* 被刪除，保留短語 *sample text*。

刪除干擾詞後，可以將相似的單詞轉換為通用形式來規範文本。例如，單詞 *run*、*runner*、*ran* 和 *running* 都是單詞 *run* 的不同形式。您可以通過詞根提取和詞形還原流程，在一個文字區塊中規範化這些單詞的所有實例。

規範化文本後，可以刪除在分析中使用但未收錄在詞典中的單詞，來對其進行標準化。例如首字母縮略詞、俚語和特殊字元。

自然語言工具包 (NLTK) Python 庫提供了一些函數，可用於消除干擾詞以及規範化和標準化文本。



創建 NLP 系統的一個首要步驟是將文本轉換為資料集合，例如 DataFrame。所有 NLP 庫都會提供協助進行此類轉換的函數。此示例介紹了如何使用 NLTK 庫中提供的 word_tokenize 函數。

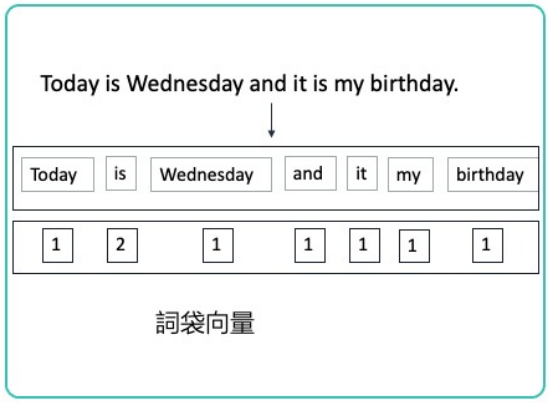
清理文本並將其載入到 DataFrame 後，可以應用一個 NLP 模型來創建特征。

常見模型包括：

- 詞袋 – 這種簡單的模型可以捕獲文檔中單詞的出現頻率。對於文檔中的每個單詞，模型都會創建一個鍵，其值即該單詞在該文檔中出現的次數。
- 詞頻和逆向文檔頻率 (TF-IDF) – *詞頻*是一個單詞在文檔中出現次數的計數。*逆向文檔頻率*是一個單詞在一組文檔中的出現次數。結合使用這兩個值來計算單詞的權重。在許多文檔中頻繁出現的單詞權重較低。

示例 NLP 模型：詞袋

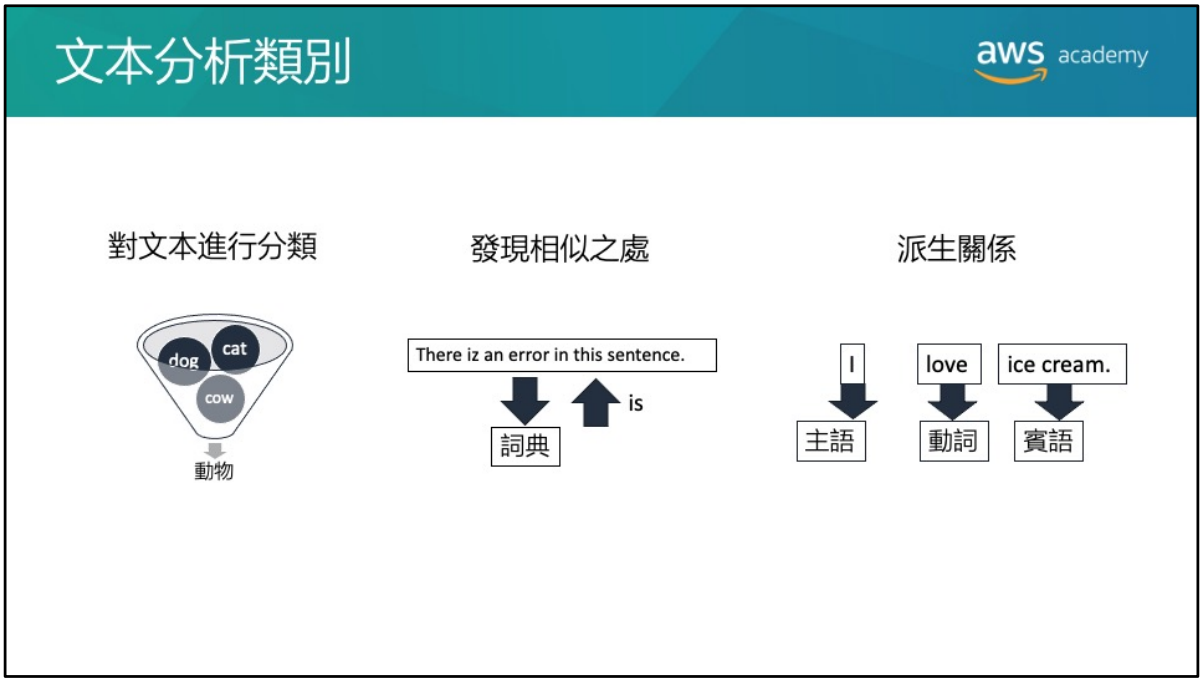
- 為每個句子或短語創建一個向量
- 根據頻率評估句子中的單詞
  - 頻率會為每個句子或短語創建一個向量



示例 NLP 模型

我們在 NLP 領域已經建立了許多模型。此示例展示了詞袋模型。詞袋是一個向量模型。向量模型將每個句子或短語轉換為向量，向量是記錄方向和大小的數學物件。在該示例中，一個簡單的句子被轉換為向量，其中每個詞都按頻率記錄。單詞 *is* 的值是 *2*，因為它在句子中出現了兩次。

詞袋通常用於將文檔分為不同的類別。詞袋還可用於派生饋入到 NLP 應用程式的屬性，例如情緒分析。



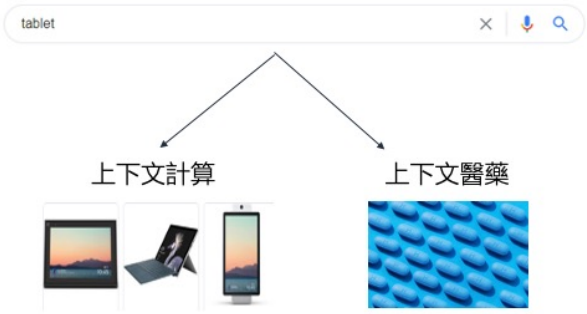
文本分析分為三大類：

- 對文本進行分類 – 此類分析類似於您在本課程中學到的其他分類系統。文本為提取特徵的流程提供輸入，然後通過 ML 演算法發送這些特徵。此演算法與分類器模型交互，以推斷分類。您可以使用 NLTK Python 庫來創建分類系統。
- 發現相似之處 – 文本匹配應用廣泛。例如，自動校正、拼寫檢查和語法檢查都基於文本匹配。編輯距離（也稱為 Levenshtein 距離）演算法非常常用。
- 派生關係 – 您可以使用名為共指消解的過程來得出文本中不同單詞或短語之間的關係。多個 NLP 系統都提供了用於派生關係的 Python 庫。

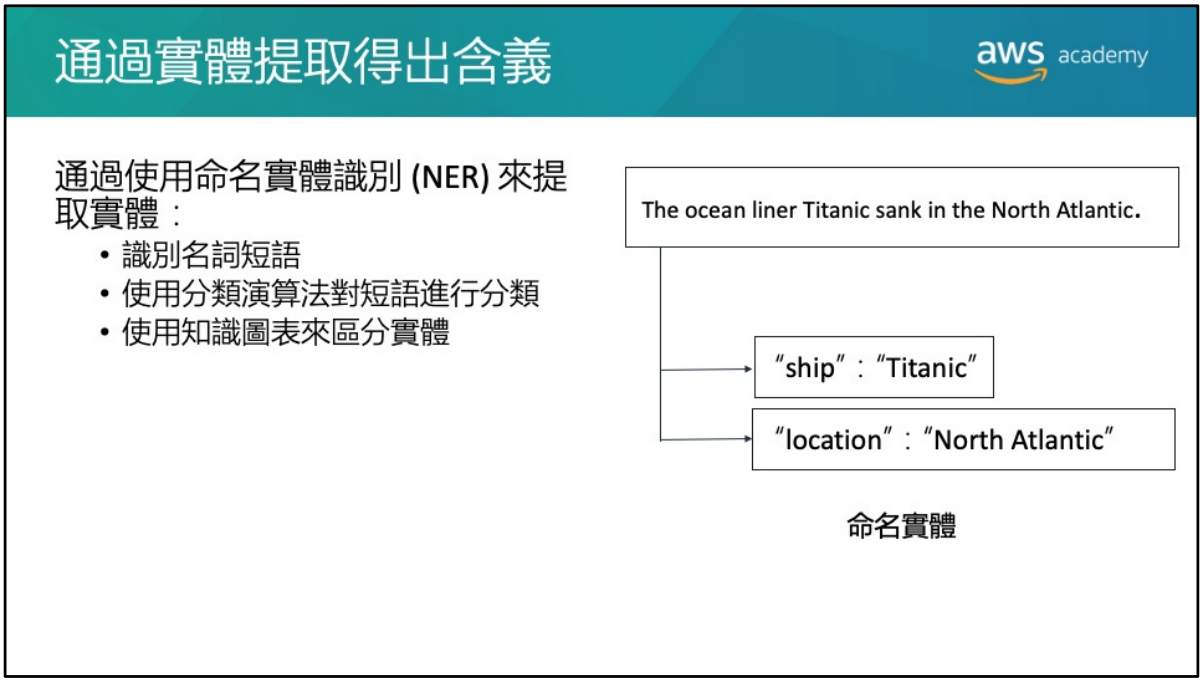
捕獲上下文

理解文本的上下文是 NLP 面臨的主要挑戰：

- 用適當的詞性來標記單詞有助於捕獲上下文
- NLP 庫可提供權杖函數，幫助進行標記



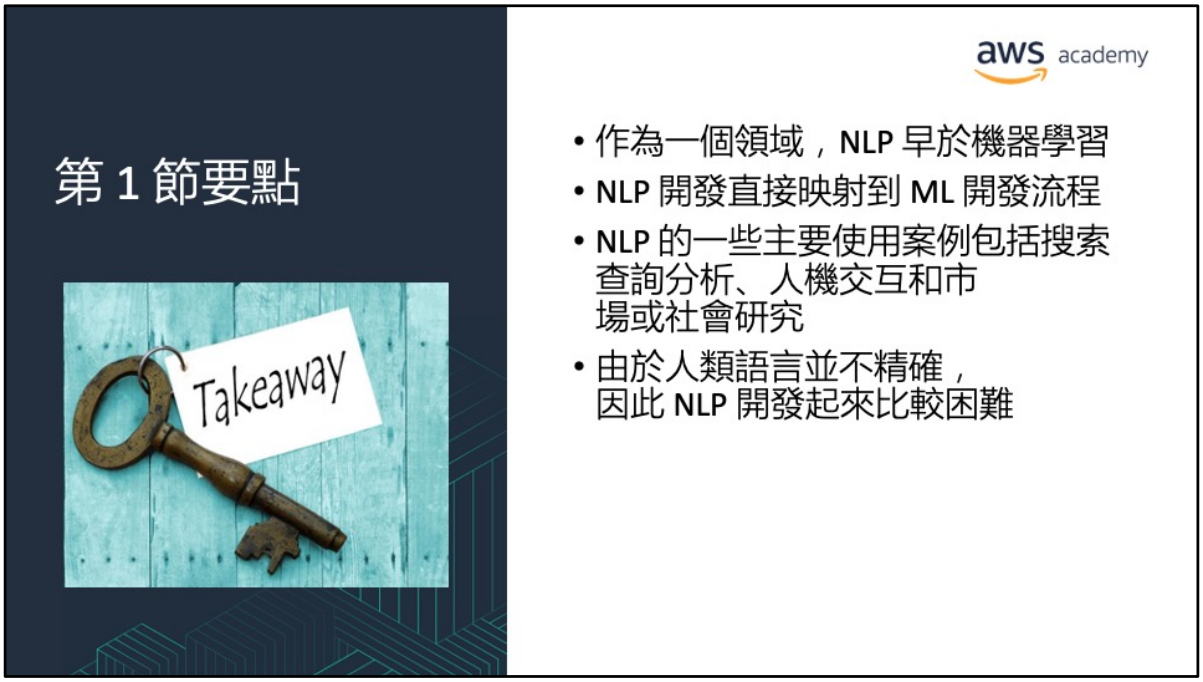
NLP 面臨的一個最大挑戰是如何描述文本的上下文。來看這樣一個例子，用戶正在搜索單詞“*tablet*”。由於單詞 *tablet* 至少有兩個不同的含義，搜尋引擎必須清楚用戶想要表達的含義。該詞可能會被進一步限定（例如，通過添加另一個詞，如 *medicine* 或 *computing*）。如果沒有其他條件，則大多數搜尋引擎都依賴最常用的上下文。



我們將提取實體的過程稱為命名實體識別 (NER)。NER 模型由以下部分組成：

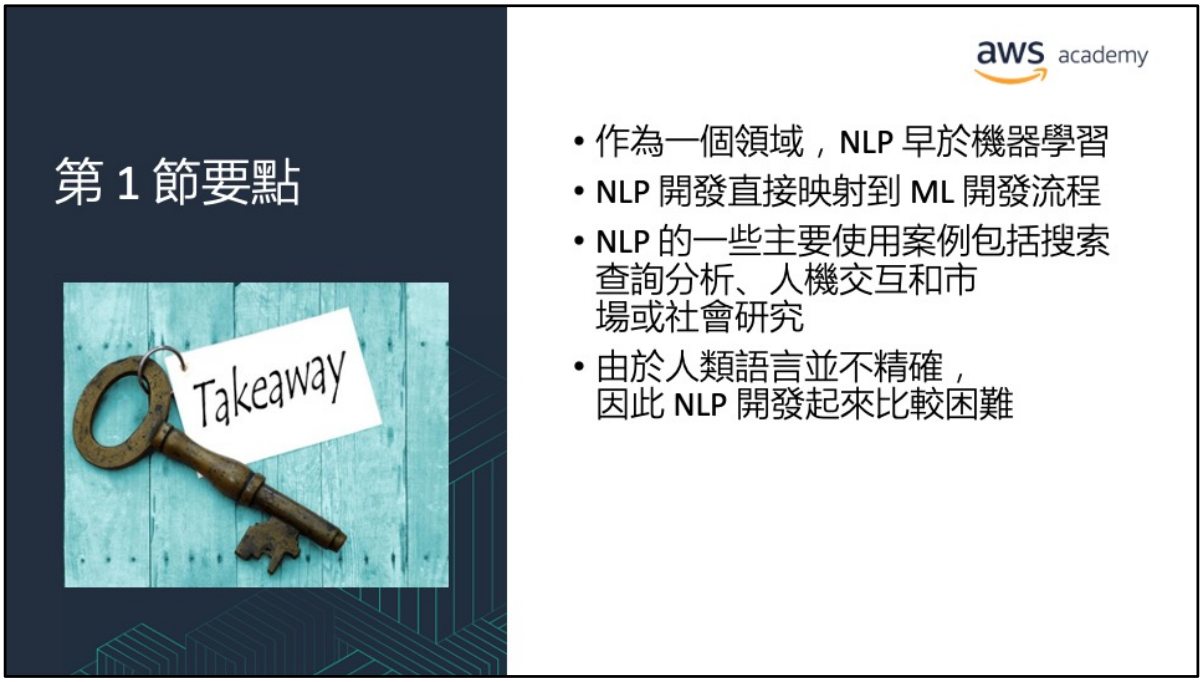
- 使用依賴項圖表和詞性標記來識別名詞短語
- 使用分類演算法（例如 Word2Vec）對短語進行分類
- 使用知識圖表來區分實體

本示例顯示了如何使用 NER 從文本中提取實體 *Titanic* 和 *North Atlantic*。提取命名實體後，可以使用知識圖表來提取含義。知識圖表將主題專業知識與機器學習相結合得出含義。Amazon 推薦引擎就是一個知識圖表示例。



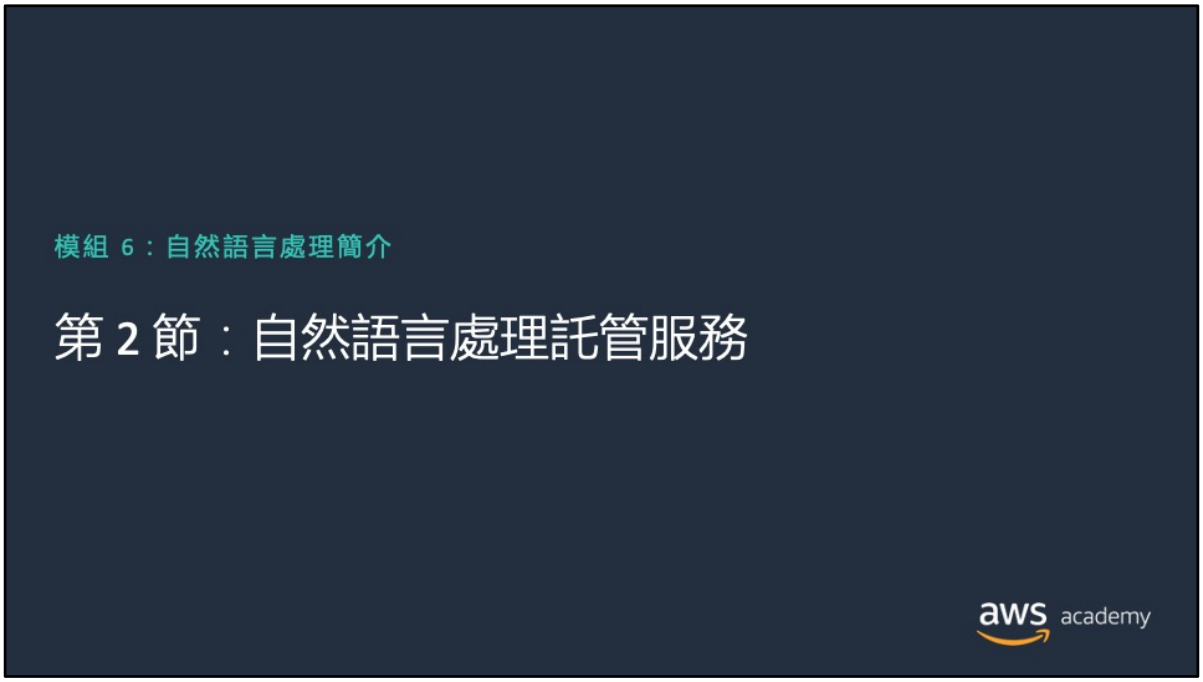
本模組中這部分內容的要點包括：

- 作為一個領域，NLP 早於機器學習。
- NLP 開發直接映射到 ML 開發流程。
- NLP 的一些主要使用案例包括搜索查詢分析、人機交互以及市場或社會研究。
- 由於人類語言並不精確，因此 NLP 開發起來比較困難。

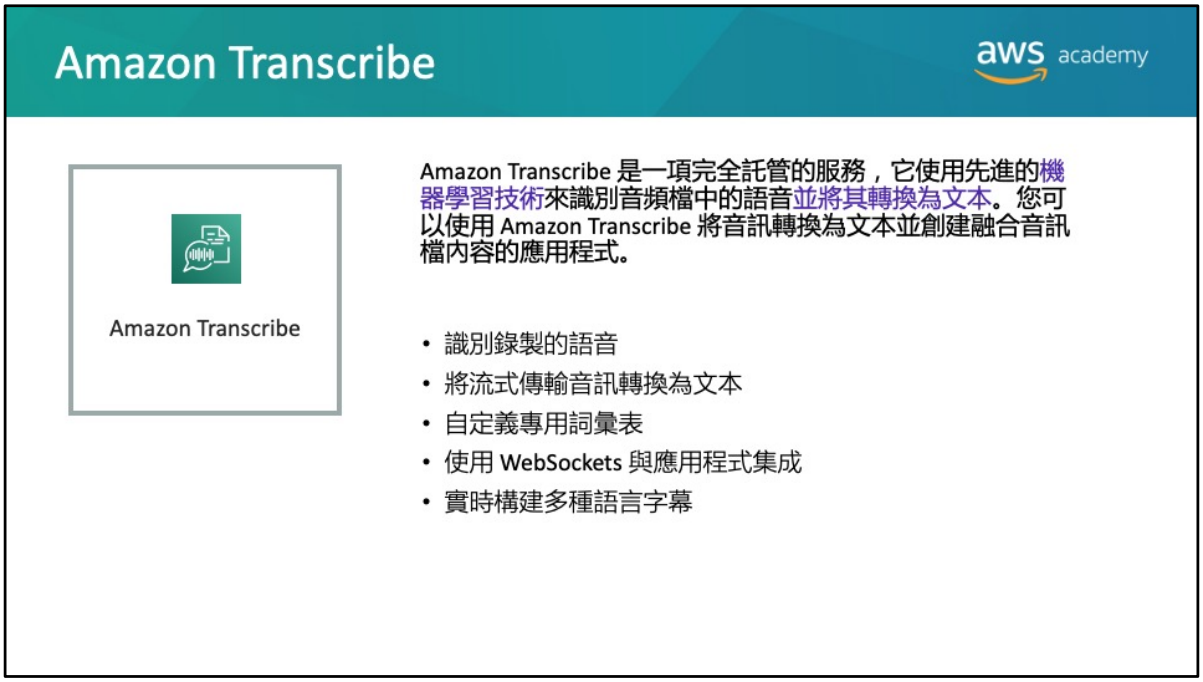


本模組中這部分內容的要點包括：

- 作為一個領域，NLP 早於機器學習。
- NLP 開發直接映射到 ML 開發流程。
- NLP 的一些主要使用案例包括搜索查詢分析、人機交互以及市場或社會研究。
- 由於人類語言並不精確，因此 NLP 開發起來比較困難。



介紹“第 2 節：自然語言處理託管服務”。在本節中，您將看到五種可用於各種 NLP 使用案例的託管 Amazon ML 服務。這些服務可以簡化 NLP 應用程式的創建流程。



Amazon Transcribe 是您將瞭解的第一項託管機器學習服務。您可以使用 Amazon Transcribe 識別音訊檔中的語音並生成轉錄。Amazon Transcribe 可以識別音訊檔中的特定語音，您可以為專門用於特定領域的術語創建自訂詞彙表。您還可以通過與 WebSockets 集成來向應用程式添加轉錄服務。WebSockets 提供了一個面向互聯網的介面，支援應用程式和 Amazon Transcribe 之間的雙向通信。

Amazon Transcribe 使用案例

醫療轉錄

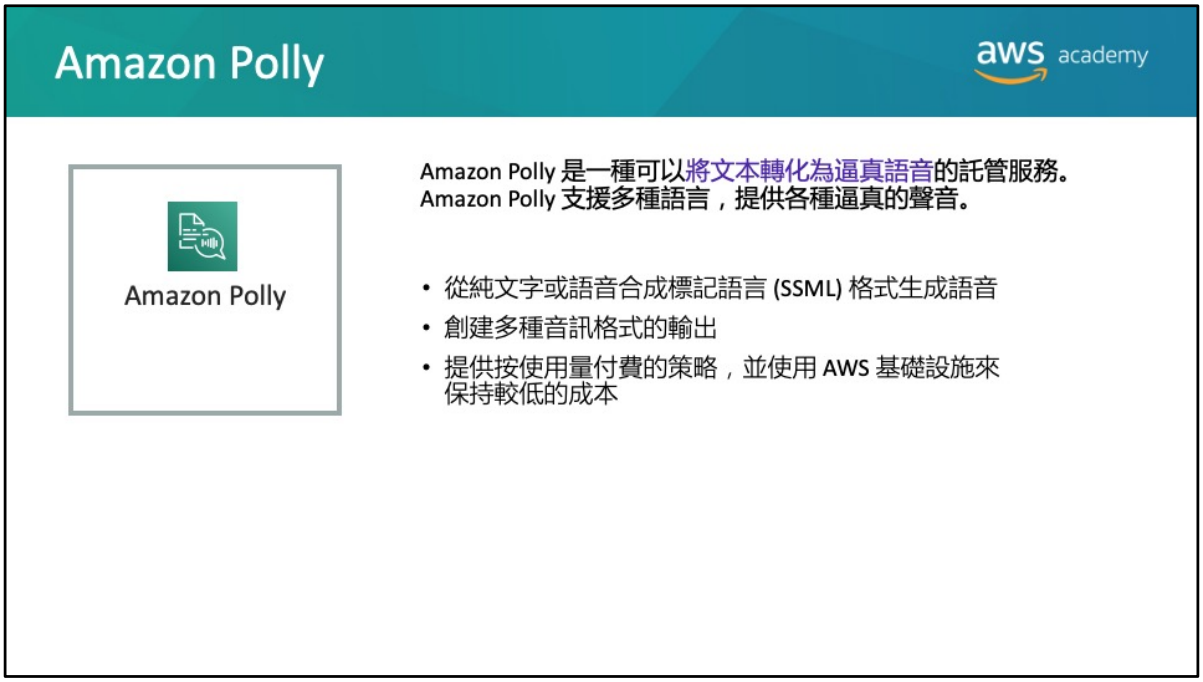
字幕

為流媒體內容添加標籤

監控呼叫中心

Amazon Transcribe 比較常見的一些使用案例包括：

- 醫療轉錄 – 醫學專家可以用語音錄製筆記，而 Amazon Transcribe 可以將其語音筆記捕獲為文本。
- 視頻字幕 – 視頻製作組織可以從視頻自動生成字幕。此外，還可以針對即時動態即時完成此操作，以添加隱藏式字幕 (CC)。
- 為流媒體內容添加標籤 – 媒體公司可以捕獲並標記內容，然後將內容饋送到 Amazon Comprehend 中進行進一步分析。
- 監控客戶呼叫中心 – 公司可以對客戶服務或銷售電話進行錄音，然後對結果進行分析，以用於培訓或戰略商機。



Amazon Polly 可將文本轉換為逼真的語音。您可以輸入純文字檔或語音合成標記語言 (SSML) 格式檔。SSML 是一種標記語言，可用於提供有關語音應如何發音的特殊指令。例如，您可能想在語音流中加入停頓。您可以添加 SSML 標記，來指示 Amazon Polly 在兩個單詞之間暫停。

您還可以將語音從 Amazon Polly 輸出到 MP3、Vorbis 和脈衝代碼調製 (PCM) 音訊流格式。

Amazon Polly 應用廣泛。Amazon Polly 的常見使用案例包括移動應用程式（例如新聞閱讀器）、遊戲、電子學習平臺以及適用于視障人士的無障礙應用。

Amazon Polly 適用於美國1996 年版健康保險流通與責任法案 (HIPAA) 和支付卡行業資料安全標準 (PCI DSS) 中的受管制工作負載。

Amazon Polly 使用案例

新聞服務製作

語言培訓

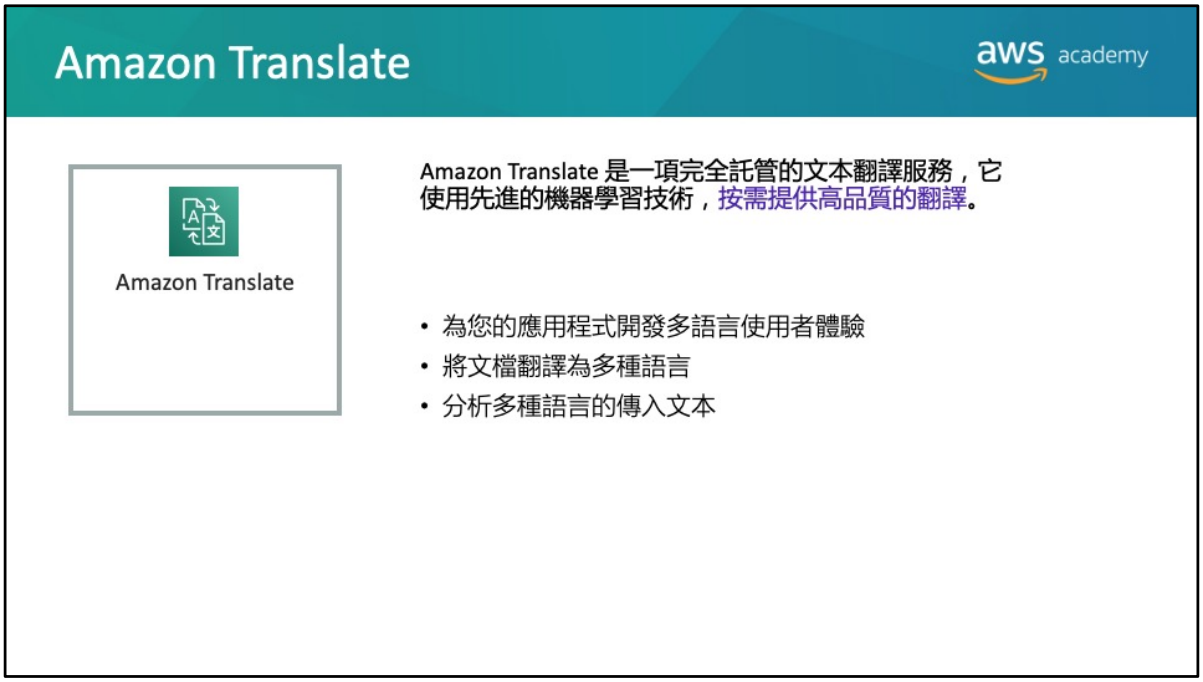
導航系統

動畫製作

Amazon Polly 比較常見的的一些使用案例包括：

- 新聞服務製作 – 大型新聞公司使用 Amazon Polly 直接將他們的文字報導轉換為聲音內容。
- 語言培訓系統 – 語言培訓公司使用 Amazon Polly 來創建學習新語言的系統。
- 導航系統 – 將 Amazon Polly 嵌入到映射應用程式設計發展介面 (API)，以便開發人員可以向基於地理位置的應用程式中添加語音。
- 動畫制作 – 動畫師使用 Amazon Polly 為角色配音。

您的講師現在將演示 Amazon Polly 或向您提供演示錄影的存取權限。



借助 Amazon Translate，您可以在應用程式中打造多語言體驗。您可以創建這樣一個系統：使用一種語言閱讀文檔，然後使用另一種語言呈現或存儲文檔。您還可以將 Amazon Translate 用於文檔分析系統。

Amazon Translate 可以與其他 Amazon ML 服務（如 Amazon Comprehend、Amazon Transcribe 和 Amazon Polly）完全集成。各項集成的用途：

- 通過與 Amazon Comprehend 集成，您可以提取命名實體、情緒和關鍵短語
- 通過與 Amazon Transcribe 集成，您可以生成多語言字幕
- 通過與 Amazon Polly 集成，您可以朗讀翻譯內容

Amazon Translate 使用案例

國際網站

軟體當地語系化

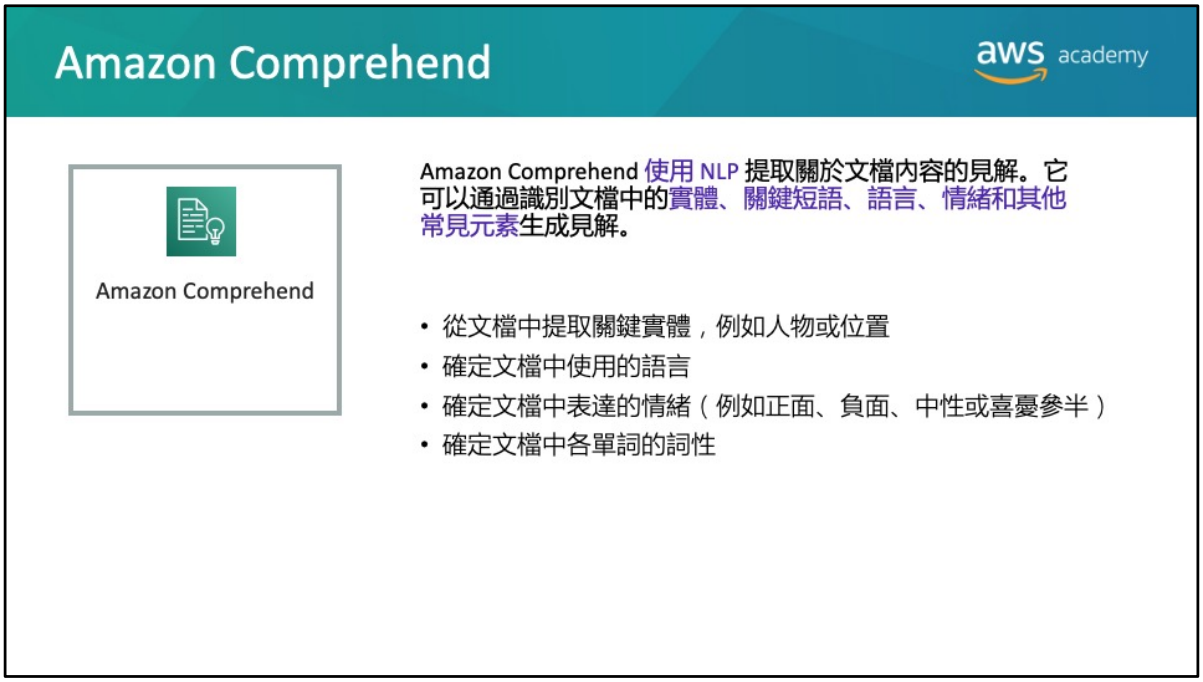
多語言 Chatbot

國際媒體

Amazon Translate 比較常見的一些使用案例包括：

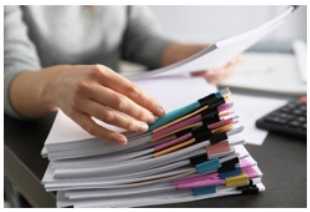
- 國際網站 – 您可以使用 Amazon Translate 快速全球化您的網站。
- 軟體當地語系化 – 當地語系化是所有面向全球使用者的軟體的主要成本。Amazon Translate 可以減少軟體發展時間並顯著降低軟體當地語系化成本。
- 多語言 Chatbot – 使用 Chatbot 為應用程式創建更具人性化的介面。借助 Amazon Translate，您可以創建一個會講多種語言的 Chatbot。
- 國際媒體管理 – 管理面向全球受眾的媒體的公司使用 Amazon Translate 降低當地語系化成本。

您的講師現在將演示 Amazon Translate 或為您提供演示錄影的存取權限。



Amazon Comprehend 採用與了許多本模組前面介紹的 NLP 技術。您可以提取關鍵實體、執行情緒分析，並用詞性標記單詞。

Amazon Comprehend 使用案例



文檔分析

欺詐檢測

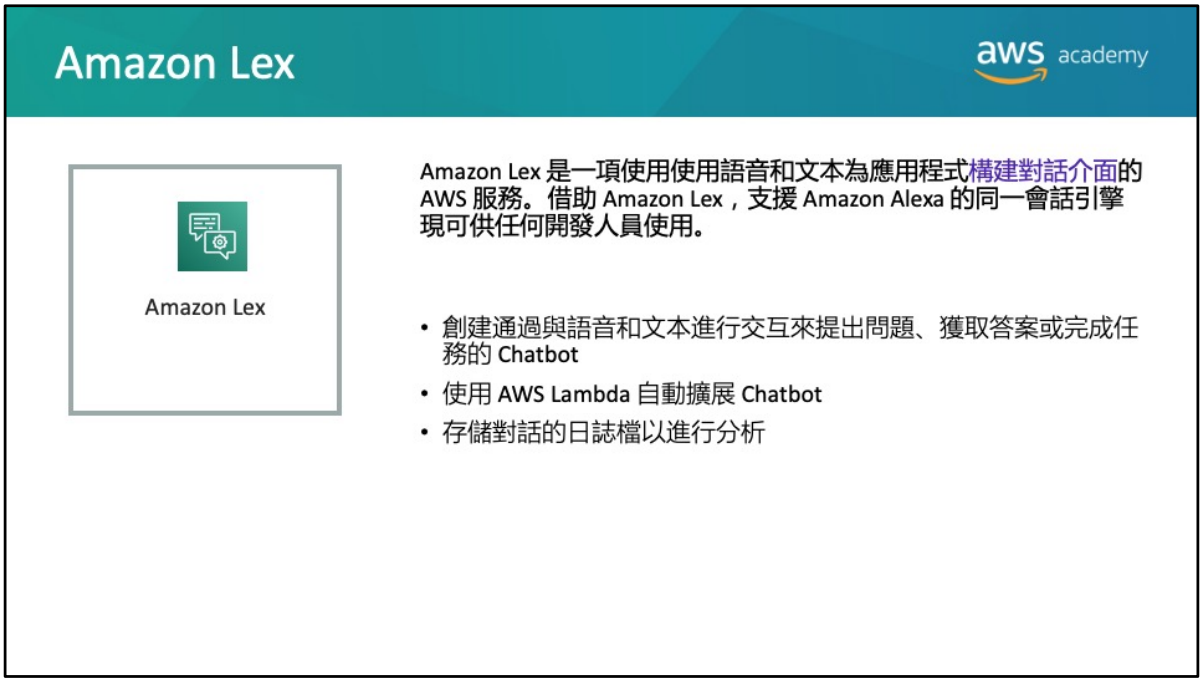
移動應用程式分析

內容管理

Amazon Comprehend 比較常見的一些使用案例包括：

- 分析法律和醫療文檔 – 法律、保險和醫療組織使用 Amazon Comprehend 來執行您在本模組中瞭解到的許多 NLP 功能。
- 財務欺詐檢測 – 銀行、金融和其他機構使用 Amazon Comprehend 來檢查非常龐大的財務交易資料集，以發現欺詐行為並尋找非法交易的模式。
- 大型移動應用程式分析 – 移動應用程式的開發人員可以使用 Amazon Comprehend 來查找其應用程式使用模式，以便他們改進設計。
- 內容管理 – 媒體和其他內容公司可以使用 Amazon Comprehend 標記內容，以進行分析和管理。

您的講師現在將演示 Amazon Comprehend 或為您提供演示錄影的存取權限。



借助 Amazon Lex，您可以在應用程式中添加人類語言前端。Amazon Lex 讓您可以使用支援 Amazon Alexa 的同一會話引擎。您可以創建 AWS Lambda 函數進行按需擴展，從而自動增加容量。此外，您可以存儲對話的日誌檔以進行進一步分析。

Amazon Lex 使用案例

庫存和銷售

互動式助理

客戶服務介面

資料庫查詢

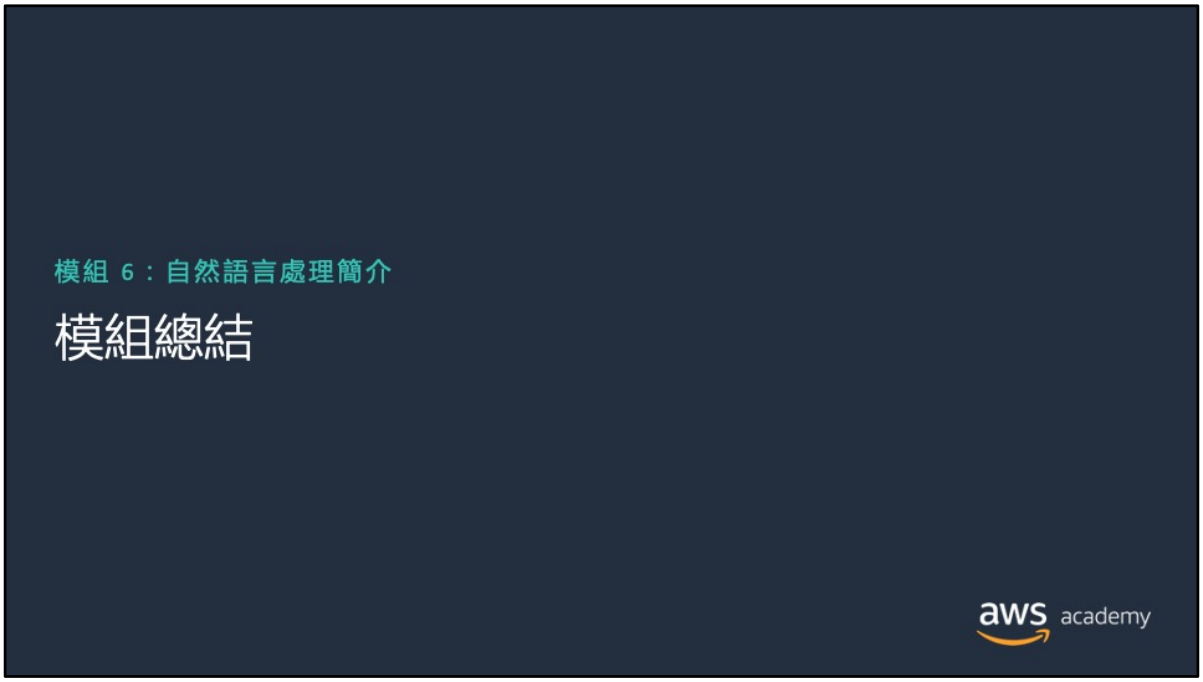
Amazon Lex 比較常見的一些使用案例包括：

- 構建用於庫存管理和銷售的前端介面 – 語音介面正變得越來越普遍。公司使用 Amazon Lex 將 Chatbot 添加到庫存和銷售應用程式。
- 開發互動式助理 – 將 Amazon Lex 與其他 ML 服務相結合後，客戶可以為許多不同行業開發更精密的助理。
- 創建客戶服務介面 – 擬人語音應用程式正迅速成為客戶服務應用程式的標配。Amazon Lex 可以節約時間和並提高這些 Chatbot 的品質。
- 使用擬人語言查詢資料庫 – 將 Amazon Lex 與其他 AWS 資料庫服務結合使用，創建具有擬人語言介面的精密資料分析應用程式。

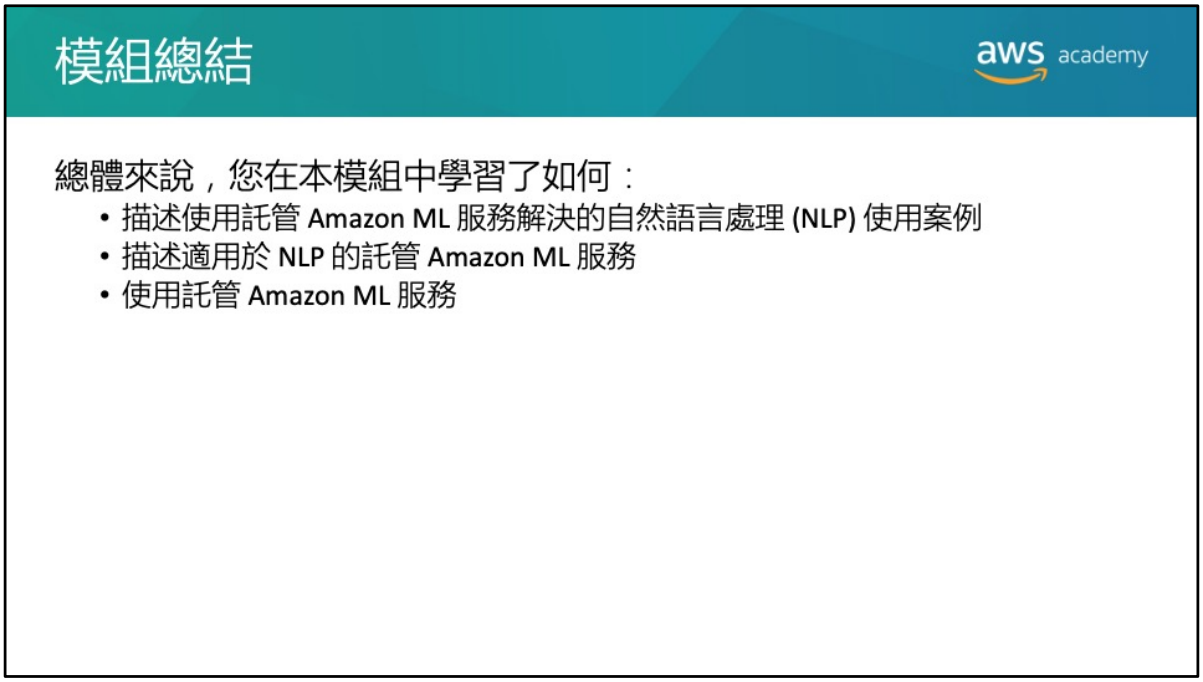
本模組中這部分內容的要點包括：

- Amazon Transcribe 可以自動將口語語音轉換為文本。
- Amazon Polly 可以將書面文本轉換為口語語音。
- Amazon Translate 可以在語言之間進行即時翻譯。
- Amazon Comprehend 可以將本模組中討論的許多 NLP 使用案例自動化。
- Amazon Lex 可以為應用程式創建擬人介面。

現在，您將完成模組 6 – 引導式實驗：創建機器人來安排預約



現在，我們來回顧和總結一下本模組，然後進行知識測驗。



總體來說，您在本模組中學習了如何：

- 描述使用託管 Amazon ML 服務解決的自然語言處理 (NLP) 使用案例
- 描述適用於 NLP 的託管 Amazon ML 服務
- 使用託管 Amazon ML 服務

完成知識測驗

現在可以完成本模組的知識測驗。

如果您想瞭解本模組所涵蓋主題的更多資訊，下面這些資源可能會有所幫助：

- Amazon Comprehend 是什麼？
- Amazon Polly 是什麼？
- Amazon Lex 是什麼？
- Amazon Transcribe 是什麼？
- Amazon Translate 是什麼？

感謝您完成本模組的學習。